Attempts at Quantitative Metazoan Metabarcoding are Difficult

John Wares and Paula Pappalardo

02 septiembre, 2015

Introduction

We start with what may seem like a trivial question: assume that you have been told that a series of fair coin flips resulted in 60% 'heads', 40% 'tails'. This is the only information given, but you already have made a judgment about how many coin flips occurred, and perhaps have generated a probability distribution in your head where the highest likelihood is for 5 or 10, rather than 50 or 100, events. This is taking advantage of what we know about the probability mass function of a binomial distribution, where the observed number of 'successes' in a series is related to the probability of success (here, presumably 50%) and the number of trials.

Here, we consider whether the same principle could be used for improving the efficiency of exploring the presence, distribution, and abundance of genetic biodiversity. Documenting the distribution and abundance of biodiversity - in many habitats, at multiple scales - is perhaps more important now than ever as scientists evaluate how populations are responding to environmental change. Though technological advances have rapidly improved some elements of this (Nagendra 2001; Bourlat et al. 2013), there are still glaring deficiencies in our ability to efficiently catalog diversity, even in small domains or limited taxonomic surveys.

The most apparent advances have been in surveys of microbial and viral diversity. Next-generation sequencing has permitted the now-commonplace exploration of fungal, bacterial, and viral diversity by generating 10^5 - 10^7 sequence reads per sample and using barcoding approaches (match of sequence to known taxonomic samples for that genomic region) to identify the taxa present and their relative abundance. While there is no doubt that this has transformed our understanding of functional ecosystem processes and microbial ecology at this scale (Nguyen et al. 2015; Turnbaugh et al. 2009; Desnues et al. 2008), there are definite limitations. For example, some taxa (e.g. Archaea) may not be as readily amplified using the same ribosomal 16S "bacteria" primers, and variation in amplification efficiency certainly exists within the Eubacteria (Acinas et al. 2005). Additionally, it is known that some bacterial genomes harbor more than one copy of this canonical locus (Kembel et al. 2012), thus muddling the relationship between read frequency and taxon frequency in a community.

The same problems exist - and are exacerbated - when studying multicellular diversity. On top of the problems of potential contamination, detecting rare taxa and/or handling singleton evidence for rare taxa, there is potentially a large variance in individual sizes of organisms. This, along with amplification variation given mismatches in the primer region, means that the relative read abundance in a NGS data set will often wildly vary (by multiple orders of magnitude) from the abundance of actual tissue in the data set (Nguyen et al. 2015; Piñol et al. 2014; Bohmann et al. 2014). Researchers tend to address this by analyzing data for simple incidence as well as relative read abundance, to identify patterns robust to either removal of information or inaccurate information (Nguyen et al. 2015).

If, however, the goal is to understand the actual relative abundance of individuals of different species in a sample - with these species harboring variation at 'barcode' loci, and often being highly divergent from one another - our question is whether there is complementary information that can be extracted from these data that does not rely on the abundance of reads that are assigned to a taxon, but relies on our understanding of diversity within populations and how that can be measured.

The summary statistics for DNA sequence diversity are well established and generally recognize the population mutation rate θ at a given locus; as a population increases in size, or as the mutation rate at that locus increases, more polymorphisms and more diversity will be found. There are limitations to this approach based on Kimura's neutral theory, as various forms of genomic selection will limit the direct relationship between population size and population diversity (e.g., Bazin, Glemin, and Galtier 2006; Wares 2010; Corbett-Detig, Hartl, and Sackton 2015). Nevertheless, these summary statistics - including Watterson's θ, a sample-normalized estimator of θ using the number of segregating sites S in a sample - may provide information necessary to generate *some* information about abundance patterns from NGS data. This information also certainly has its limits: nucleotide diversity (π) requires information on polymorphic site frequencies that will be biased by differential amplification across individuals, as well as relatively uninformative - or diminishing returns - as the number of sampled individuals increases (Wakeley 2008). Haplotype diversity (H) is likely sufficient to set a minimum boundary on the number of individuals sampled, and H along with S may carry enough information to generate a probabilistic distribution associated with larger numbers of individuals.

Here we present the mathematical considerations necessary to develop these quantitative tools to estimate the relative abundance of species in a barcoding sample of unknown individuals. The tools combine previous information on genetic diversity in the field population with observed properties of the sample, such as the number of haplotypes and the number of segregating sites for each species. We then evaluate the situations in which there is sufficient power to make meaningful statements about relative abundance from polymorphism data alone.

Methods

The approach here is identifying information that can comfortably be used as prior information to establish the posterior probability of observing polymorphism data from an *unknown* number of input individuals for a taxon. Any type of sampling information may help to set an upper limit: for example, if it is known that only 200 individual specimens were originally used for isolation of DNA, then the maximum number of total individuals inferred from this approach should be 200. This itself is not a numerical advance in biology, but limits our prior belief nonetheless.

There are also clear minimum bounds that can be established for the abundance of a taxon. Considering DNA sequence haplotypes as our most basic information, we ask how many *distinct* haplotypes are recovered in the data that match a particular taxon? For a haploid mitochondrial marker like the oft-applied cytochrome oxidase I (COI), this number is the minimum number of individuals present (if the number happens to be zero, it is also likely to be the maximum number of individuals in the sample!).

