Attempts at Quantitative Metazoan Metabarcoding are Difficult

John Wares and Paula Pappalardo

27 August, 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. Most notably, on top of the problems of potential contamination, detecting rare taxa and/or handling singleton evidence for rare taxa, and the potentially large variance in individual sizes of organisms, 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). This is caused primarily by shifts in amplification efficiency given mismatches in the primer region, and is often dealt with 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, our 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 - the 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 have some information about the probability associated with larger numbers of individuals.

Here we present the mathematical considerations necessary to develop these quantitative tools, and 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 0, it is also likely to be the maximum number of individuals in the sample!).

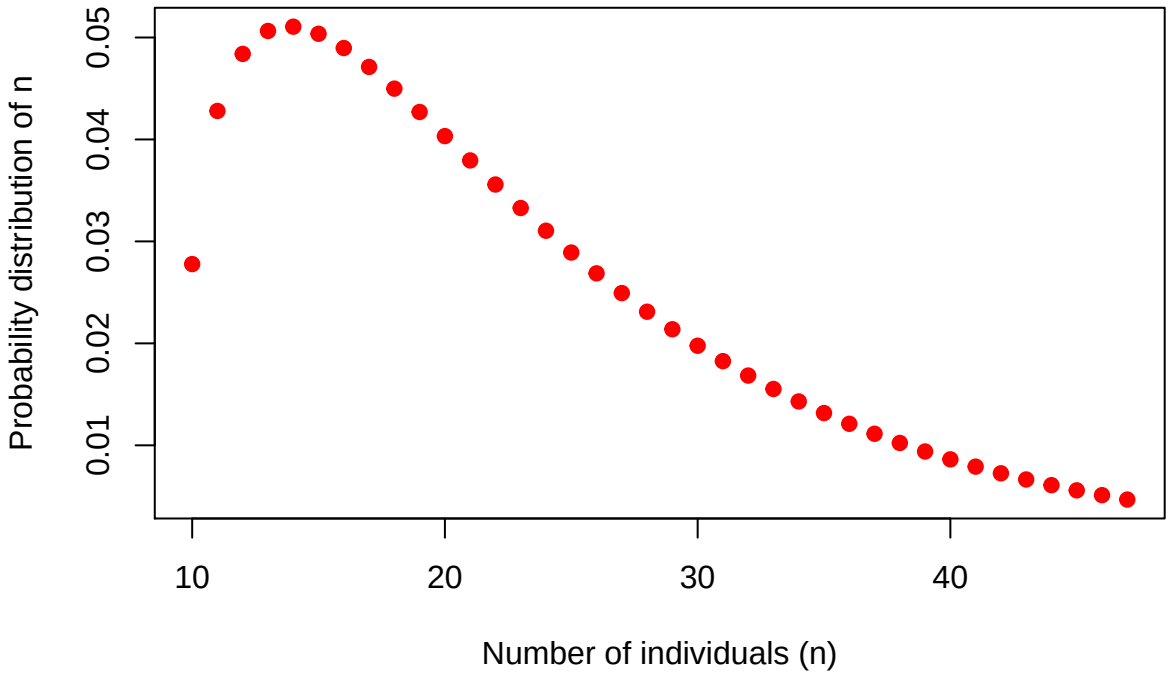
Haplotype Diversity

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 useful, 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 defined by the reciprocal of the number of haplotypes.



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. A problem lies in the willful abuse of the Gamma distribution without a better understanding of how haplotype diversity H and the sample size N are 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.

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: $4N_e u$

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 the relationship between Ewens' sampling theory and our understanding of H is close. 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 skirt around the problem of different haplotype frequencies in an empirical data set.

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 (Wakeley 2008).

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

Figure Xa 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 Xb 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.

