Quantitative Metazoan Metabarcoding

John Wares and Paula Pappalardo

06 April, 2015

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 (presumably 50%) and the number of trials.

Here, we argue that the same principle can 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 (remote sensing forest study), 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 - 10^6 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 at microbial ecology at this scale (Nguyen et al. 2015)(Moran, Rohwer, Knight refs), 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 standard 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). 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 both 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). 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) 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

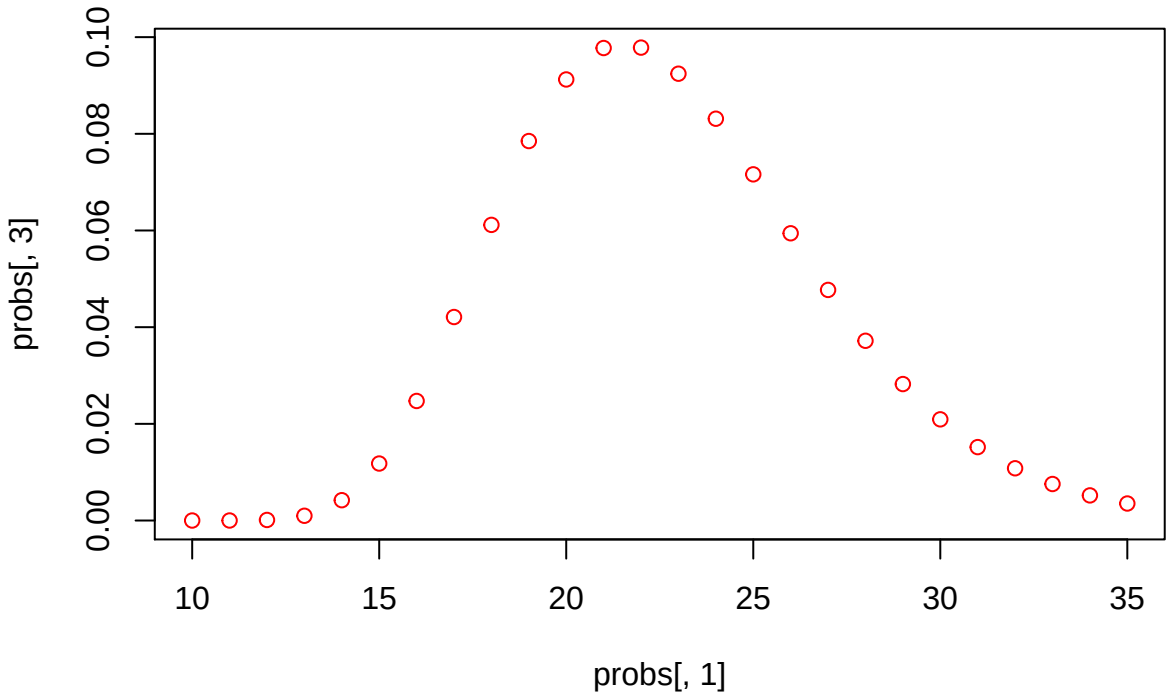
The approach here is identifying information that can comfortably be used as prior information to establish the posterior probability of observing polymorphism data from an *unknown* number of input individuals for a taxon. Some information is clearly limiting: for example, if it is known that only about 200 individual specimens were originally used for isolation of DNA, then the maximum number of total individuals recovered from this approach should be about 200. This is perhaps not exciting 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-used 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). The number of haplotypes recovered from a sample of a species can be summarized with "haplotype diversity", H, defined by Nei and Tajima (1981) as

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

and representing the probability that sampling a new individual will result in sample of a new haplotype.

An example of how H could be used is shown below for a sample in which 10 haplotypes are observed, *from a population or taxon with prior information about H*, here with $H = 0.75$. 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 number of haplotypes and scale defined by our prior knowledge of H.



So, observing 10 haplotypes for this taxon, we might feel comfortable believing there are between 16 and 31 actual individuals that were sampled. The 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 do not trust the proportion/frequency representation of an allele in our sample.

Another way to approach this is through 'true diversity' indices (reviewed in Sherwin 2010), as this family of statistics indicates the number of equally abundant types that have an average frequency equal to the observed average of types. This would mean that instead of taking H as a given from a previously-studied taxon or population, the extant data would be used to calculate