We suggest three methods that could help to estimate the number of individuals for a particular species in a metabarcoding sample: 1) an inference based on haplotype diversity of the field population and the observed *number of haplotypes* in the sample, 2) an inference based on the expectations of Ewens'sampling theory and the *number of haplotypes* observed in the sample; and 3) an inference based on a prior assumption of θ in the field population and the observed *number of segregating sites* in the sample .

To evaluate the potential usefulness of each method for recovering the abundance of input individuals, we simulated populations evolving under a Wright-Fisher neutral model. We performed the simulations with Hudson's *ms* program (Hudson 2002) using the *gap* (Zhao 2015) package in R (R Core Team 2015). We simulated 3 populations, using three different population mutation rates (θ of 2, 10, and 20). For each population we estimated haplotype diversity using the PopGenome (Pfeifer et al. 2014) package in R.

From the simulated populations we took "field samples" of different sizes (n), sampling without replacement. We replicate the sampling experiment 100 times, to be able to assess variation of sampling. For each replicate, we calculate the number of haplotypes and the number of segregating sites, which represent our observed values in the simulated samples. The sampling size, known to us from this design, is what we are going to predict using the reversed inferences described below for each method. All the analysis of the simulated populations was done in R (R Core Team 2015). Detailed information on simulations and R code is presented in the supplementary material (*for now in file "runningSimulations"*).

Haplotype Diversity

In addition to the simple number of haplotypes observed at a barcode marker, we may also attempt to estimate the number of individuals that harbored those haplotypes. Here, we assume that there is previous information on haplotype diversity (H) from the natural populations of the species (or distinguishable populations) that are present in the barcoding sample. The "haplotype diversity", H, defined by Nei and Tajima (1981) as

$$H = \frac{N}{N-1}(1 - \sum_{i=1} x_i^2)$$

represents the probability that sampling a new individual will result in sample of a new haplotype. N is the number of haplotypes, and x_{i} is the sample frequency of the i_{th} haplotype.

An example of how H could be used is shown below for a sample in which 10 distinct haplotypes are observed, and the *prior information about H* for a particular taxon is $H = 0.7$. In addition to assuming that prior information about the population is appropriately comparable, here we assume a minimum of 10 individuals, and that what we do not know can be modeled by a Gamma distribution with the shape defined by the reciprocal of haplotype diversity (so that low diversity provides little information, high diversity suggests that the number of individuals is closer to the observed number of haplotypes), and the rate is defined by the reciprocal of the number of haplotypes.

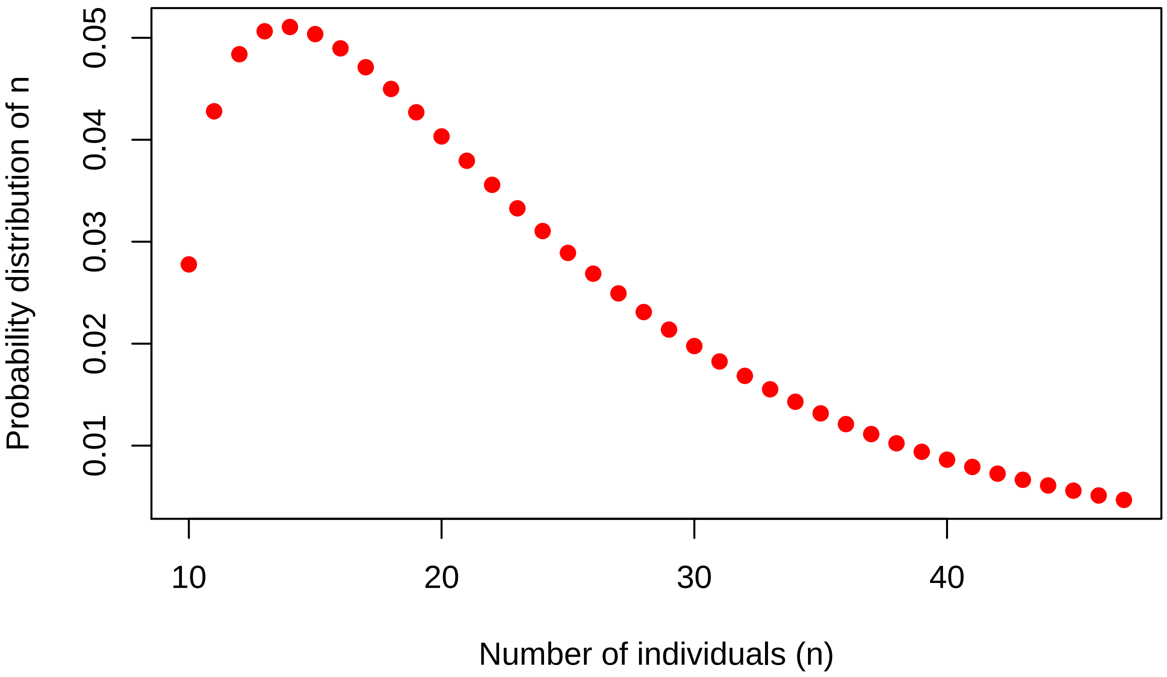


Figure 1. Probability distribution of observing n individuals in a sample in which 10 haplotypes are observed, and the haplotype diversity of the field population is $H = 0.7$.