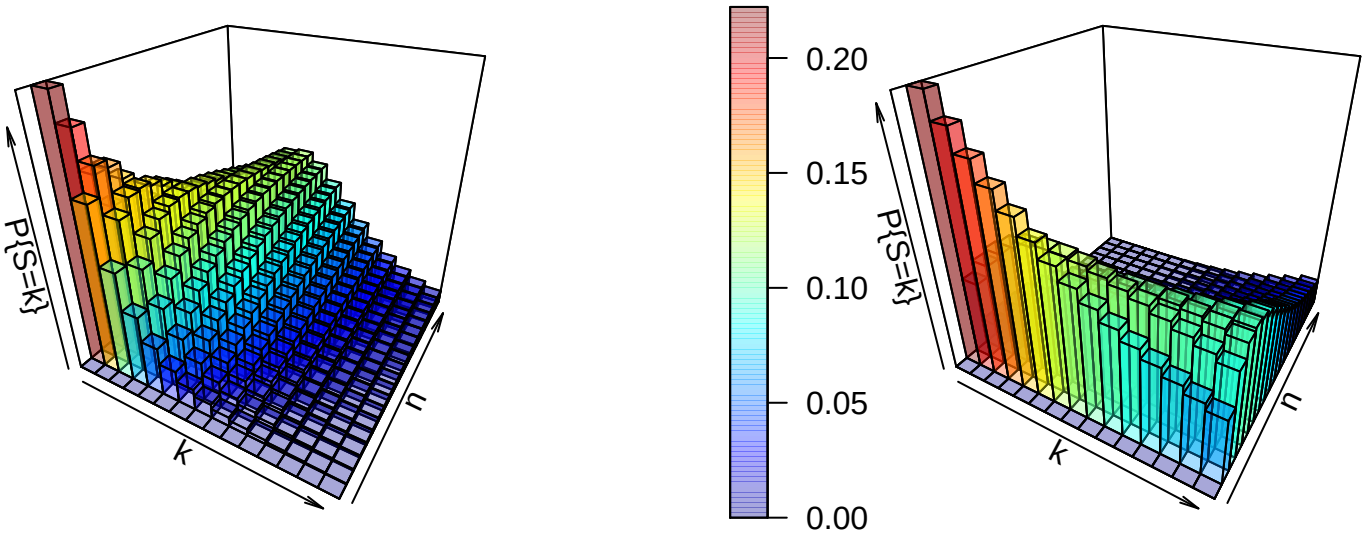


Figure 2. *Probability surface of observing a number of segregating sites* k* for a given sample size n when θ is set. In (a), $\theta = 2$; in (b), $\theta = 10$.

Evaluation of Inferences

We have suggested methods that could help to estimate the relative number of individuals in a metabarcoding sample: 1) an inference on the input sample size based on haplotype diversity, 2) inference of possible input sample sizes given the simple *number* of haplotypes observed based on Ewens' sampling theory; and 3) working from a prior assumption of /*theta* and the observed number of segregating sites in an empirical metabarcoding dataset.

To evaluate the potential usefulness of each method for recovering the abundance of input individuals, we simulated populations with Hudson's *ms* program (Hudson 2002) using the *gap* package in R (**???**). We simulated 6 populations, using three different population mutation rates (/*theta* of 2, 10, and 20), and also considering populations with or without growth. The growth rate was modified in simulations to target a Tajima's D value (-1.0) similar to those observed in empirical data for marine invertebrates (Wares 2010). *perhaps we take growth out of methods as well*

From these six 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 try to predict using the reversed inferences described above. All analysis of the simulated populations was done

in R (**???**). Detailed information on simulations and R code is presented in the supplementary material (*for now in file "runningSimulations"*).

