

Article

Can theory improve the scope of quantitative metazoan metabarcoding?

John P. Wares ^{1,*} and Paula Pappalardo ^{2,†}

¹ Department of Genetics and Odum School of Ecology, University of Georgia, Athens, Georgia, USA 30602; jpwares@uga.edu

² Odum School of Ecology, University of Georgia, Athens, Georgia, USA 30602; paulapappalardo@gmail.com

[†] These authors contributed equally to this work.

* Author to whom correspondence should be addressed; E-Mail: jpwares@uga.edu; Tel.: +1-706-542-8000.

Academic Editor:

Received: / Accepted: / Published:

Abstract: Using second-generation sequencing approaches to quantify biodiversity has a number of hurdles, in particular that the number of reads for a given taxon may not be proportional to the number of individuals of that taxon in a sample. Here, we consider whether summary statistics generated in the course of population genetic analyses may be useful in reverse inference of the number of individuals input to an assay. Although our results show that these statistics – particularly the number of observed haplotypes given the overall level of diversity in the population – may be informative, there remain significant concerns about the ability to ‘metabarcoding’ a sample and infer relative species abundance.

Keywords: biodiversity; species diversity; population genetics; summary statistics; coalescent theory; metabarcoding

1. 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. A large error from our expectations is what suggests the low sample size from that distribution.

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 [1, 2], 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 (NGS) 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 [3-5], 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 [6]. Additionally, it is known that some bacterial genomes harbor more than one copy of this canonical locus [7], 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 [3, 8, 9]. 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 [3].

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, how that pertains to the effective population size, 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 (N), or as the mutation rate μ at that locus increases, more polymorphisms and more diversity will be found (θ is proportional to $N\mu$). 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

[10-12]. Nevertheless, these summary statistics - including Watterson's θ , a sample-normalized estimator that uses the number of segregating sites S in a sample - may provide information necessary to generate *some* inference of relative abundance from NGS data. This information also 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 [13]. Haplotype diversity (H) is likely sufficient to set a minimum boundary on the number of individuals sampled, and H along with S may provide enough information to generate a probabilistic distribution associated with larger numbers of individuals.

Here we present the mathematical considerations necessary to develop estimates of the relative abundance of species in a barcoding sample of unknown number of individuals. The tools combine previous information on genetic diversity in the sampled population (that is, a species or population that has previously been analyzed at the same locus) 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.

2. Experimental Section

Our approach is to identify information that can be used as prior information to establish the likelihood 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 (of all taxa) were originally used for isolation of DNA, then the maximum number of total individuals inferred from this approach should be 200 for any single included population. This itself is not a major advance in biology, but a limit on the inference 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 sample of genetic diversity, 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 prior estimates of haplotype diversity of a particular population and the observed *number of haplotypes* in a matched sample from that population, 2) an inference based on the expectations of Ewens' [14] sampling theory given a prior estimate of θ and the *number of haplotypes* observed in the sample; and 3) an inference based on a prior estimate 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 [15] using the *gap* [16] 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 then calculated summary statistics using the *PopGenome* [17] package in R.

From the simulated populations (single simulations of 1000 alleles given θ of 2, 10 and 20) we took "field samples" of different sizes ($n = 2, 4, 8, 16, 32, 64, 128$), sampling without replacement. We replicate the sampling experiment 100 times for each combination of θ and n , to be able to assess variance associated with the sampling effort. For each replicate, we calculate the number of haplotypes and the number of segregating sites, which represent our observed values in the simulated samples. From these distributions of observed values (haplotypes and segregating sites) in each combination of theta and sampling size we selected the values within the percentiles 0.25 and 0.75 to use in back calculations of the number of individuals in that sample. The distributions are presented in the Supplementary Material (Figures S2 and S3). The sampling size, known to us from this design, is what we attempt 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), with the exception of the estimation of predicted segregating sites according to Wakeley [13] that were performed in the software Mathematica (<https://www.wolfram.com/mathematica/>) given floating-point errors in R. Detailed information and the R code used to performed simulations is presented in the Appendix S1 in the Supplementary Material.