So, observing 10 haplotypes for this taxon, and given a relatively haphazard use of the Gamma to obtain a useful probability shape given assumptions about how informative haplotype diversity is, we might feel comfortable believing (with a 95% interval) there are between 10 and 47 actual individuals that were sampled, with a highest likelihood solution of [**14 CHECK THIS DID YOU FIGURE OUT BETTER WAY FOR 95%?**]. A concern here lies in the willful abuse of the Gamma distribution without a better understanding of how haplotype diversity H and the sample size N may be actually related through the frequency of haplotypes - remember, at this point we are assuming we cannot trust the proportion/frequency representation of an allele in our sample.

For each of the 100 replicates in each sampling size within the three simulated populations we used the corresponding haplotype diversity for that population and the number of haplotypes observed in that replicate to estimate the probability distribution of the sampling size using a gamma function as defined above (shape defined by the reciprocal of haplotype diversity and the rate defined by the reciprocal of the number of haplotypes). From each probability distribution we recorded the sampling size with the highest probability (and the confindence intervals?) to compare with the experimental/simulated sampling sizes. Finally, we calculated the difference between the "real" sampling size -the one from our simulations- and the sampling size for the sample estimated using the gamma distribution method, as a measure of the precision of our method.

Sampling theory

Ewens (1972) developed a sampling theory of selectively neutral alleles, that based in the number of samples and the mutation parameter θ, allows one to estimate the expected number of different alleles (here, we address alleles from a haploid genome, i.e. haplotypes) in a sample. Assuming a sample of n individuals, the mean number of haplotypes in a sample can be approximated by:

$$E(h) = \frac{\theta}{\theta} + \frac{\theta}{\theta + 1} + ... + \frac{\theta}{\theta + n - 1}$$

where, h is the number of different haplotypes in the sample, n is the number of individuals in the sample, and θ: $4\mathrm{N_e u}$

If θ is very small, the expected number of haplotypes should be quite low, approaching 1. On the other hand, if θ is extremely large, the number of haplotypes should tend to n as noted above; of course there is a close relationship between Ewens' sampling theory and our understanding of H. Using this equation, we can estimate the distribution of the number of haplotypes for different sampling sizes, with a variance:

$$Var(h) = E(h) - [\frac{\theta^2}{\theta^2} + \frac{\theta^2}{(\theta + 1)^2} + ... + \frac{\theta^2}{(\theta + 2n - 1)^2}]$$

In general, the variance increases with θ for n of biological interest. Ewens' (1972) derivations rely on the assumption that the sample size is much lower than the actual population size. Considering this approach, rather than one based in haplotype diversity H, may allow us to avoid the problem of uncertain haplotype frequencies in an empirical data set.

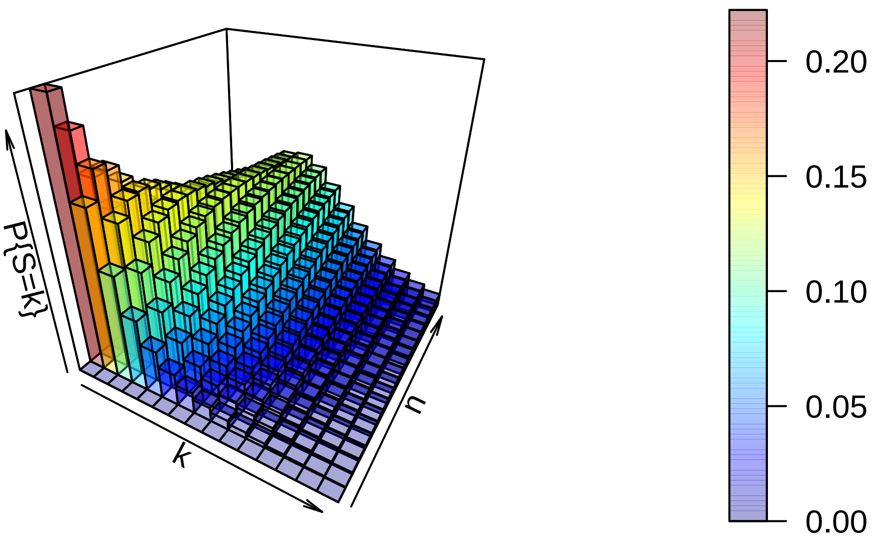
For each of the three populations with theta equal to 2, 10 or 20, and through the range of sampling sizes considered in this study (2 to 128) we applied Ewens (1972) formula to estimate the expected number of haplotypes (and the variance) for each sampling size. We then compared the observed number of haplotypes in each sample with the expected number of haplotypes by Ewens's formula for each sampling size and estimated the "observed" sampling size in our sample. The accuracy of the method was calculated as the difference between the "real" sampling size - the one from our simulations - and the number of individuals for the sample estimated using Ewens's method.