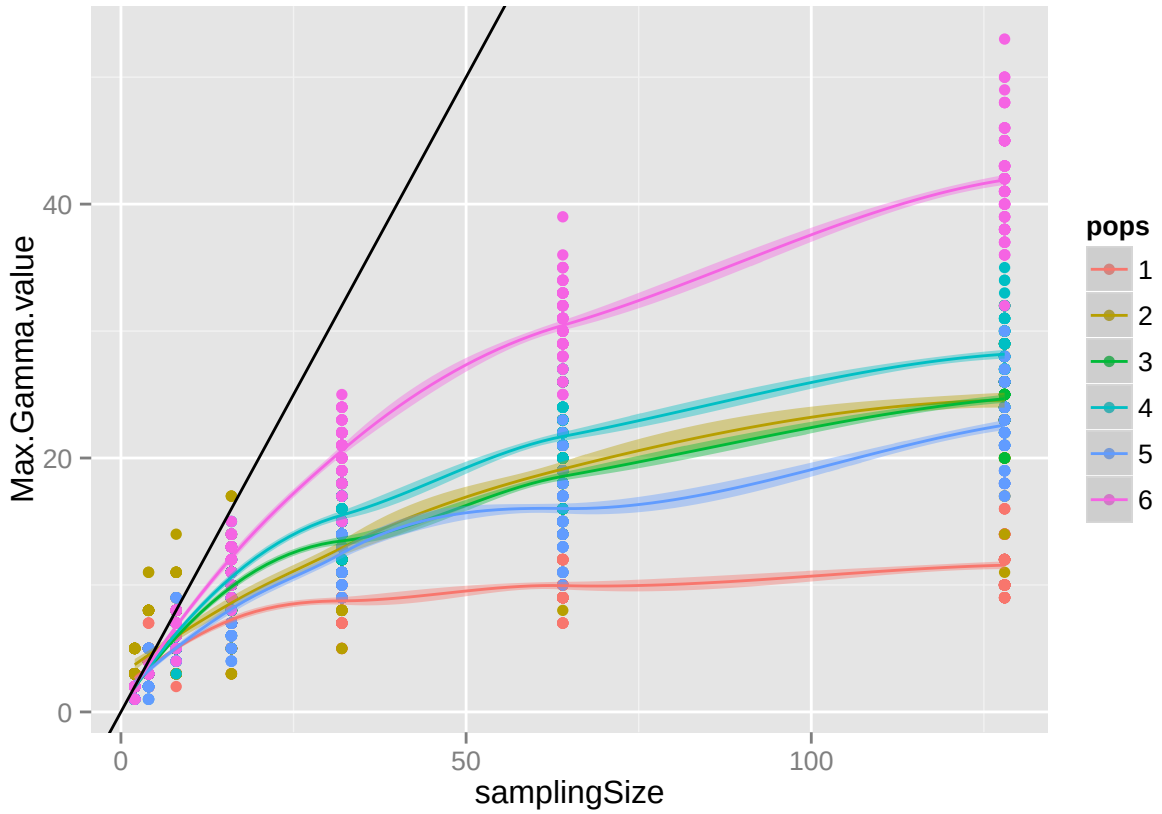
Results

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. Populations with simulated growth generate data consistent with the typical signature of purifying selection and/or demographic change seen in many metazoan populations (**???**).

