ESTIMATING POPULATION DIVERSITY WITH UNRELIABLE LOW FREQUENCY COUNTS

JOHN BUNGE*
Department of Statistical Science, Cornell University,
Ithaca, NY 14853, USA
**E-mail: jab18@cornell.edu*
www.northeastern.edu/catchall

DANKMAR BÖHNING
Department of Mathematics and Statistics, University of Reading,
Reading RG6 6BX, UK
E-mail: d.a.w.bohning@reading.ac.uk

HEATHER ALLEN
Food Safety and Enteric Pathogens Research Unit, National Animal Disease Center,
Agricultural Research Service, Ames, Iowa, 50010, USA
E-mail: heather.allen@ars.usda.gov

JAMES A. FOSTER
Department of Biological Sciences, University of Idaho,
Moscow, ID 83844, USA
E-mail: foster@uidaho.edu

We consider the classical population diversity estimation scenario based on frequency count data (the number of classes or taxa represented once, twice, etc. in the sample), but with the proviso that the lowest frequency counts, especially the singletons, may not be reliably observed. This arises especially in data derived from modern high-throughput DNA sequencing, where errors may cause sequences to be incorrectly assigned to new taxa instead of being matched to existing, observed taxa. We look at a spectrum of methods for addressing this issue, focusing in particular on fitting a parametric mixture model and deleting the highest-diversity component; we also consider regarding the data as left-censored and effectively pooling two or more low frequency counts. We find that these purely statistical "downstream" corrections will depend strongly on their underlying assumptions, but that such methods can be useful nonetheless.

Keywords: microbial diversity, mixture model, species problem, capture-recapture, left-censored data.

1. Introduction

Classical population diversity estimation is based on the assumption that the observed data counts — so many taxa observed once, twice, etc., in the sample — are correctly recorded. One then uses these "frequency count" data $(i, f_i), i = 1, 2, 3, \ldots$, where f_i is the number of taxa occurring i times in the sample, to estimate the total number of taxa in the population. There is a substantial literature on this problem, which can be framed either in terms of taxa or species, or in terms of the number of individuals in the population. In the latter case we have the capture-recapture situation, where one wishes to estimate the population size. Here we

will take diversity to be the total number of taxa or classes, denoted by C, but our discussion is equally applicable to the capture-recapture problem.

It may be argued that "bounds on this quantity $[C]$ may be highly uncertain because a small fraction of the environment could be composed of a vast number of different species."[1] We will proceed nonetheless under the assumption that such estimation is in principle statistically feasible;[2] in this case, however, the quality of the observed data counts is crucial to the accuracy and precision of the diversity estimate. Ordinarily one assumes that the sample consists of an (approximately) random selection of individuals from the population; the sampled individuals are then sorted or divided into classes, and the sizes of the sample classes are recorded. The numbers of sample classes of various sizes constitute the frequency count data. In some situations, though, the initial identification of the sampled individuals may be at fault, or the classification into sample classes may be questionable. This is the case, for example, when dealing with high-throughput DNA sequencing data, which are prone to errors of various types. These errors may arise at various stages, in particular the identification of the sequences and the clustering algorithms used to combine sequences into clusters or taxa may be questionable. The end result is that the number of low frequency counts — especially the singletons (f_1) — may be "artificially" inflated compared to what would be obtained by a data-collection process with a lower error rate.

The question then arises: Is it possible to *statistically* discount or down-weight the low frequency counts in the data (which are quite influential in estimating total diversity)? Could such a procedure be based solely on the available data, without knowledge of the mechanisms that produced the errors? Put this way the answer would seem to be No; and certainly the best procedure would be to correct the errors at the source before they enter the data stream. However, there are some *ex post facto* statistical analyses that are worth considering, if not as a complete fix then at least as a benchmark, to provide diversity estimates under different sets of assumptions. Such adjustments could also provide lower bounds for the total diversity that are comparatively insensitive to the errors in question. In this paper we describe the spectrum of available statistical approaches to this issue, examining two in detail: fitting a model with several diversity components and deleting the component corresponding to highest diversity; and declaring the low frequency counts uncertain in measurement and "left-censoring" the data. We examine these methods using a high-diversity dataset derived from a phage metagenome, and we compare numerical results based on this dataset. We conclude that some of these procedures may be useful, but that the investigator must take careful account of their underlying assumptions and the implications thereof.