Segregating Sites

As noted above, there are specific probability distributions associated with a sample of sequences, the number of segregating sites S, and a prior assumption of θ under the standard coalescent model (Wakeley 2008).

$$P(S=k)=\sum_{i=2}^{n}(-1)^i\binom{n-1}{i-1}\frac{i-1}{\theta+i-1}(\frac{\theta}{\theta+i-1})^k$$

Figure 2a illustrates this distribution for θ=2. This represents a low-diversity population, and unless few segregating sites are observed there may be a broad range of sample sizes consistent with such an observation. Figure 2b illustrates the same probability distribution, but assuming θ=10. When the prior knowledge or assumption of diversity is higher, there tends to be a sharper distribution on n for a given k.



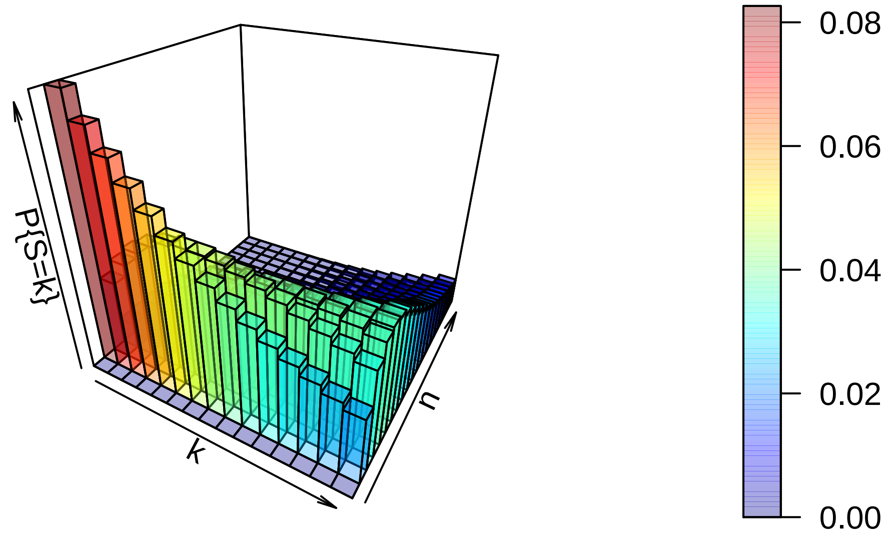


Figure 2. Probability surface of observing a number of segregating sites k for a given sample size n when θ is set. In (a), $\theta = 2$; in (b), $\theta = 10$.

Results

The summary statistics of the simulated population data are presented in Table 1. As expected, the haplotype diversity, number of haplotypes and number of segregating sites are higher as θ increases.

Table 1. Summary information on the simulated "field" populations that were used in this study.

Population	Theta	Haplotype diversity	Tajima's D	Number of haplotypes	Number of segregating sites
Population 2	2	0.34	0.26	11	12
Population 4	10	0.93	0.26	47	71
Population 6	20	0.96	1.13	79	164

Haplotype diversity and gamma estimation

Overall, using haplotype diversity and our educated guess at how this diversity reflects the input tends to greatly underestimate the simulated sample, at least for larger sampling size (Figure 3). For the smaller sampling sizes the difference between the predicted sampling size and the simulated sampling size is close to zero, meaning the method provides little error; the difference is also smaller when θ is larger (Figure 3).

The probability distributions using this method and the predicted sampling sizes against the observed haplotypes are provided in the supplementary material.