2.1. 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 [18] as

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

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

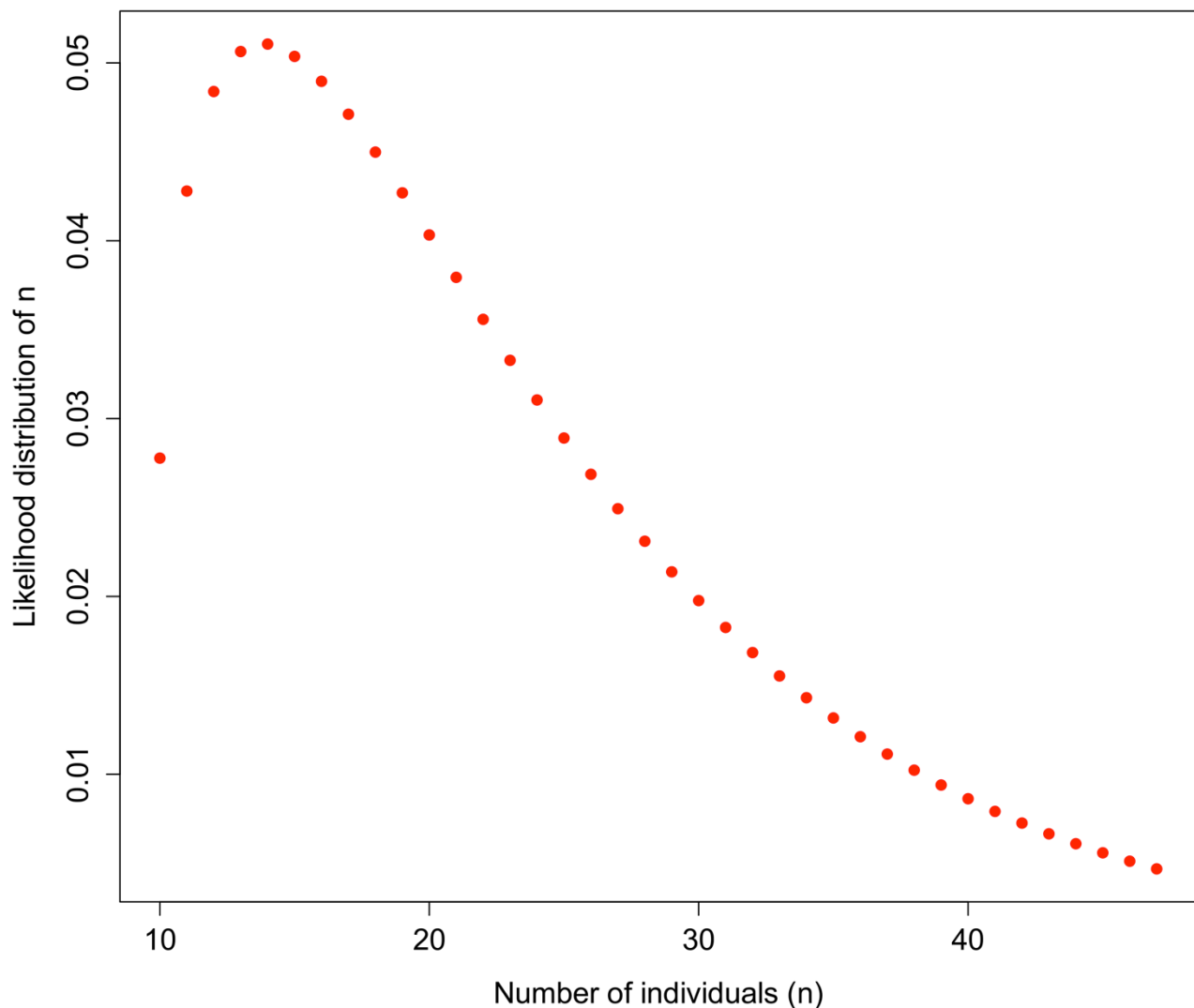


Figure 1. Likelihood of identifying n individuals in a sample in which 10 haplotypes are observed, and the haplotype diversity of the originally observed population is $H = 0.7$. Here, a gamma function is applied to represent the likelihood such that the distribution is flat at low values of H and a sharper distribution with high values of H , bounded by the actual observed number of haplotypes.

An example of how H could be used is shown in Figure 1 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.

So, observing 10 haplotypes for this taxon, and using the gamma to obtain a useful probability shape based on assumptions about how informative haplotype diversity is, we might feel comfortable believing 14 individuals were sampled (the highest likelihood solution). 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 likelihoods of the number of individuals in the sample (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 likelihood distribution we recorded the sampling size with the highest probability to compare with the simulated sampling size in each replicate. Finally, we calculated the difference between the "real" sampling size -the one from our simulations- and the sampling size inferred using the gamma distribution method, as a measure of the precision of our method.

2.2. Sampling Theory

Ewens [14] developed a sampling theory of selectively neutral alleles, that based on 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:

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

where, h is the number of different haplotypes in the sample, n is the number of individuals in the sample, and $\theta = 4N_e u$

If θ is very small, the expected number of haplotypes should be quite low regardless of the number of individuals sampled. 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:

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

In general, the variance increases with θ for n of biological interest. Ewens' 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' 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.

2.3. Segregating Sites

Under the standard coalescent model there are specific probability distributions associated with a sample of sequences, the number of segregating sites S , and a prior assumption of θ :

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

Figure 2a illustrates this distribution for $\theta=2$, segregating sites k from 1 to 20, and number of individuals n from 1 to 20. 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 $\theta=10$, with a range of both k and n from 1 to 50. When the prior knowledge or assumption of diversity is higher, the range of segregating sites that could give a more accurate inference of n is higher, with a sharper distribution on n for lower to intermediate k .