Haplotype diversity and gamma estimation

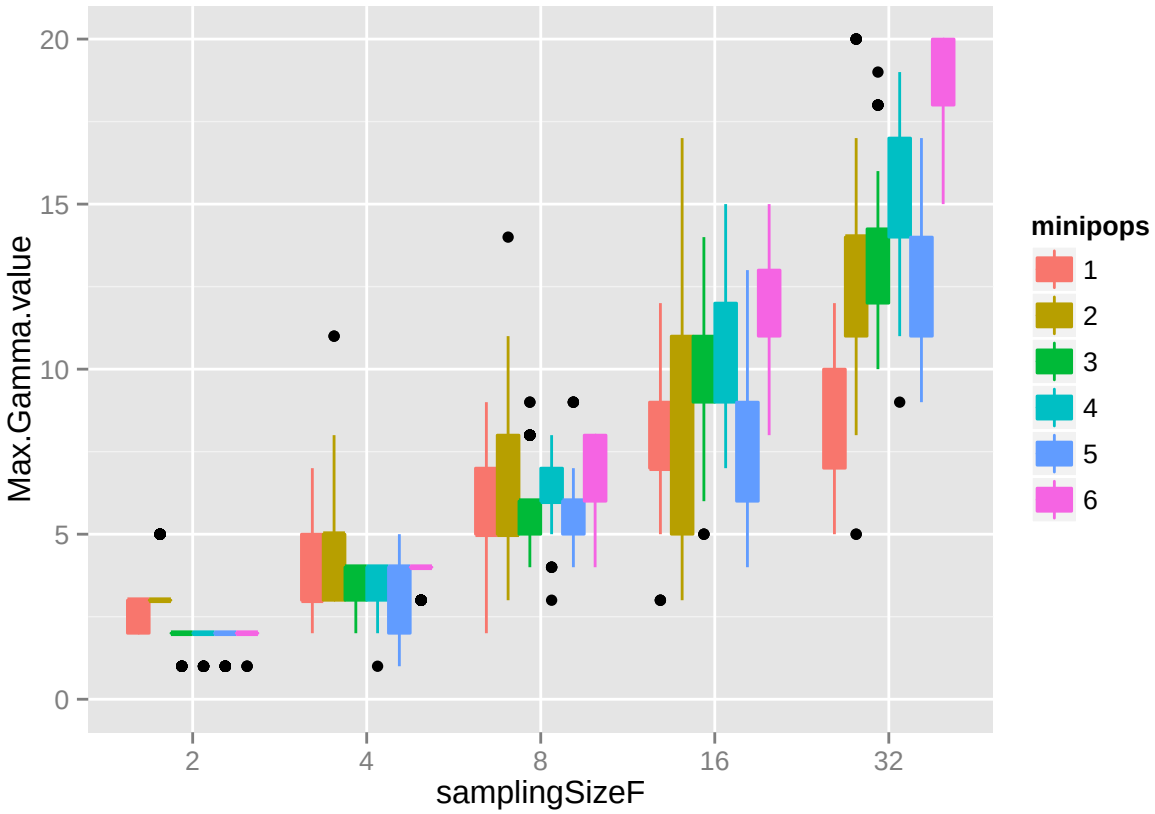
We found that the approach using a Gamma distribution to estimate the number of input individuals does a better job of estimating the simulated sampling sizes in low input sample sizes (Figure X2a, lower than 16-32 depending the population) and when θ is larger (θ=10 or 20, Figure X2b). The effect of growth - intended to reflect the typical mutational diversity in metazoan populations - varied with θ (Figure X2b). Overall, using haplotype diversity and an educated guess at how this diversity reflects the input tends to greatly underestimate the simulated sample (Figure X2a). The probability distributions for the six populations and the 7 sampling sizes considered (2,4,8,16,32,64 and 128) are presented in Supplement X.

```
#install.packages("ggplot2")
library(ggplot2)

# now we have the dataframe "gammaData" with all the values we need to plot
#gammaData<-read.csv("gammaData.csv",header=T)
gammaData<-read.csv("gammaData_13Jul.csv",header=T) #7/31 is this correct to change this file name? wou

# subsetting small sampling sizes to see better what happen there
minigamma<-gammaData[gammaData$samplingSize<33,]

# making factors for the ggplot2 pltos
factor(gammaData$Pop)->gammaData$pops
factor(minigamma$Pop)->minigamma$minipops
factor(minigamma$samplingSize)->minigamma$samplingSizeF

# plotting with ggplot, need to fix legends
#http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/
# Fig X2a
ggplot(gammaData, aes(x = samplingSize,y=Max.Gamma.value,fill=pops,color=pops)) + geom_point()+stat_smo
```




```
# Fig X2b
# thinking how to plot the subset of samples sizes where gamma works a little bit better....
ggplot(minigamma, aes(x = samplingSizeF,y=Max.Gamma.value,fill=minipops,color=minipops)) + geom_boxplot
```

```
## Warning: Removed 55 rows containing non-finite values (stat_boxplot).
```



Haplotype number

Paula here put plots of Ewens analytical results?