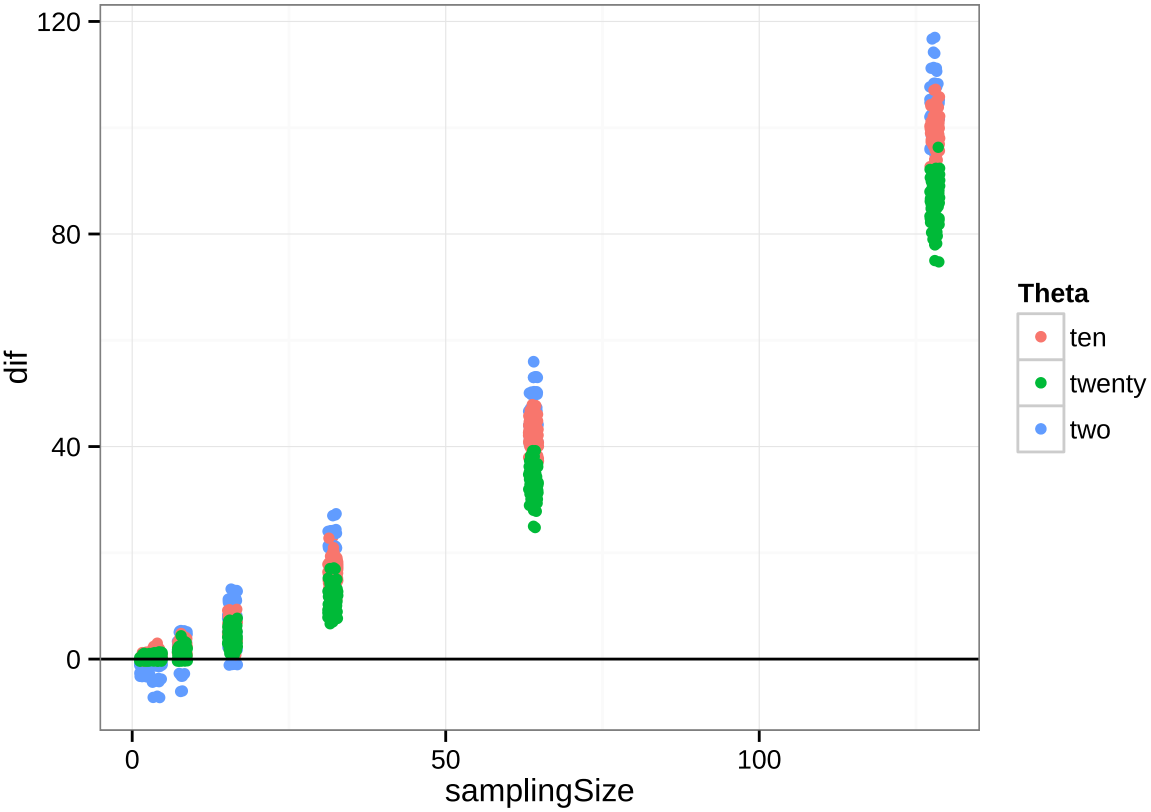


Figure 3. Difference between the simulated sampling size and the sampling size for the sample predicted using the gamma distribution method. The abline at zero shows the ideal situation with the estimation equal to the simulated sampling size.

Sampling theory

Using Ewens's sampling theory and backcalculating the sampling size from the number of haplotypes gives a difference between the simulated and predicted sampling size that is centered to zero. The efficacy of this method (the difference between observed and predicted sampling size) decreases with an increase in sampling size and increases is the original population has a larger theta (from two to twenty).

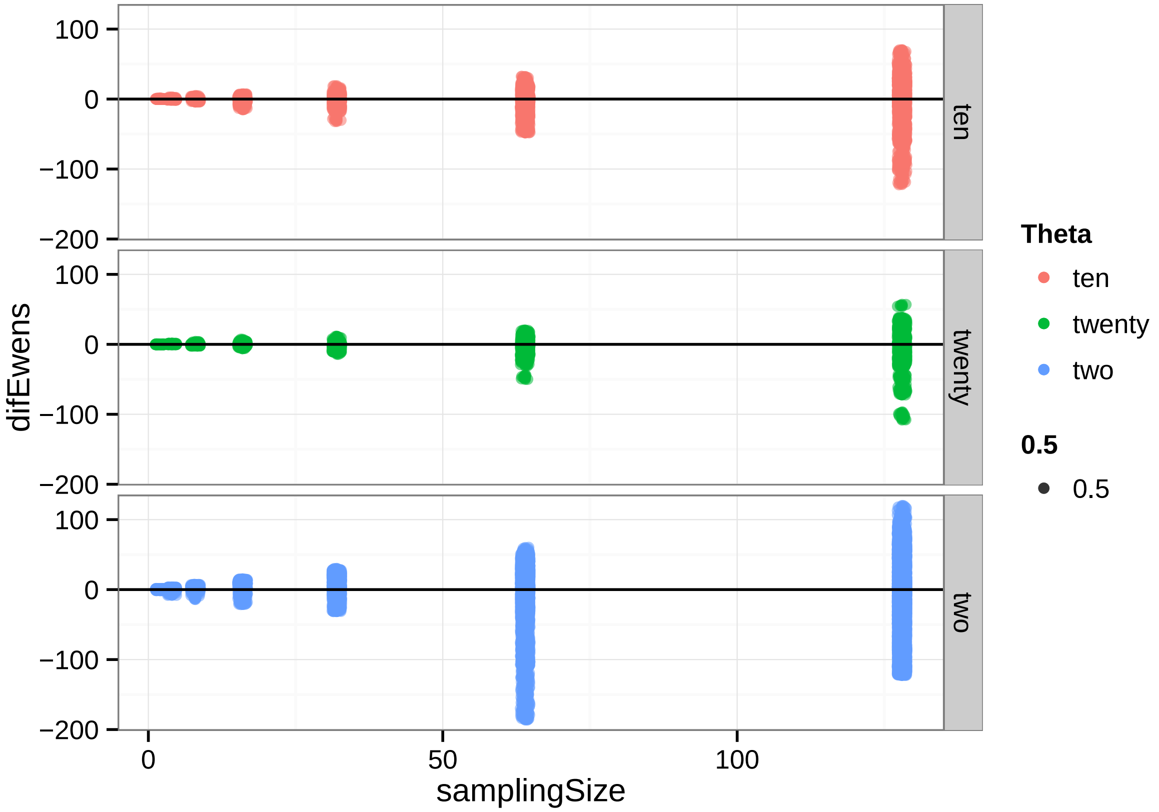
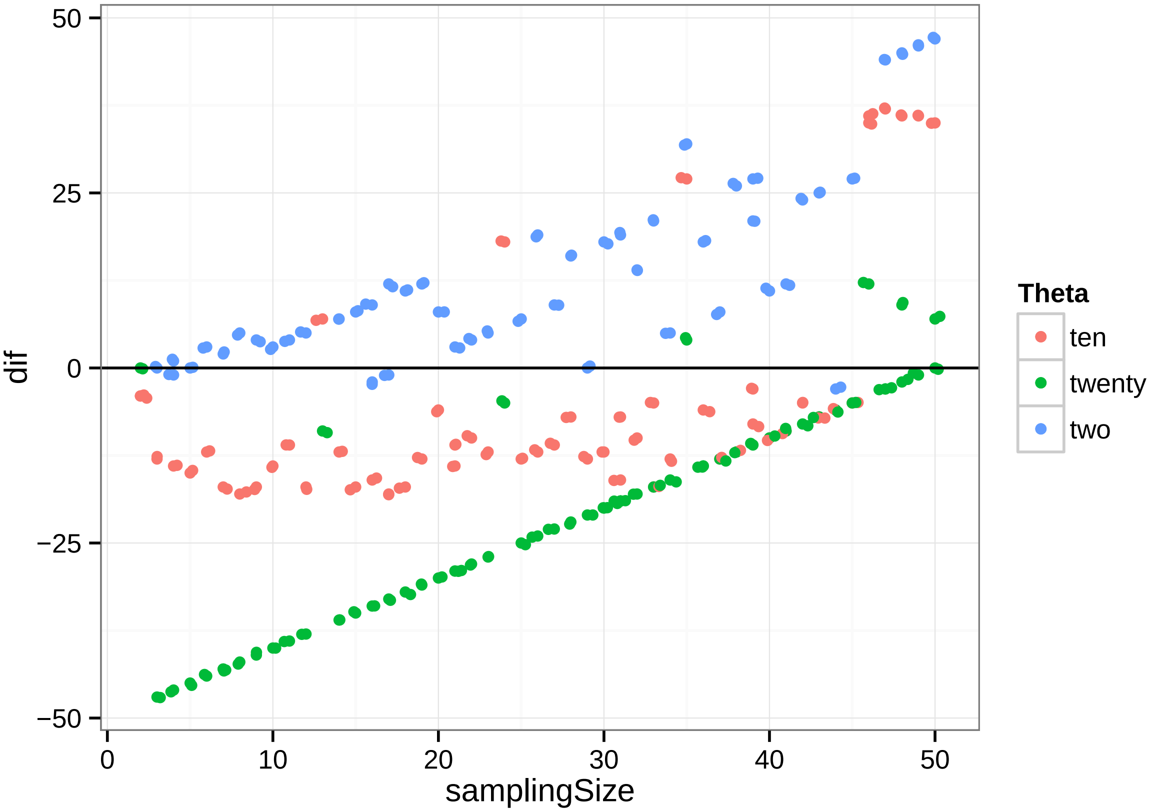