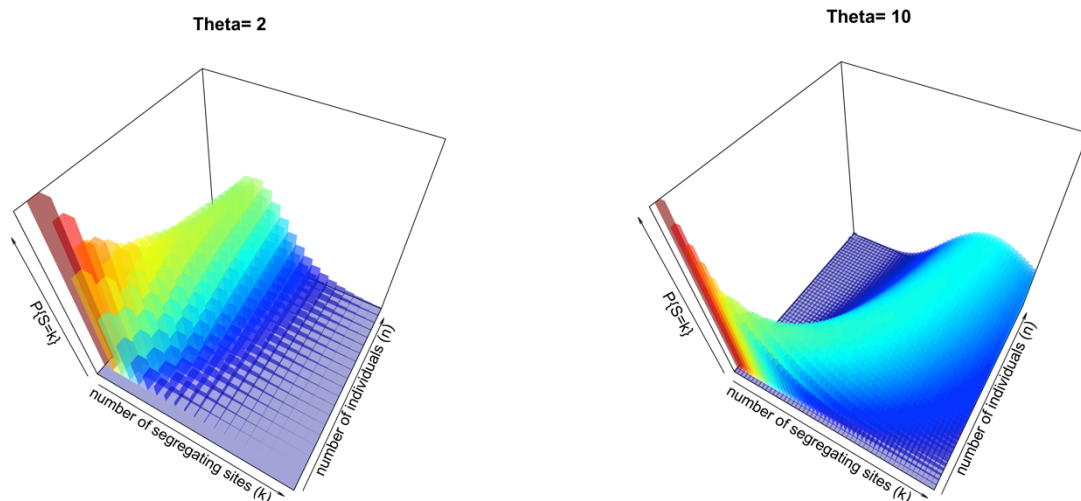


Figure 2. Probability surface of observing a number of segregating sites k for a given sample size n when θ is set. In (a), $\theta = 2$; in (b), $\theta = 10$.

For each of the three populations with theta equal to 2, 10 or 20, and with a range of sampling sizes from 2 to 128, we applied Wakeley's [13] equation 4.3 (presented above) to estimate the expected number of segregating sites for each sampling size. We then compared the observed number of segregating sites in each sample with the expected number of segregating sites 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 inferred number of individuals in the sample.

2.3. Comparing Methods

To summarize and compare the three methods we calculated the mean of the difference between simulated and predicted for each combination of θ and type of method.

3. Results and Discussion

The summary statistics of the simulated populations are presented in Table 1; the R code used to performed the simulations and estimate the summary statistics is presented in Appendix S1. As expected, the haplotype diversity, number of haplotypes and number of segregating sites increase with θ .

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

Population	Theta	Haplotype diversity	Number of haplotypes	Number of segregating sites
Population 1	2	0.34	11	12
Population 2	10	0.93	47	71
Population 3	20	0.96	79	164

3.1. Haplotype diversity and gamma estimation

Overall, using haplotype diversity (and our educated guess at how this diversity reflects the input) tends to underestimate the simulated sample (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 (Figure 3a). But for larger sampling sizes (Figure 3b) the predicted value greatly overestimates the number of individuals and the error is big. The efficiency of the method improves for larger θ , theta=20 has smaller differences between simulated and predicted sampling size (Figure 3). The probability distributions using the gamma method and the predicted sampling sizes for each combination of theta and sampling size are provided in the supplementary material (Figure S3).

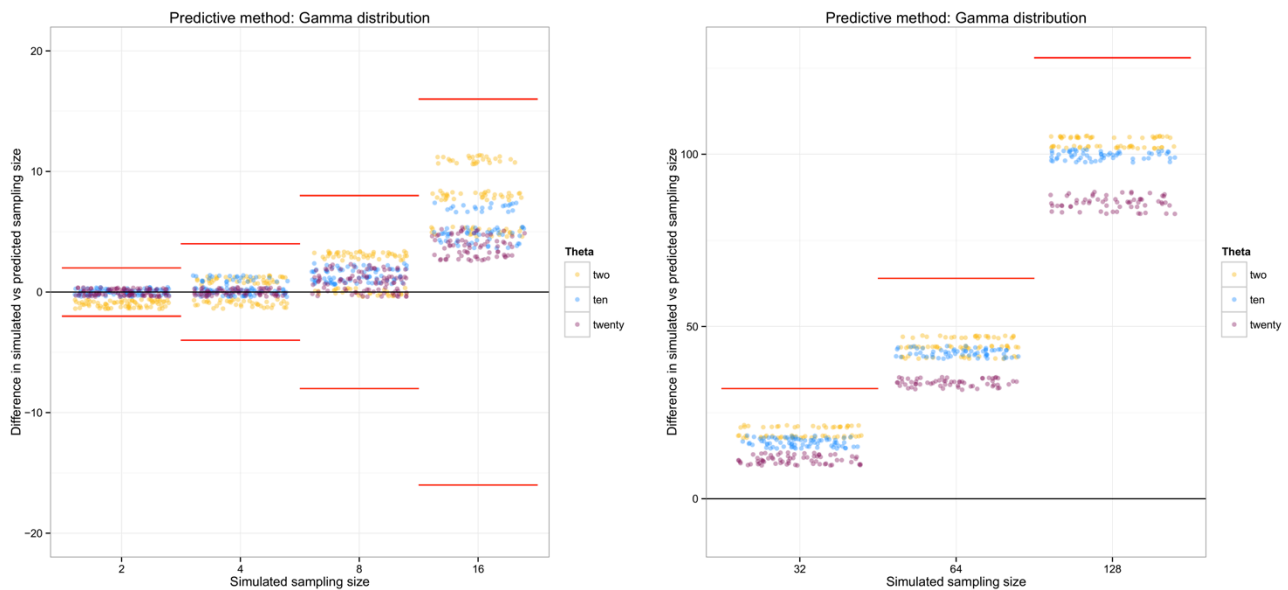


Figure 3. Difference between the simulated sampling size and the predicted sampling size using the gamma distribution method. (a) sampling sizes of 2, 4, 8 and 16; (b) sampling sizes of 32, 64 and 128. The horizontal line at zero shows the ideal situation with the estimation equal to the simulated sampling size. The red lines show the limit where the difference between simulated and predicted equals the sampling size; that is, beyond the red lines the error is too large to make a sensible prediction. Please note that the data are jittered to enhance visualization, and that the y-axis differs between (a) and (b).