2. Error correction at the source

The ideal way to deal with uncertain low frequency counts is to render them certain; i.e., to correct any errors in the counts before they are analyzed. This would naturally take place at the initial data collection or generation stage. In the case of high-throughput sequencing, there have been remarkable recent developments in this regard such as the programs PyroNoise and AmpliconNoise, which aim at "separately removing 454 sequencing errors and PCR single base errors."[3] It is arguably true that corresponding advances, or at least standardization, are

also needed at the computational processing stage (alignment and clustering of sequences). These matters are non-statistical and hence we do not pursue them further here, but they are the subject of current research in bioinformatics.

3. Lower bounds

One conceptually simple approach is to require only lower bounds for the total diversity. Indeed since the beginning of statistical population diversity estimation, long before the appearance of high-throughput sequencing, researchers have been interested in comparing formal estimates of diversity (based on frequency count data regarded as error-free) with lower bounds or benchmarks for the total diversity. In fact, it has been argued mathematically that under the least restrictive nonparametric assumptions, only lower bounds (as opposed to estimates or upper bounds) are possible. Here we assume that we are willing to make the minimal structural modeling assumptions required to estimate C (and we do not enter further into this particular mathematical issue). However, it is possible to define a maximally biased model, namely, the equal-class-sizes model, i.e., the assumption that the population is *equally* subdivided into taxa or classes C. It is known that estimates of C made under this assumption are maximally biased downward, in a certain sense.[4] At least two such estimates are available: parametric, which is the maximum likelihood estimator of C under the unmixed Poisson model (each class contributes a Poisson number of representatives to the sample, and the mean number of representatives is equal across classes); and the nonparametric Good-Turing estimate, which is derived differently but still falls under the equal-sizes assumption. These estimates are typically reported along with the standard errors that would be appropriate under the equal-sizes assumption. It may be of some value to consider these downwardly-biased estimates when the low frequency counts are questionable, even though they are based on different modeling assumptions (correct data, equal class sizes). We look at numerical results below for our example dataset.

4. Deleting the high-diversity component of a mixture model

A promising but potentially rather extreme approach to the problem of uncertain low frequency counts consists of statistically reducing the lowest-abundance counts. This is based on a finite mixture model for estimating C, i.e., a model consisting of a convex combination (weighted average) of several components, which correspond to several different levels of diversity in the population.[2] A three-component model, for example, is more flexible than a simple one-component model, because it essentially allows for three categories or levels of diversity. Each class in the population contributes representatives to the sample according to one of three (mean) rates: the lowest rate corresponds to the rarest or smallest classes and the highest rate to the largest or most abundant classes. (Some mathematical details for the particular finite mixture model used here are given in the Appendix.) Such a model can be fitted directly to frequency count data by maximum likelihood, and each of the three components (in this example) is then identified by its mean rate and by the proportion of the population to which it corresponds. If the low frequency counts are supposed to be uncertain, one can then delete (mathematically subtract) the highest diversity or lowest abundance component

from the fitted model, and base the projection of total diversity on the remaining, lower diversity components. This is mathematically straightforward, but as we will see in the numerical example below, the resulting reduction in the estimate of total diversity may be as much as two orders of magnitude, leading to conceptual questions about what exactly is the target of estimation under the reduced (highest diversity component deleted) model.