Figure 4. Difference between the simulated sampling size and the sampling size for the sample predicted using Ewens's sampling theory. The abline at zero shows the ideal situation in which the estimation equals the simulated sampling size. The panels separate the three different theta: two, ten and twenty.

Theta and segregating sites approach

The implementation of Wakeley's (2008) formula for estimation of the relationship between S, n, and θ was only tested for sampling sizes lower than 50, as we noted instabilities in estimation at larger sample sizes. Because of that, instead of using 2,4,8,16,32 as before, we did the full range from 2 to 50 sampling size.

John, my code finally worked but this figure looks funny. I will check again tonight...



Do we still use this? I cannot see PDF yet today.

Remember that the input data for this single taxon included `#r actual` individuals, the vertical line in plot above. What is likelihood function? Product of the two distributions? That is too stark in areas where they don't really overlap probabilities. Shouldn't be ZERO there?

Discussion

What we have shown is, in effect, the high variance in genealogical and mutational data associated with the coalescent process in population genetics (**???**). Though our early efforts suggested a broad utility in ranking the abundance of taxa in a mixed sample of metabarcode data, the result of our extended simulations indicate a preponderance of high-variance, downward-biased results in estimating the number of individuals in a sequence data set. In considering basic haplotype diversity H, the observed number of haplotypes, as well as the number of segregating sites S, our attempts to use genetic diversity tend to greatly underestimate the simulated sample of individuals, at least for larger sampling sizes and/or low values of θ. If the goal is to improve our ability to quantitatively describe the biodiversity in a system using metabarcoding approaches, we show that such an approach is of poor utility unless the diversity of the system is high and the number of individuals input to the metabarcoding analysis is modest. Given the additional uncertainty associated with assumptions of comparable diversity from prior evaluation of each population, there are no benefits in cost or estimation over traditional barcoding of individual specimens.

The statistics we evaluate are not independent from one another; they are pertain to the same genealogical process assumed to underlie a sample of DNA sequence data and are different ways of summarizing this coalescent process. Although some methods, e.g. approximate Bayesian computation, have been used to infer the demographic history of a sample of sequences using the aggregate of summary statistics available (e.g. Ilves et al. 2010), the relationship shown here appears to be too tenuous to make an advance in our ability to estimate relative abundance of taxa from such metabarcode data. Some of the error or bias in estimation we note from our simulation work reflects common problems in sampling data and exploring them with summary statistics. Felsenstein (1992) had noted that a high variance (in fact, driven by bimodal distribution of resultant statistics) in the number of segregating sites S would be expected with low sample size. Effectively, the comparison of a small number of sequences in a high-diversity system has a large probability of pairwise contrasts across the oldest node of a genealogy (Felsenstein 1992), and at very small sample sizes there is a potential that two closely-related sequences are sampled rather than reflect the TMRCA of the genealogy.