3.2. Sampling theory

Using Ewens's sampling theory to estimate the number of individuals from the number of haplotypes gives a difference between the simulated and predicted sampling size that is centered on zero and that only has larger error for $\theta = 2$ (Figure 4). The precision of the method increases for larger values of θ .

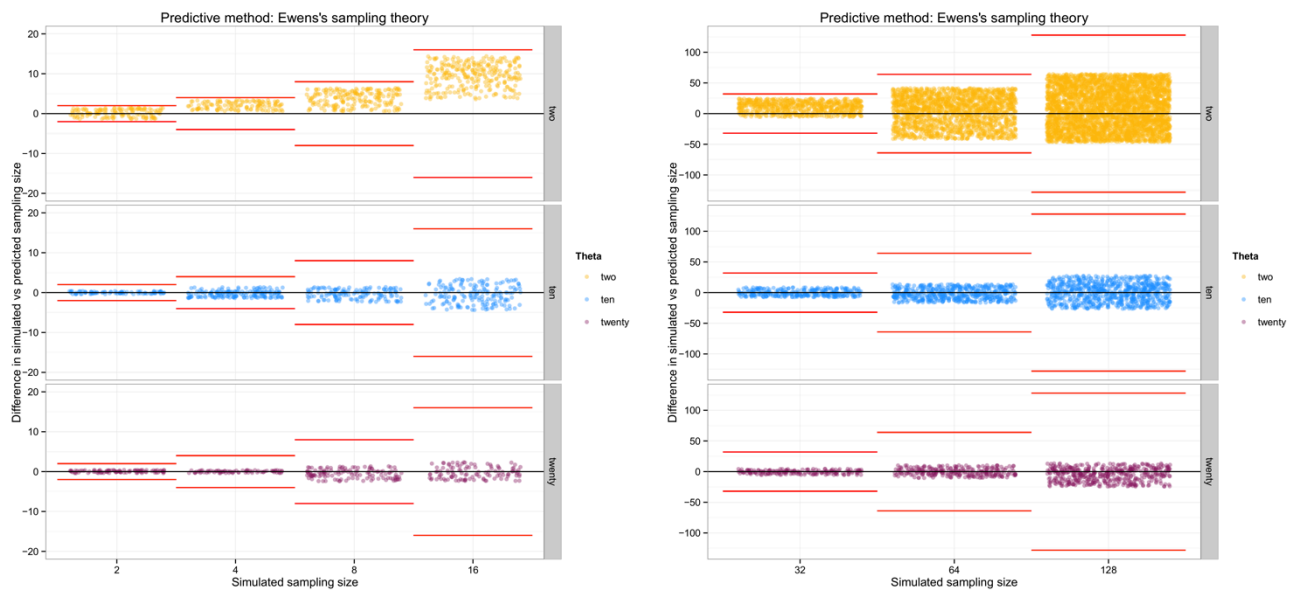


Figure 4. Difference between the simulated sampling size and the sampling size for the sample predicted using Ewens's sampling theory. (a) sampling sizes of 2, 4, 8 and 16; (b) sampling sizes of 32, 64 and 128. The panels separate the three different values of θ : two, ten and twenty. The horizontal line at zero shows the ideal situation in which the estimation equals the simulated sampling size. The red lines show the limit where the difference between simulated and predicted equals the sampling size; that is, beyond the red lines the error is too large to make a sensible prediction. Please note that the data are jittered to enhance visualization, and that the y-axis differs between (a) and (b).

3.3. *Theta and the segregating sites approach*

The implementation of Wakeley's formula for estimation of the relationship between the number of segregating sites k , the number of individuals n and θ presented the larger differences of precision between populations with different θ . The estimation of n for $\theta=20$ are totally flawed, with predicted values that overestimate the real n and errors that are far larger than the sampling size (Figure 5). The estimations of n tend to be underestimated in $\theta=2$ and overestimated in $\theta=10$, in particular for smaller sampling sizes (Figure 5a). The differences in the estimation between θ of 2 and 10 are reduced for larger sampling sizes (Figure 5b).