Theta and segregating sites approach

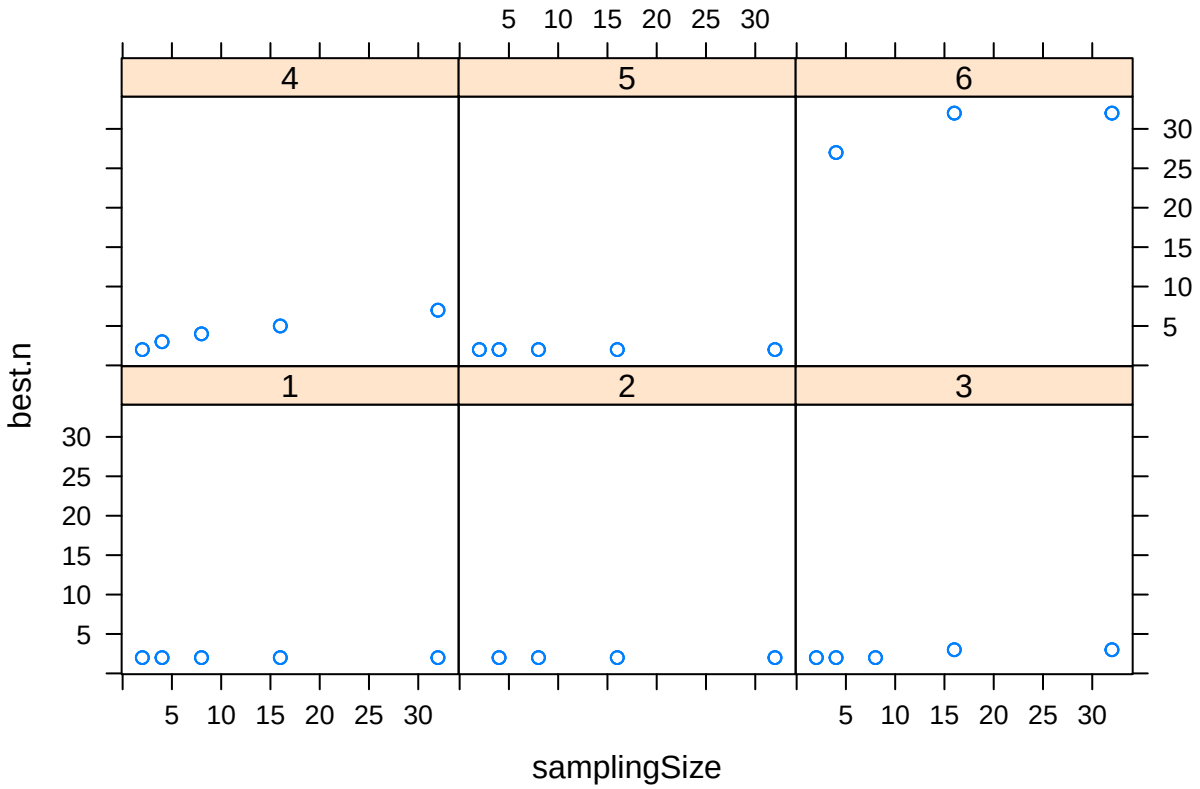
The implementation of Wakeley's (2008) formula crushed for n larger than 50, so we were only able to test our sampling sizes of 2,4,8,16 and 32 with this method.

```
#install.packages("vioplot")

library(lattice)
library(plyr)
library(vioplot)
```