Additionally, our approach is predicated on the idea that prior analysis of a given population - a genetically discrete and relatively homogeneous evolutionary unit - will effectively suggest the diversity to be found in subsequent samples. There are certainly instances where the diversity at a barcode locus has been so extraordinarily high that haplotype diversity approached 1, and the number of haplotypes recovered in a sample was very close to the number of individuals in that sample, such as the barnacle *Balanus glandula* (**???**,(**???**),(**???**)). However, this same example of a hyperdiverse barnacle also requires recognition that there are at least 2 distinct evolutionary lineages in this taxon with broadly overlapping geographic ranges (**???**,(**???**)), which dramatically affects our understanding of the diversity recovered as well as the underlying genealogical process and association with regional diversity.

This leaves metabarcode research with three options: (1) continue to individually sequence using Sanger methods; (2) only use metabarcode data for presence/absence of a taxon; (3) in cases where the amplification bias may be considered negligible, as with closely-related lineages, the frequency of reads may be useful for approximating the *relative* but not absolute abundance of lineages in a sample. It should be noted that the problem we face - unknown input to the diversity observed - is a similar problem that biologists have handled studying species introductions (Wares et al. 2005), now exacerbated by the confounding issues of next-generation sequencing.

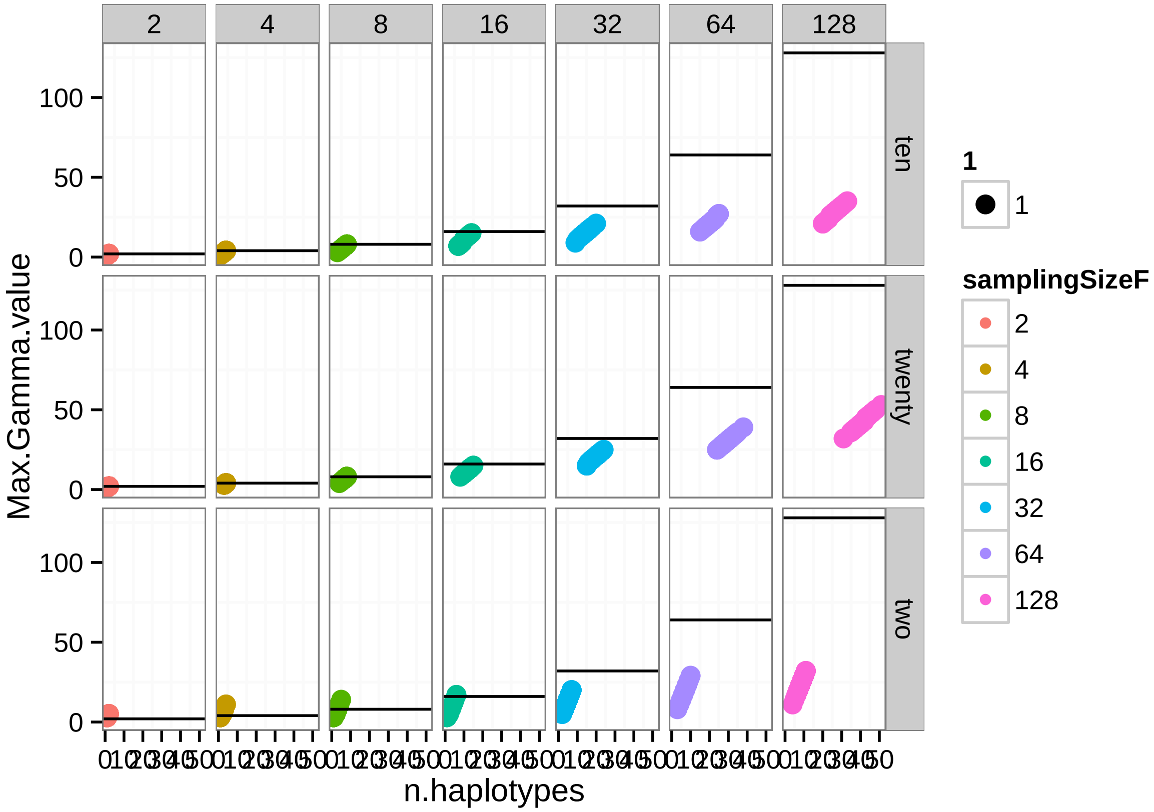
Though there are concerns about how well read/sequence frequency reflects the relative abundance of populations in an environmental sample - driven largely by differential amplification success of target genomes (Nguyen et al. 2015; **???**; Bohmann et al. 2014), it is worth noting that among high-θ populations there may still be comparisons appropriate in a relative sense: greater haplotypic diversity from a metabarcode sample would suggest more individuals of that species were in the sample. In this way we can evaluate order-of-magnitude results, and have less need for prior information from a population. As a complementary recognition that the number of haplotypes provides us with more information than simple presence/absence, we may start to improve on our ability to recover actual ecology from actual molecules.