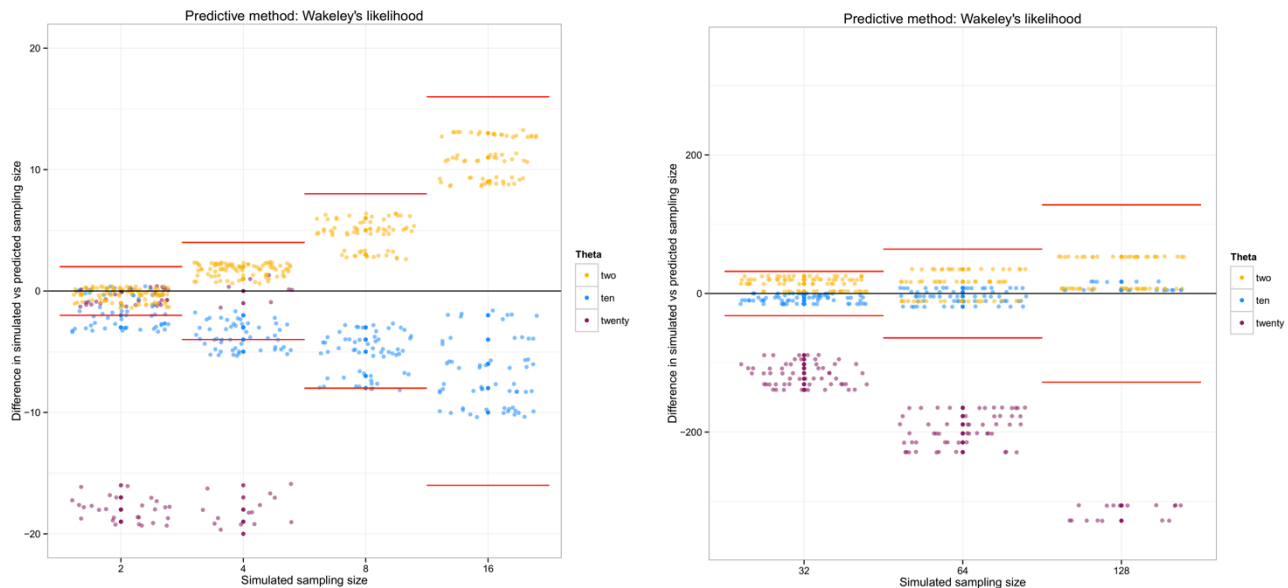


Figure 5. Difference between the simulated sampling size and the sampling size for the sample predicted using Wakeley's segregating sites theory. (a) sampling sizes of 2, 4, 8 and 16; (b) sampling sizes of 32, 64 and 128. The zero line shows the ideal situation in which the estimation equals the simulated sampling size. The red lines show the limit where the difference between simulated and predicted equals the sampling size; that is, beyond the red lines the error is too large to make a sensible prediction. Please note that the data are jittered to enhance visualization, and that the y-axis differs between (a) and (b).

3.4. Comparing methods

In Figure 6 we can see that the method of sampling theory is the most efficient, with the mean difference between simulated and predicted close to zero, particularly for θ of 10 and 20. The gamma method is only useful for small sampling sizes (2, 4 and 8), and the method of segregating sites works relatively fine for intermediate diversity levels, but less well when diversity is extremely high or low.