5. Left-censored data

The final approach we considered is left-censoring the frequency count data, as follows. Suppose we assume that the number of classes observed in the sample is correct, but that there is uncertainty regarding the exact values of the counts for certain classes, especially the low frequency counts. That is, for each observed class in the sample we declare the possibility of "measurement error," so that a singleton could perhaps have been a doubleton, or a doubleton a singleton. In other words, we separate the observed frequency counts at some level L into a pooled or total low frequency count, say $f_L^* = f_1 + f_2 + \ldots + f_L$, vs. the higher frequency counts $f_{L+1}, f_{L+2}, f_{L+3}, \ldots$. In general statistical terms this is called left-censored data (at L). Under this scenario the observed number of classes in the sample, say c, is preserved exactly, but we suppose that (for sample class sizes L and below) the number of individuals contributing to the observation of a given class is uncertain. This has the effect of essentially reducing the apparent diversity of the sample, because although there are still c classes, the number of counts of size L and below are combined. In particular, the number of singletons is no longer "known" and contributes only to the total f_L^*. This seems not unreasonable, but on the other hand it has the effect of rendering the total number of "individuals," namely $1f_1 + 2f_2 + 3f_3 + \ldots$, uncertain, which may or may not be logical in a given application. Maximum likelihood estimation of C based on left-censored data is relatively straightforward in this problem, although it does not admit a closed-form formula but requires numerical optimization (see the Appendix for a brief mathematical outline under the single exponential model). Robb and Böhning[5] dealt in detail with a closely related version of this problem for the capture-recapture setting. The reduction in the estimate of total diversity is less dramatic relative to the unadjusted estimate than under the previous approach (deleting the high-diversity component). However, the structural assumptions underlying this procedure may or may not be reasonable from the investigator's point of view.

6. Possible Bayesian extensions

It is possible to lower the estimate of total diversity by establishing a prior distribution on the population diversity C. This is (part of) the Bayesian approach, and there are two basic viewpoints. Some investigators prefer a method that puts minimal constraints on C, and in fact mathematically minimizes the prior information assigned to C. Such a method is called noninformative or objective Bayesian, and has been studied in this context by Barger and Bunge,[6] Farcomeni and Tardella,[7] and others. We do not pursue this further here because in our present setting we are interested in specifying prior information (also, no relevant computer software is readily available). The alternative, then, is an informative or even subjective Bayesian approach, in which the investigator specifies more or less strong prior assumptions

regarding C. Perhaps the simplest version of this is to set an upper bound $C_{\max}$ for C, which can sometimes be done on the basis of biological or chemical considerations; for example, the taxonomic diversity of a microbial population cannot exceed the number of cells in it. The prior distribution of C is typically taken to be approximately flat, meaning that all values up to $C_{\max}$ are equally likely *a priori*. This has been explored by Manrique-Vallier and Fienberg[8] and others. Alternatively one may opt for an informative or even subjective prior distribution on C with no upper bound. This may decrease rapidly, for example, meaning that the larger the value of C, the less likely it is thought to be *a priori*. In this case the influence of the prior on the final diversity estimate may be considerable, and corresponding sensitivity analysis is called for. Again such methods have been dealt with to some extent in theory,[6] but no software is readily available. A further extension would be to apply an informative prior to the components of a mixture model, partially but not entirely downweighting the high-diversity component. This remains to be explored. In summary, the Bayesian approach may be promising in terms of controlling the estimate of C, but it is an open area for research in this regard and is beyond our scope here.

7. Example, numerical results and remarks

Our sample dataset is based on a phage metagenome.[9] More specifically, the frequency counts were derived from a contig spectrum from a swine fecal metagenome; the contig spectrum was generated using Circonspect via the CAMERA pipeline.[10] For the purpose of discussion here we can assume that we are interested in estimating the taxonomic diversity of this metagenome. The frequency counts are given in Table 1. The total number of observed taxa