Acknowledgments

Idea brought about by extended problem-solving session with J. Drake and colleagues in the Odum School of Ecology and Department of Genetics at The University of Georgia (UGA). Our work was improved greatly by suggestions from C. Ewers-Saucedo and K. Bockrath. This work was supported by funding from NSF (OCE-1029526) and the UGA Department of Genetics.

Figure captions

Supplementary material



Supplementary Figure 1. Predictions of sampling size using the gamma distribution method for each population (with thetas 2,10 and 20) against the observed number of haplotypes in the sample. The ablines in each panel represent the "real" sampling size of that sample.

Literature Cited

Acinas, S. G., R. Sarma-Rupavtarm, V. Klepac-Ceraj, and M. F. Polz. 2005. "PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S RRNA Clone Libraries Constructed from the Same Sample." Journal Article. *Appl. Environ. Microbiol.* 71: 8966–69.

Bazin, E., S. Glemin, and N. Galtier. 2006. "Population Size Does Not Influence Mitochondrial Genetic Diversity in Animals." Journal Article. *Science* 312 (5773): 570–72. <Go to ISI>://000237296700043.

Bohmann, K., A. Evans, M. T. P. Gilbert, G. R. Carvalho, S. Creer, M. Knapp, D. W. Yu, and M. de Bruyn. 2014. "Environmental DNA for Wildlife Biology and Biodiversity Monitoring." Journal Article. *Trends in Ecology & Evolution* 29 (6): 358–67. doi:Doi 10.1016/J.Tree.2014.04.003.

Bourlat, S. J., A. Borja, J. Gilbert, M. I. Taylor, N. Davies, S. B. Weisberg, J. F. Griffith, et al. 2013. "Genomics in Marine Monitoring: New Opportunities for Assessing Marine Health Status." Journal Article. *Marine Pollution Bulletin* 74 (1): 19–31. doi:Doi 10.1016/J.Marpolbul.2013.05.042.

Corbett-Detig, R. B., D. L. Hartl, and T. B. Sackton. 2015. "Natural Selection Constrains Neutral Diversity Across a Wide Range of Species." Journal Article. *PLoS Biol* 13 (4): e1002112. doi:10.1371/journal.pbio.1002112.

Desnues, C., B. Rodriguez-Brito, S. Rayhawk, S. Kelley, T. Tran, M. Haynes, H. Liu, et al. 2008. "Biodiversity and Biogeography of Phages in Modern Stromatolites and Thrombolites." Journal Article. *Nature* 452 (7185): 340–U5. doi:Doi 10.1038/Nature06735.

Ewens, W. J. 1972. "The Sampling Theory of Selectively Neutral Alleles." Journal Article. *Theor. Pop. Biol.* 3: 87–112.

Felsenstein, J. 1992. "Estimating Effective Population Size from Samples of Sequences: Inefficiency of Pairwise and Segregating Sites as Compared to Phylogenetic Estimates." Journal Article. *Genet Res* 59 (2): 139–47. http://www.ncbi.nlm.nih.gov/pubmed/1628818.

Hudson, R. R. 2002. "Generating Samples Under a Wright-Fisher Neutral Model of Genetic Variation." Journal Article. *Bioinformatics* 18 (2): 337–8. http://www.ncbi.nlm.nih.gov/pubmed/11847089.

Kembel, S. W., M. Wu, J. A. Eisen, and J. L. Green. 2012. "Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance." Journal Article. *PLoS Comput Biol* 8 (10): e1002743. doi:10.1371/journal.pcbi.1002743.

Nagendra, H. 2001. "Using Remote Sensing to Assess Biodiversity." Journal Article. *International Journal of Remote Sensing* 22 (12): 2377–2400. doi:Doi 10.1080/01431160117096.

Nguyen, N. H., D. Smith, K. Peay, and P. Kennedy. 2015. "Parsing Ecological Signal from Noise in Next Generation Amplicon Sequencing." Journal Article. *New Phytol* 205 (4): 1389–93. doi:10.1111/nph.12923.

Pfeifer, Bastian, Ulrich Wittelsbuerger, Sebastian E. Ramos-Onsins, and Martin J. Lercher. 2014. "PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R." *Molecular Biology and Evolution* 31: 1929–36. doi:10.1093/molbev/msu136.

Piñol, J., G. Mir, P. Gomez-Polo, and N. Agusti. 2014. "Universal and Blocking Primer Mismatches Limit the Use of High-Throughput DNA Sequencing for the Quantitative Metabarcoding of Arthropods." Journal Article. *Mol Ecol Resour*. doi:10.1111/1755-0998.12355.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, et al. 2009. "A Core Gut Microbiome in Obese and Lean Twins." Journal Article. *Nature* 457 (7228): 480–4. doi:nature07540 [pii] 10.1038/nature07540.

Wakeley, J. 2008. *Coalescent Theory: An Introduction*. Book. Greenwood Village, CO: Roberts & Co.

Wares, J. P. 2010. "Natural Distributions of Mitochondrial Sequence Diversity Support New Null Hypotheses." Journal Article. *Evolution* 64: 1136–42.

Zhao, J. H. 2015. *Gap: Genetic Analysis Package*. http://cran.r-project.org/package=gap.