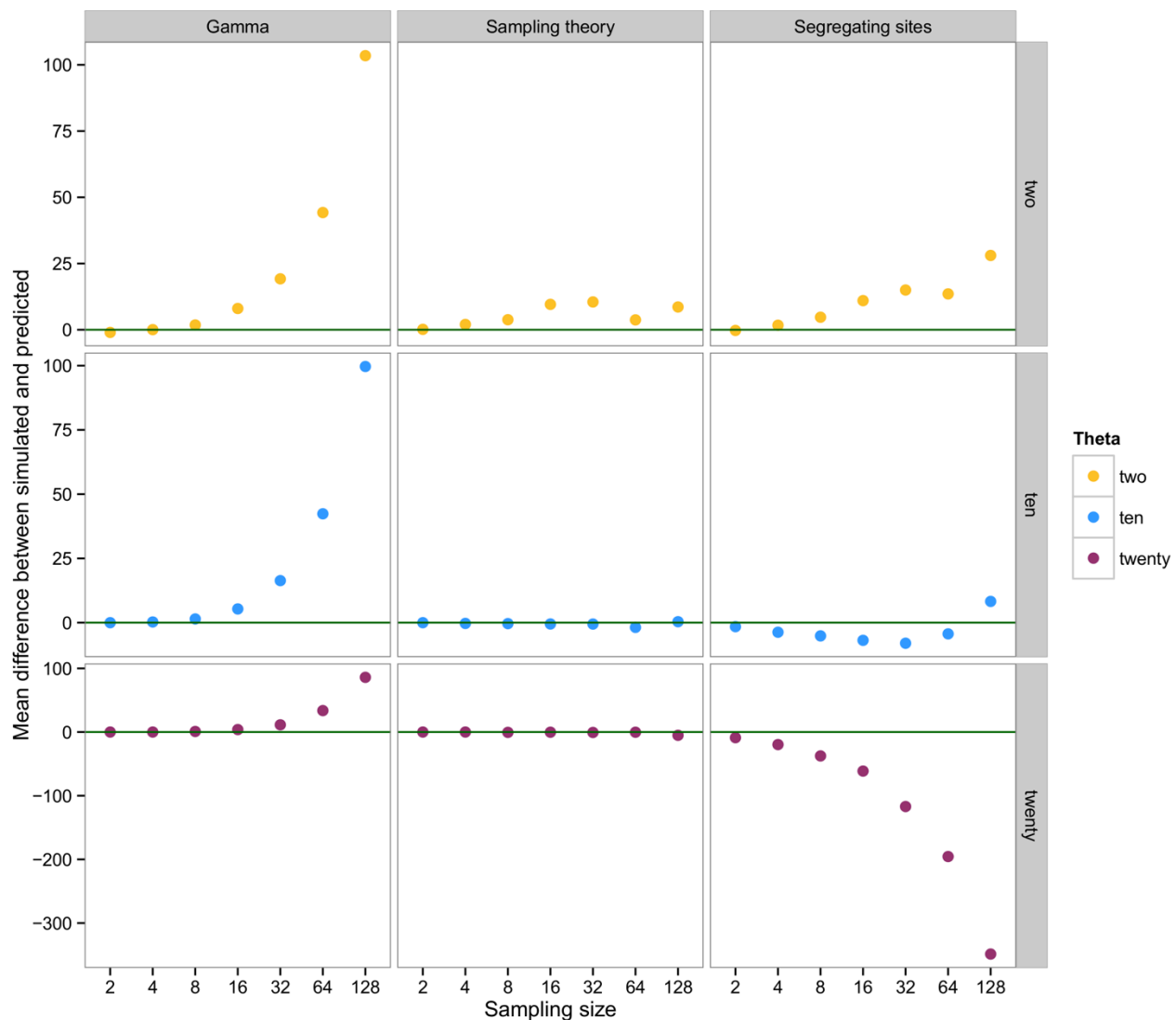


Figure 6. Mean difference between the simulated and the predicted sampling size for each combination of method (Ewens, Gamma, Wakeley) and θ (2, 10, 20). The zero line (in green) shows the ideal situation in which the estimation equals the simulated sampling size. Please note that the y-axis differ between θ panels.

4. Conclusions

What we have shown is, in effect, the high variance in genealogical and mutational data associated with the coalescent process in population genetics [19]. 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 , 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 . Fundamentally, this problem can be seen in Figure 2 for $\theta=10$; although there is an observable and relatively sharp distribution of probability relating k and n , the surface is quite wide with respects to n as k increases for a given θ .

The one component of these summary statistics that indicates an unbiased relationship between θ , n , and the number of haplotypes is, of course, a simple reflection of Ewens' [14] sampling theory. This suggests that *if* we already have a good understanding of θ for a particular population (for example if metabarcoding efforts are documenting the mix of individuals in a system that is already well-characterized for most species), then observing a number of haplotypes in a metabarcode sample allows some nominal indication of how many individuals are likely to have been included. Overall, however, if our goal is to improve the 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 (i.e., Sanger) barcoding of individual specimens.

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 [20] 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 [20], 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, the saturating relationship between sample size and observed diversity has been pointed out by Wakeley [13] as a feature of efficient estimation of θ from a natural sample; turning the crank on this equation backwards results in an inefficient inference of n from the other parameters.

Additionally, our approach is predicated on the idea that prior analysis of a given population - a genetically discrete and relatively homogeneous evolutionary unit, which can be difficult to delimit in and of itself - 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* [21-23]. 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 [23], which dramatically affects our understanding of the diversity recovered as well as the underlying genealogical process and association with regional diversity.

The statistics we evaluate are not independent from one another; they 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 [24], 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. It does seem 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. From the metabarcoding data themselves, each discrete population offers an estimate of θ and a number of observed, distinct haplotypes; this information is likely sufficient to bin abundances into more inclusive groupings of "common", "intermediate", and "rare" (e.g. a simplified Preston or Whittaker plot [25]).

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. 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 [3, 8, 9], it is worth noting that at least one recent study found a strong correlation between the proportion of metabarcoding reads from a taxon and the relative abundance of that taxon given point counts [26], as well as a strong correlation with taxon frequency from individual barcode data. If this relationship is supported in further studies, and if we can use information such as the number of haplotypes in a taxon as complementary information for minimum abundance, then 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 Ana Bento, Christine Ewers-Saucedo and Katie Bockrath. We kindly thank Toby Brett for his help to code in Mathematica. Annette Govindarajan and Stephen Spear gave us feedback that improved the quality of the manuscript. This work was supported by funding from NSF (OCE-1029526) and the UGA Department of Genetics.

Author Contributions

Initial idea and methodology developed by JPW; programming and additional contributions by PP. Both authors contributed equally to writing and interpretation of results.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