Table 1. Phage metagenome frequency count data

i	f_i	i	f_i	i	f_i	i	f_i
1	4736	12	8	23	2	34	1
2	521	13	7	24	3	35	1
3	152	14	6	25	3	36	1
4	69	15	5	26	1	37	1
5	46	16	4	27	2	38	1
6	27	17	4	28	1	39	1
7	21	18	3	29	2	40	1
8	18	19	3	30	2	41	1
9	16	20	3	31	1	43	1
10	10	21	3	32	1	45	1
11	9	22	2	33	1	52	1

is $c = 5703$. It is clear even without graphing the data that the sample diversity is high: for instance, the number of singletons is almost an order of magnitude higher than the number of doubletons. There is some basis to believe that the experimental and bioinformatic procedures that generated these data are prone to erroneous inflation of the low frequency counts,[3] so this dataset is a good test-bed on which to compare the damping approaches described above. The output from the program CatchAll v.3.0[11] that is relevant for our purposes is shown in Table 2. (Here for simplicity we analyze the full dataset, i.e., all frequency counts up to

and including the maximum f_{52}, rather than checking for right-hand outlier cutoff values τ as discussed elsewhere.[11]) Note first that the best estimate selected by CatchAll, under the

Table 2. Phage data analysis. Est Div = estimated total diversity; SE = standard error; LCB/UCB = lower/upper 95% confidence bounds

Method	Est Div	SE	LCB	UCB
Poisson	8730	103	8535	8938
GoodTuring	11690	346	11050	12407
ThreeMixedExp	67792	8656	53009	87195
Discounted: TwoMixedExp	1727	221	1410	2305

assumption that the data are error-free, is the (finite) mixture of three exponentially-mixed Poisson components, or equivalently a mixture of three geometric distributions. This yields an estimated total diversity of 67792 (SE 8656). This value is believed on scientific grounds to be too high, resulting from an unknown number of possibly artifactual or erroneous singleton (and perhaps doubleton and tripleton) counts. Considering lower bounds as discussed in Section 3 above, Table 2 displays the results for the Poisson maximum likelihood estimate (MLE) and the Good-Turing nonparametric estimate. Both of these are based on the structural assumption of equal class sizes in the population, but the Poisson estimate is the formal MLE under that model and the Good-Turing estimate is a simple nonparametric approximation thereof (which may possess certain robustness properties under mild departures from the equal-class-sizes model). Both of these estimators assume that (all of) the frequency count data are correct, but are maximally biased downward if the true population does not have equal class sizes — which it certainly does not, irrespective of the correctness of the data. Thus they constitute lower bounds for the total diversity if the data are correct, and the same is true *a fortiori* if the sample diversity is incorrectly inflated. It is interesting to note that the two estimates do not agree exactly; this is due to the different assumptions of the two estimators (parametric/nonparametric), and also to the very high number of singletons in this particular dataset. (The SE's and 95% confidence intervals associated with these estimates are not too meaningful in this case: they are values obtained under the assumption of equal class sizes in the population.)

The best parametric model selected for the phage data by CatchAll's selection routine is the mixture of three components. Figure 1 shows the frequency count data and the fitted model. This model specifies three levels of abundance in the population: high, with mean sampling rate 9.66 and proportion 0.004; medium, with mean sampling rate 1.160 and proportion 0.022; and low, with mean sampling rate 0.076 and proportion 0.975. This means that the highest abundance classes enter the sample at approximately $9.66/0.076 \approx 127$ times the rate of the lowest abundance classes on average. The high-abundance classes account for 0.4% of the population, the medium-abundance classes 2.2%, and the low-abundance classes 97.5% (allowing for rounding error). Figure 2 shows these three components with a logarithmic vertical axis for frequencies 0–10 (f_0 is the estimated zero count, i.e., the estimated number of unobserved taxa). If we delete the low-abundance (high diversity) component of the fitted