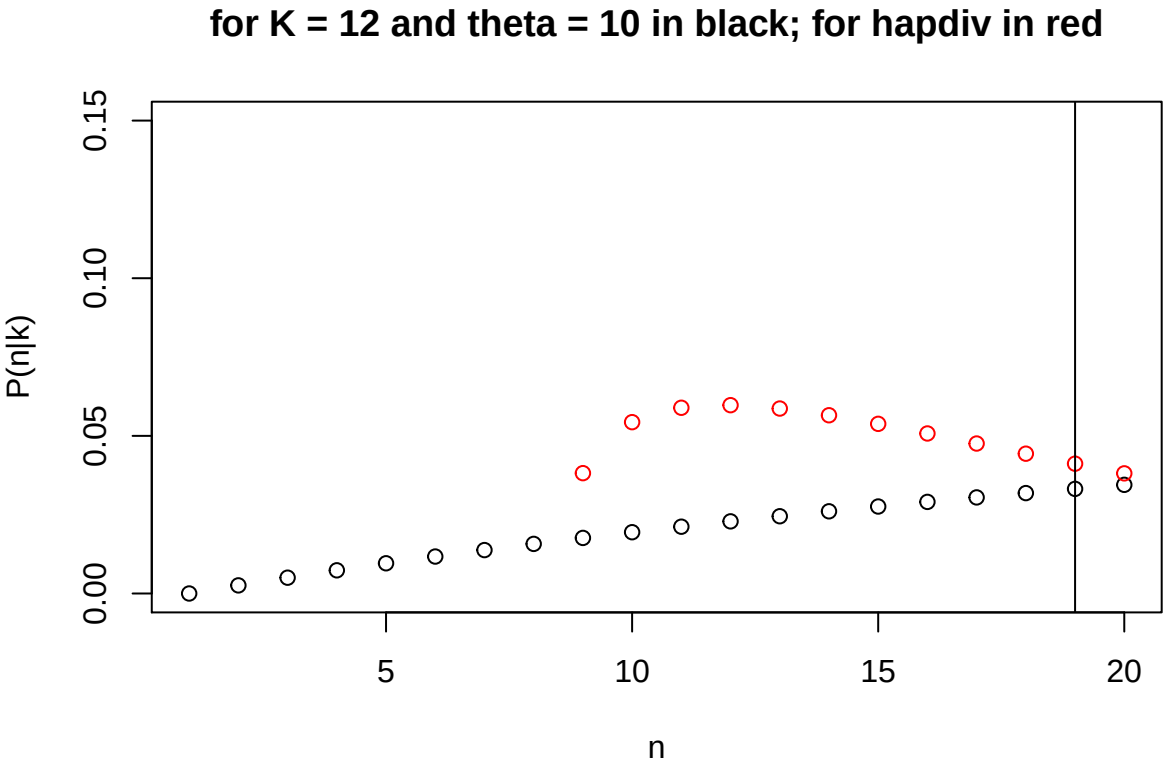
```
## Loading required package: sm
## Package 'sm', version 2.2-5.4: type help(sm) for summary information
```

```
# Load Wakeley data and observed segregating sites
obsdata<-read.csv("numberSegSites.csv",header=T)
wakdata<-read.csv("WakeleyData.csv",header=T)

# Summarize the info on observed segregating sites
infoObs<-ddply(obsdata, c("Pop","samplingSize"), summarise, medianOb = median(n.seg.sites,na.rm=T),meanO
```

```r
# keep only the n<=32
miniObs<-infoObs[which(infoObs$samplingSize<33),]

# Summarize Wakeley's info
miniwak<-wakdata[which(wakdata$n<33),]

#----make loop to find the best n's-----
thetas<-c("two","ten","twenty")
a<-c(rep("two",135),rep("ten",135),rep("twenty",135))
b<-c(seq(1,135,1),seq(1,135,1),seq(1,135,1))
goodwak<-as.data.frame(cbind(a,b))
names(goodwak)<-c("theta","seg.sites")

for(i in thetas){
  for(j in 1:nrow(goodwak)){ #Paula this wasn't working should 'mydata' be 'goodwak'??
    goodwak$seg.sites[j]->mysegsite
    mysubset<-subset(miniwak,miniwak$Theta==i & miniwak$segSites==mysegsite)
    max(mysubset$prob)->maxprob
    which(mysubset$prob==maxprob)->best
    mysubset$n[best]->best.n
    best.n->goodwak$best.n[j]
    }
}
# now "goodwak" has theta,seg.sites and best.n

# Now we want to merge the observed segregating sites with the segregating sites used to estimate Wakel
bymedian<-merge(miniObs,goodwak,by.x="medianOb",by.y="seg.sites",all.x=T)
bymedian[complete.cases(bymedian$best.n),]->bymedian1
bymedian1$Pop<-factor(bymedian1$Pop)
xyplot(best.n~samplingSize|Pop,data=bymedian1)
```



John's part below:



Remember that the input data for this single taxon included 19 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?

Need to then adapt this to a sample from 4-5 species for which there is known information? Or do simulated data (X species, vector of theta in simulation and hapdiv when all said and done, they are after all, related) and adjust the evenness in series of plots to see if evenness/richness gets recovered appropriately, given confidence intervals after all...

n.b. Marc Feldman was concerned about the two statistics double-dipping on the same theory... is this ABC, is this borrowing strength, is this inappropriate?? **AH. but certainly *S* and *H* are not independent observations thus cannot simply use product of the two to generate a Likelihood** perhaps identify which is better, useful, or are neither sufficient?