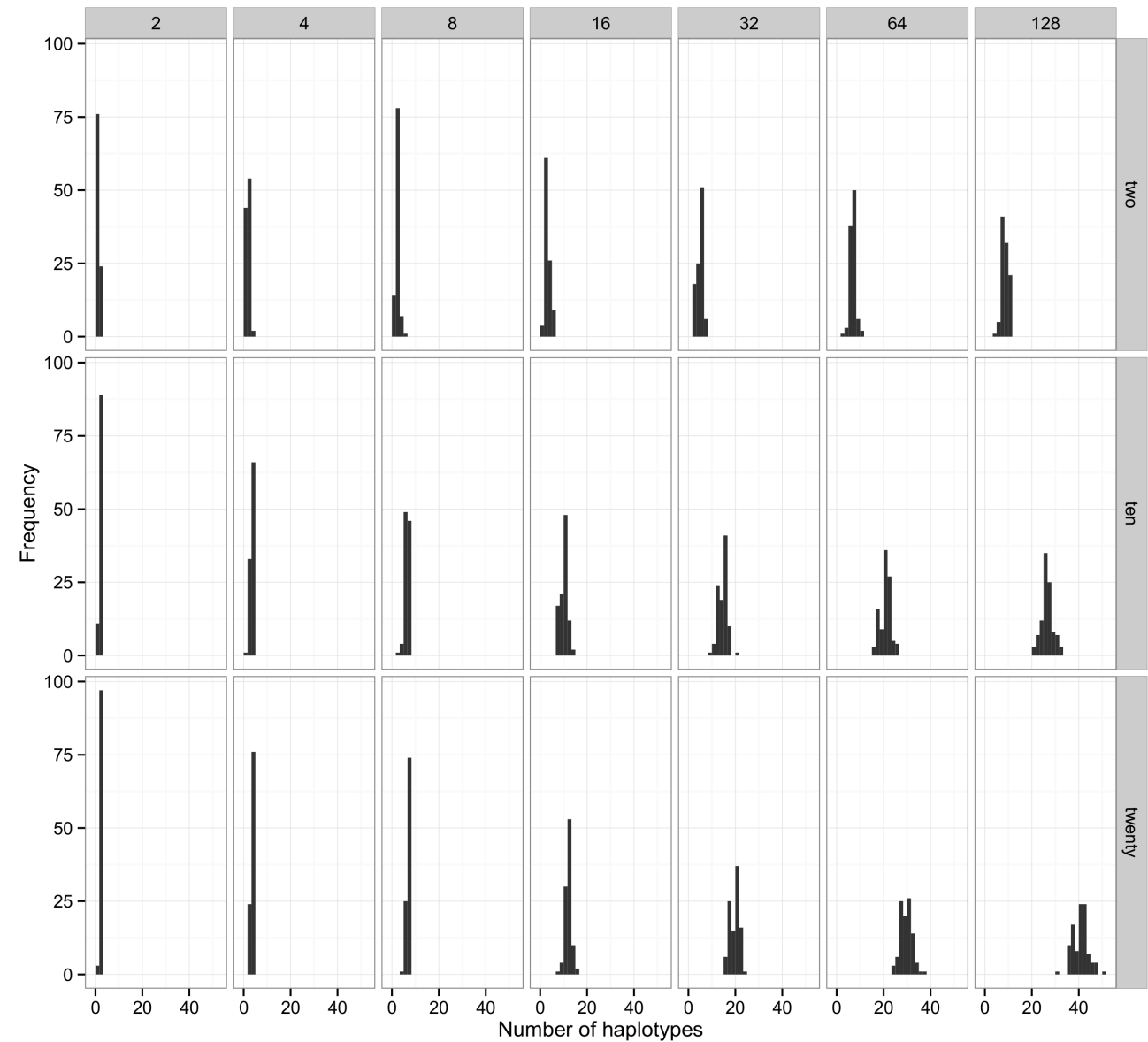
1. Bourlat, S.J., et al., *Genomics in marine monitoring: New opportunities for assessing marine health status*. Marine Pollution Bulletin, 2013. **74**(1): p. 19-31.
2. Nagendra, H., *Using remote sensing to assess biodiversity*. International Journal of Remote Sensing, 2001. **22**(12): p. 2377-2400.
3. Nguyen, N.H., et al., *Parsing ecological signal from noise in next generation amplicon sequencing*. New Phytol, 2015. **205**(4): p. 1389-93.
4. Turnbaugh, P.J., et al., *A core gut microbiome in obese and lean twins*. Nature, 2009. **457**(7228): p. 480-4.
5. Desnues, C., et al., *Biodiversity and biogeography of phages in modern stromatolites and thrombolites*. Nature, 2008. **452**(7185): p. 340-U5.
6. Acinas, S.G., et al., *PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample*. Appl. Environ. Microbiol., 2005. **71**: p. 8966-8969.
7. Kembel, S.W., et al., *Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance*. PLoS Comput Biol, 2012. **8**(10): p. e1002743.
8. Pinol, J., et al., *Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods*. Mol Ecol Resour, 2014.
9. Bohmann, K., et al., *Environmental DNA for wildlife biology and biodiversity monitoring* (vol 29, pg 358, 2014). Trends in Ecology & Evolution, 2014. **29**(8): p. 485-485.
10. Bazin, E., S. Glemin, and N. Galtier, *Population size does not influence mitochondrial genetic diversity in animals*. Science, 2006. **312**(5773): p. 570-572.
11. Wares, J.P., *Natural distributions of mitochondrial sequence diversity support new null hypotheses*. Evolution, 2010. **64**: p. 1136-1142.
12. Corbett-Detig, R.B., D.L. Hartl, and T.B. Sackton, *Natural Selection Constrains Neutral Diversity across A Wide Range of Species*. PLoS Biol, 2015. **13**(4): p. e1002112.
13. Wakeley, J., *Coalescent Theory: An Introduction*. 2008, Greenwood Village, CO: Roberts & Co.
14. Ewens, W.J., *The sampling theory of selectively neutral alleles*. Theor. Pop. Biol., 1972. **3**: p. 87-112.
15. Hudson, R.R., *Generating samples under a Wright-Fisher neutral model of genetic variation*. Bioinformatics, 2002. **18**(2): p. 337-8.
16. Zhao, J.H., *GAP: Genetics Analysis Package*. 2015.
17. Pfeifer, B., et al., *PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R*. Molecular Biology and Evolution, 2014. **31**(7): p. 1929-1936.
18. Nei, M. and F. Tajima, *Genetic Drift and Estimation of Effective Population-Size*. Genetics, 1981. **98**(3): p. 625-640.
19. Kingman, J.F.C., *The coalescent*. Stochastic Processes and their Applications, 1982. **13**: p. 235-248.
20. Felsenstein, J., *Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates*. Genetical Research, 1992. **59**: p. 139-147.
21. Wares, J.P., S.D. Gaines, and C.W. Cunningham, *A comparative study of asymmetric migration events across a marine biogeographic boundary*. Evolution, 2001. **55**(2): p. 295-306.
22. Wares, J.P. and C.W. Cunningham, *Diversification predates glaciation in Balanus glandula*. The Biological Bulletin, 2005. **208**: p. 60-68.
23. Sotka, E.E., et al., *Strong genetic clines and geographical variation in gene flow in the rocky intertidal barnacle Balanus glandula*. Molecular Ecology, 2004. **13**(8): p. 2143-2156.