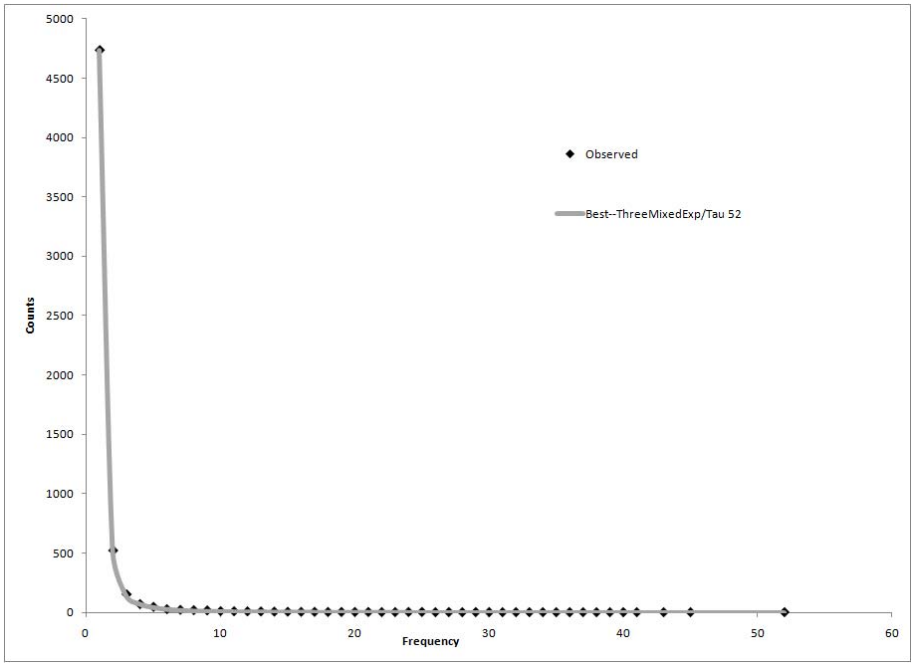

Fig. 1. Observed phage frequency count data vs. best fitted model.

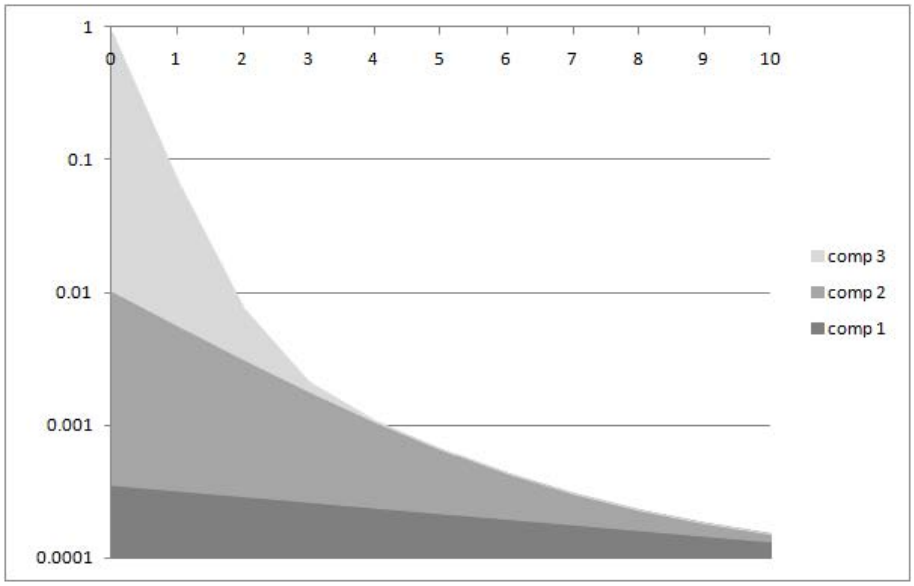

Fig. 2. Three components of fitted mixture model, logarithmic vertical axis, frequencies 0–10.

model (component 3 in Figure 2) and project the total population diversity based on the two remaining components, we obtain the results shown in the last row of Table 2. The estimated total diversity is now 1727, corresponding to the (sum of) the two higher-abundance components of the model, a reduction of 97.5% from the original unadjusted estimate, or more than an order of magnitude. This reduction may be too severe, since it deletes the entire low-abundance component, but there is no way, based on the data alone, to distinguish between

real and erroneous low frequency counts.

Finally we consider left-censoring. No general software has yet been developed for this so for simplicity in exploring the procedure we use a single-component exponential (geometric) parametric model, as the computational complexity for the higher-order mixture models is considerable. This is a low but not minimum diversity model for the frequency count data, so it produces estimates that are above the Poisson and Good-Turing values; however like those models it tends to fit real data poorly (because it allows for insufficient diversity), and produces estimates below those of the best selected model (the three-component mixture in this case). It is however sufficient for our demonstration here. Table 3 shows the estimated total diversity at several different censoring points. As noted above $f_L^* = f_1 + f_2 + \ldots + f_L$, so for $L = 1$ there

Table 3. Estimated total diversity at different left censoring points under single geometric model. L = censoring point, Est Div = estimated total diversity.