Discussion

High variance (bimodality) in number of segregating sites with low sample size follows predictions of Felsenstein (1992), who suggests that phylogenetic approaches to genealogical diversity likely to be more efficient than pairwise and segregating sites methods. Effectively, the comparison of a small number of sequences in a high-diversity system has a large probability of comparisons across the deepest node of a genealogy, but with very small samples also the potential that two closely-related sequences are sampled.

5-3-15 we are now pretty sure that the S=k method is too insensitive (as often the case iwth coalescent, and we see that above samples of 10 the values are relatively insensitive, see Vince Buffalo's "dancing genealogies" for illustration of why even if peaks look sharper for theta 10 the possibilities are massive for the empirical stuff that fits....). So this is concerning for even being able to tell orders of magnitude apart, e.g. 10 of species A and 100 of species B probably won't happen.

So new structure of paper we think: show that haplotypes are the defined minimum number of individuals, this is the dumb but easy and factual way. Then the gamma distribution, which is a shot in the dark, makes sense but parameterization is hard to think about. Then the 'smart' coalescent way, but we figure out it is not very sensitive.

And then part B of Results would be that we show empirical values of #haplotypes for samples of n given theta, identify whether this helps us parameterize the slightly-smarter approach. We may also ponder the mathematical relationship between theta and hapdiv, but of course for a given theta (truth) there is a range of theta (estimated from empirical) and a range of hapdiv (estimated from empirical)

Returning to the coin flip, it is worth evaluating wherein lies the strength of inferential signal. 50% gives NO information, could be 2 or infinite flips. So it is deviation that is signal in the coin flip example. Similarly, for a system of diversity such as this we need the *potential* for diverse outcomes to evaluate: low theta means nearly all sample sizes are possible, for example. In this sense, developing this with a mind for species that are broadly distributed and highly abundant is likely a more effective strategy than endemic small populations.

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) exacerbated by the confounding issues of next-generation sequencing.

Talk a bit about how barcode frequency information maybe isn't as far off when dealing with closely related taxa? We aren't throuwing out frequency, goal here is to look solely at complementary information.

In the end, Discussion :: though there are concerns about read frequency... at a minimum haplotpye number bounds the minimum. In this way we feel better about order of magnitude results, and less need for prior information from a population. This has been less evaluated in microbial/viral samples as "population" is perhaps less defined in those communi9ties than in eukaryote or metazoan communiteis, e.g. through gene flow. But as a complementary recognition that the number of haplotypes tells us some information for sure, and prior information about that population may also provide additional information, we may start to improve on our ability to recover actual ecology from actual molecules.

5-3-15 we are now pretty sure that the S=k method is too insensitive (as often the case iwth coalescent, and we see that above samples of 10 the values are relatively insensitive, see Vince Buffalo's "dancing genealogies"

for illustration of why even if peaks look sharper for theta 10 the possibilities are massive for the empirical stuff that fits. . . .). So this is concerning for even being able to tell orders of magnitude apart, e.g. 10 of species A and 100 of species B probably won't happen.

An important note is that we need to be confident that our data from the field population (the haplotype diversity or theta) is representing a "true" population, and it is not including different populations

another thing to consider with prior knowledge: presumably different species amplify with different efficiency, but WITHIN a population - well, perhaps you can evaluate (a) the frequency of haplotypes, and thus (b) the site frequency spectrum - in other words, you could calculate haplotype diversity as well and see how well it matches prior estimates, and also you could then assume that things like Taj D could be calculated? Could that perform any better at getting you back to actual-n? Probably not.

Acknowledgments

Idea brought about by extended problem-solving session with J. Drake, helped greatly by C. Ewers-Saucedo and K. Bockrath. Work supported by funding from NSF-OCE-Chile, OVPR, and UGA Department of Genetics.

Figure captions

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.

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.

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.