24. Ilves, K., et al., *Colonization and/or mitochondrial selective sweeps across the North Atlantic intertidal assemblage revealed by multi-taxa approximate Bayesian computation*. Molecular Ecology, 2010. **19**: p. 4505-4519.
25. Magurran, A.E., *Measuring Biological Diversity*. 2004, Malden, MA: Blackwell Publishing.
26. Leray, M. and N. Knowlton, *DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity*. Proc Natl Acad Sci U S A, 2015. **112**(7): p. 2076-81.

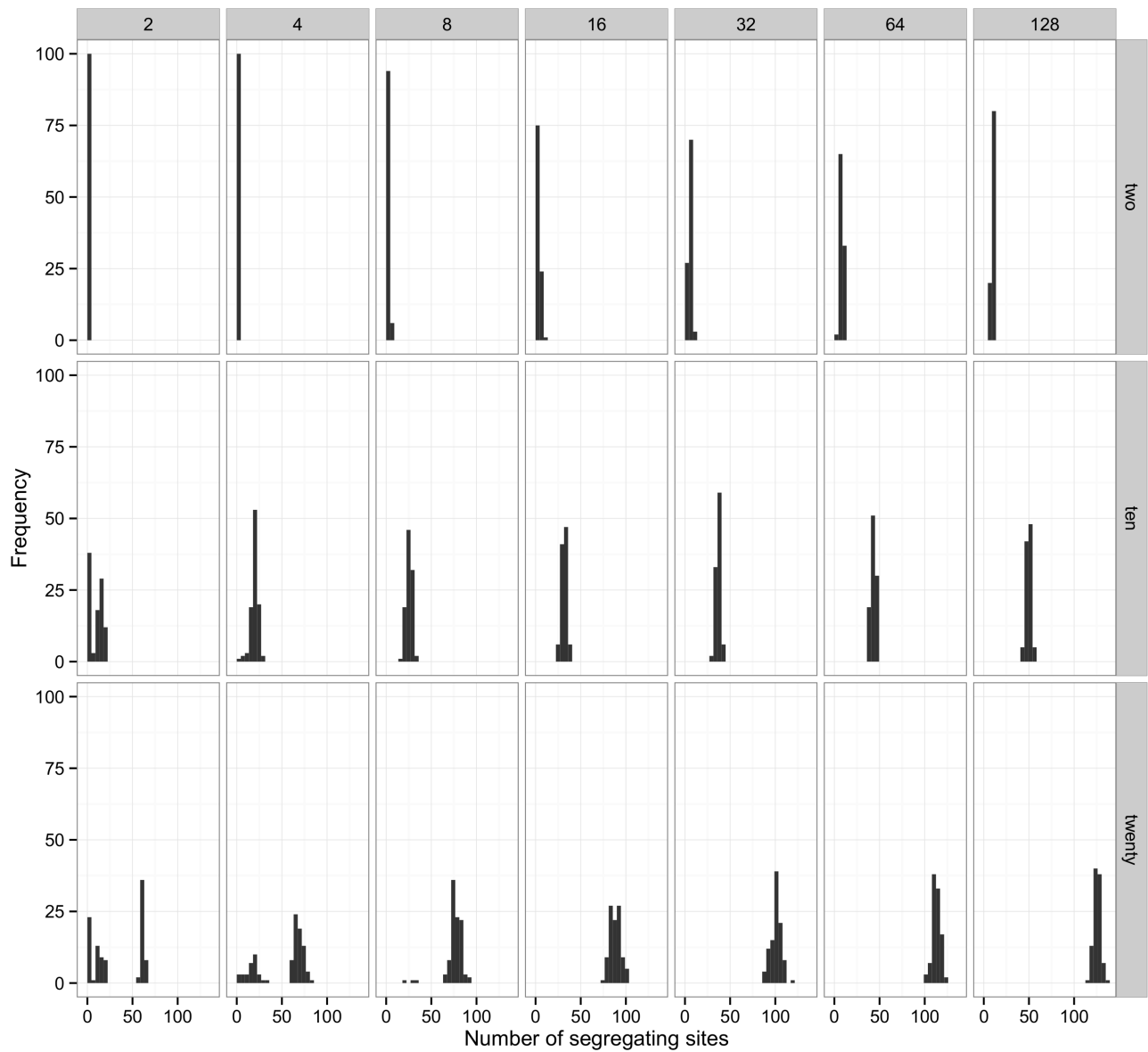
© 2015 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).

Supplementary Figures

Supplementary Figure 1. Distribution of the number of haplotypes for each combination of theta (two, ten, twenty) and sampling size (2, 4, 8, 16, 32, 64, 128).



Supplementary Figure 2. Distribution of the number of segregating sites for each combination of theta (two, ten, twenty) and sampling size (2, 4, 8, 16, 32, 64, 128).



Supplementary Figure 3. Probability distribution of the number of individuals back calculated using the gamma approach. Each panel shows a combination of simulated sampling size and the mutation rate in the original populations.