L	1	2	3	4	5	10
Est Div	14880	12693	11267	10339	9712	8202

is no censoring; for $L = 2$ the singletons and doubletons are pooled, for $L = 3$ the singletons, doubletons and tripletons are pooled, and so on. The unadjusted estimate under the single geometric model (14880) is between those derived from the equal-abundance (8730) and the selected three-component mixture models (67792), as expected. Also as expected the estimate declines as the censoring point L increases, i.e., as low frequency information is removed from the data. Figure 3 displays the fitted values for the first 10 frequency counts $(f_1, \ldots, f_{10})$ for the best model (mixture of three exponentials/geometrics), the single geometric with no censoring, the single geometric with censoring point $L = 2$, and the equal-abundance unmixed Poisson models.

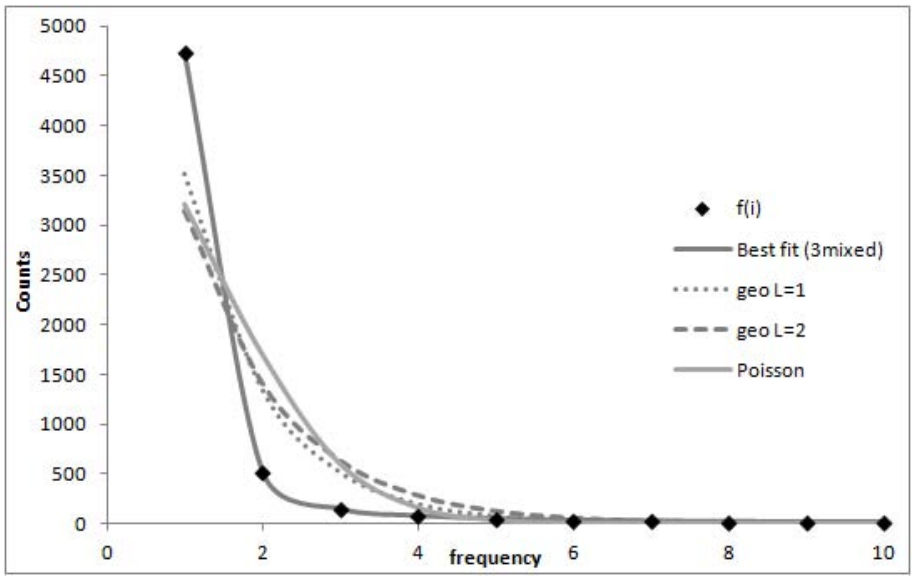

Fig. 3. Best fitted vs. discounted models, frequencies 1–10.

It is tempting to look for convergence of the estimates as L increases, and this is logical in a sense since if L is set to its maximum (the maximum frequency in the data, here 52) then all of the data are pooled into a single class. However it is not clear how we would interpret such convergence statistically. Thus this method presents us with the need to choose a cutoff L more or less arbitrarily. Furthermore it is not obvious that this method is reasonable in the microbial diversity application, even though it does damp the estimate of total diversity (as a function of L), because it essentially declares that every observed taxon does exist, and that in fact the low sample frequency classes might have been observed with higher frequency had the data been correctly recorded. On the other hand the censoring approach might be reasonable in the capture-recapture setting, where one might suppose that apparent singletons (say) should have been recorded more often but were inadvertently missed. In summary, this method is intriguing but presents conceptual and statistical difficulties at present, especially for microbial diversity estimation.

In summary, then, it is not surprising that any proposed *ex post facto* statistical approach to compensate for incorrect or uncertain low frequency counts will depend strongly on its underlying structural assumptions. Another approach, which has not yet been attempted, would be to construct a probabilistic model for the process by which errors in the low frequency counts are generated, and to incorporate this in diversity estimation. However, in models of this type any error-generation process parameter, say $p := P$(incorrect generation of a singleton), will typically be statistically confounded with the parameters of the model used to estimate diversity (the mean sampling rates, etc.), leading again to the problem of being unable to separate true from spurious counts. Still, it is not impossible that progress could be made, either in the Bayesian direction mentioned in Section 3, or via error-process modeling, or by some combination of these, and this is a topic for further research. The gold standard remains, however, to obtain correct data at the outset, or to correct it at the source.

References

1. M. Lladser, R. Gouet and J. Reeder, *PLoS ONE* **6**, (2011). doi:10.1371/journal.pone.0021105
2. J. Bunge and K. Barger, *Biometrical Journal* **50**, 971 (2008).
3. C. Quince, L. Anders and R. J. Davenport, *BMC Bioinformatics* **12**, 38 (2011).
4. D. Böhning and D. Schön, *J. Roy. Statist. Soc. Ser. C* **54**, 721 (2005).
5. M. Robb and D. Böhning, *Biometrical Journal* **53**, 75 (2011).
6. K. Barger and J. Bunge, *J. Bayesian Analysis* **5**, 765 (2010).
7. A. Farcomeni and L. Tardella, *Test* **19**, 187 (2010).
8. D. Manrique-Vallier and S. Fienberg, *Biometrical Journal* **50**, 1051 (2008).
9. H. Allen, T. Looft, D. Bayles, S. Humphrey, U. Levine, D. Alt, and T. Stanton (2011). Submitted.
10. S. Sun *et al.*, *Nucleic Acids Research* **39**, 546 (2011).
11. J. Bunge, *Pac. Symp. Biocomput.*, 121 (2011).

Acknowledgments

We thank Rob Knight for his ideas in regard to discounting uncertain low frequency counts. This research was funded in part by National Science Foundation grant DEB-08-16638 to JB. This research was conducted using the resources of the Cornell Center for Advanced

Computing, which receives funding from Cornell University, the National Science Foundation, and other leading public agencies, foundations, and corporations. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S Department of Agriculture. USDA is an equal opportunity provider and employer.

Appendix A. Some mathematical details

Deleting the high-frequency component. The three-component mixture model used in the numerical example is

$$P(X=j;\theta)=\theta_4\left(\frac{1}{1+\theta_1}\left(\frac{\theta_1}{1+\theta_1}\right)^j\right)+\theta_5\left(\frac{1}{1+\theta_2}\left(\frac{\theta_2}{1+\theta_2}\right)^j\right)+(1-\theta_4-\theta_5)\left(\frac{1}{1+\theta_3}\left(\frac{\theta_3}{1+\theta_3}\right)^j\right),$$

$j=0,1,2,\ldots,,\theta_1,\theta_2,\theta_3>0$, $0<\theta_4,\theta_5<1$, where X is the number of representatives contributed to the sample by an arbitrary species. Assuming without loss of generality that the highest-diversity component is the first listed, then the discounted estimate of diversity is $(1-\hat{\theta}_4)\hat{C}$, with associated standard error equal to $(1-\hat{\theta}_4)$SE, where $\hat{\theta}_4$ is the MLE of θ_4, $\hat{C}$ is the unadjusted (conditional) MLE of C, and SE is the unadjusted SE of $\hat{C}$.[11]

Left-censoring. The single geometric model discussed in the text is $P(X=j;p)=(1-p)p^j$, with j and X as above, $0<p<1$. The zero-truncated version of this is $P(Y=j;p)=(1-p)p^{j-1}$, $j=1,2,\ldots$. The likelihood of a dataset $f_1,f_2,\ldots$ censored at L is then

$$\ell(p):=\left((1-p)p^0+(1-p)p^1+\ldots+(1-p)p^{L-1}\right)^{f_1+f_2+\ldots+f_L}\prod_{j>L}\left((1-p)p^{j-1}\right)^{f_j},$$

and the estimates given in the example are derived by maximizing ℓ with respect to p for the given dataset.