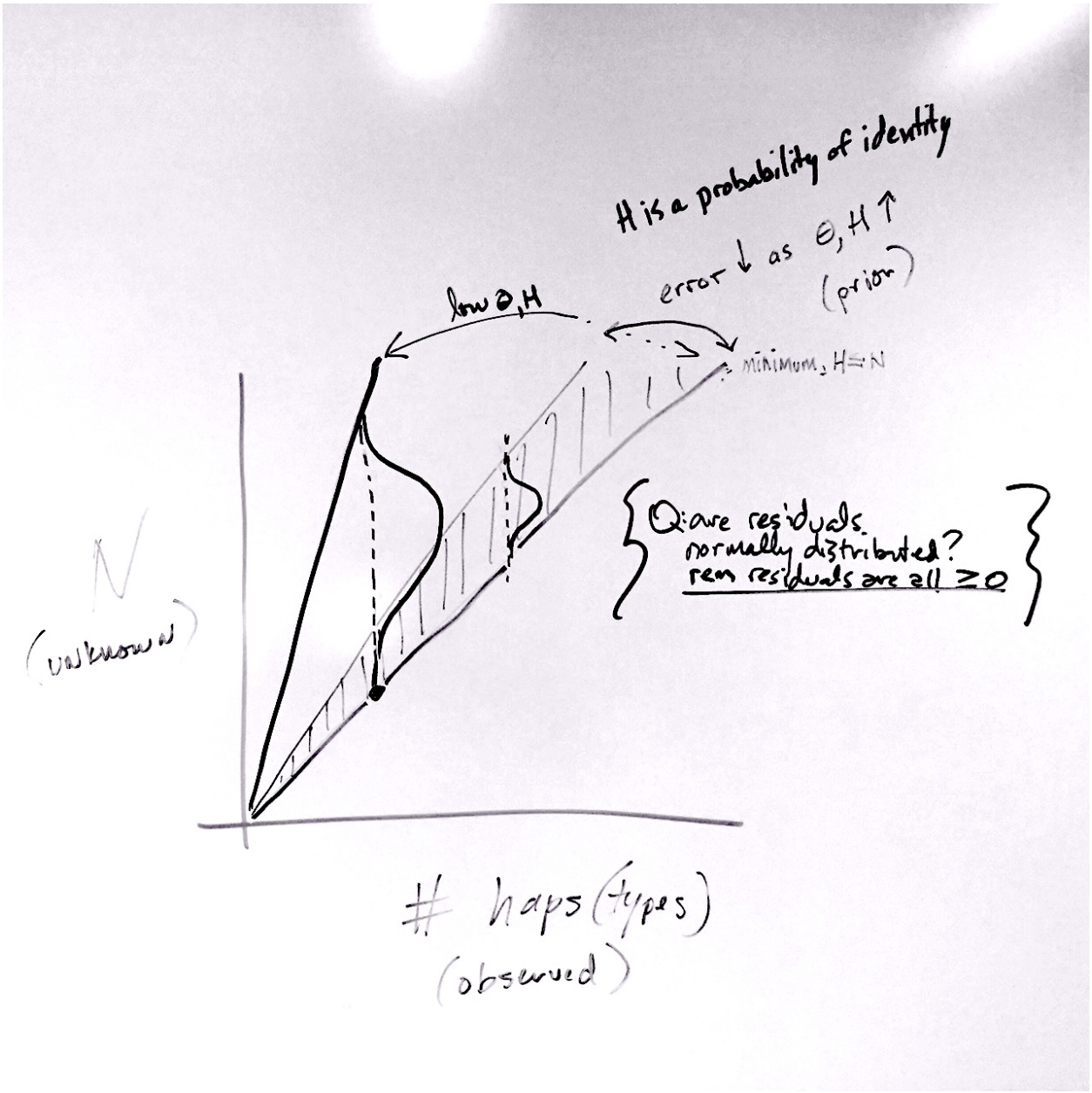
$$^{q}D = (\sum_{i=1}^{R} p_i^q)^{1/(1-q)}$$

where R is richness (number of types). As the Simpson (1949) index is used to indicate the probability that two individuals taken at random from the data are of the same (haplo)type, a comparable statistic to H is reached by taking the inverse Simpson index, or ^{2}D, and then transforming into the Gini-Simpson index as $1 - 1/^{2}D$, again the probability that two individuals have distinct haplotypes.

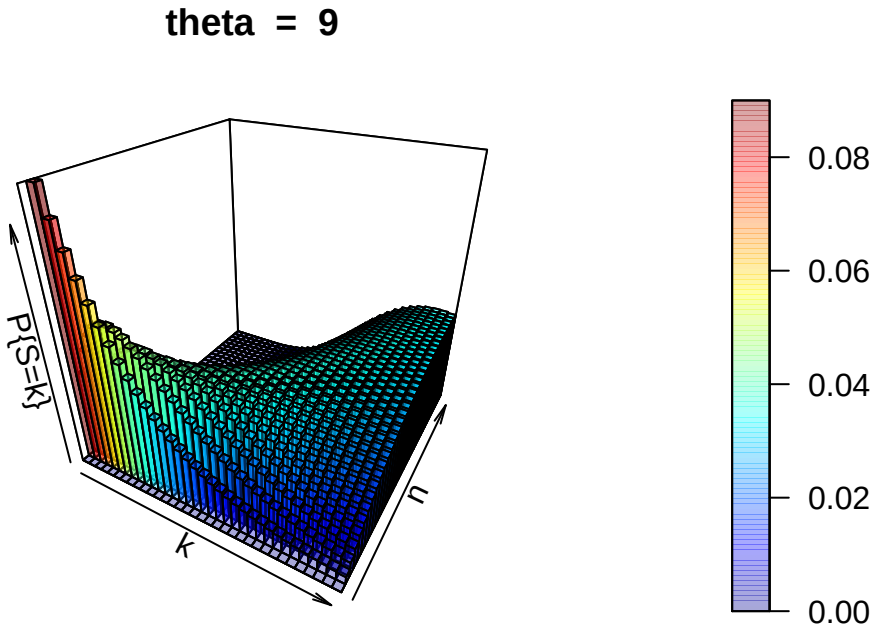
Evaluating a sample of sequences from the barnacle *Notochthamalus scabrosus*, where 20 individuals were haphazardly sampled from the data of Laughlin et al. (2012), we see that these data would traditionally report haplotype diversity H of 0.7578947, from 10 observed haplotypes (most dominant haplotype at frequency 0.5), and in this instance the Gini-Simpson index is equal to 0.72. Now we have another statistic that can be calculated from previous data on the population, that focuses on the number of dominant haplotypes. Here, the inverse Simpson index ^{2}D is 3.5714286, the effective number of haplotypes in the system.

Now I'm not sure what to do with this next. It isn't quite there, because now need to be able to pull a statistic from our UNKNOWN sample to improve on ability to recover the likely individuals that went in. It will still be true that for a population with low *H*, we have little information. (1) worth asking if frequency is OK within species? (2) even if you use frequency and recover the same basic value, there is no connection to N in this statistic on its own. So perhaps hard to improve on some set of assumptions around the Gamma?

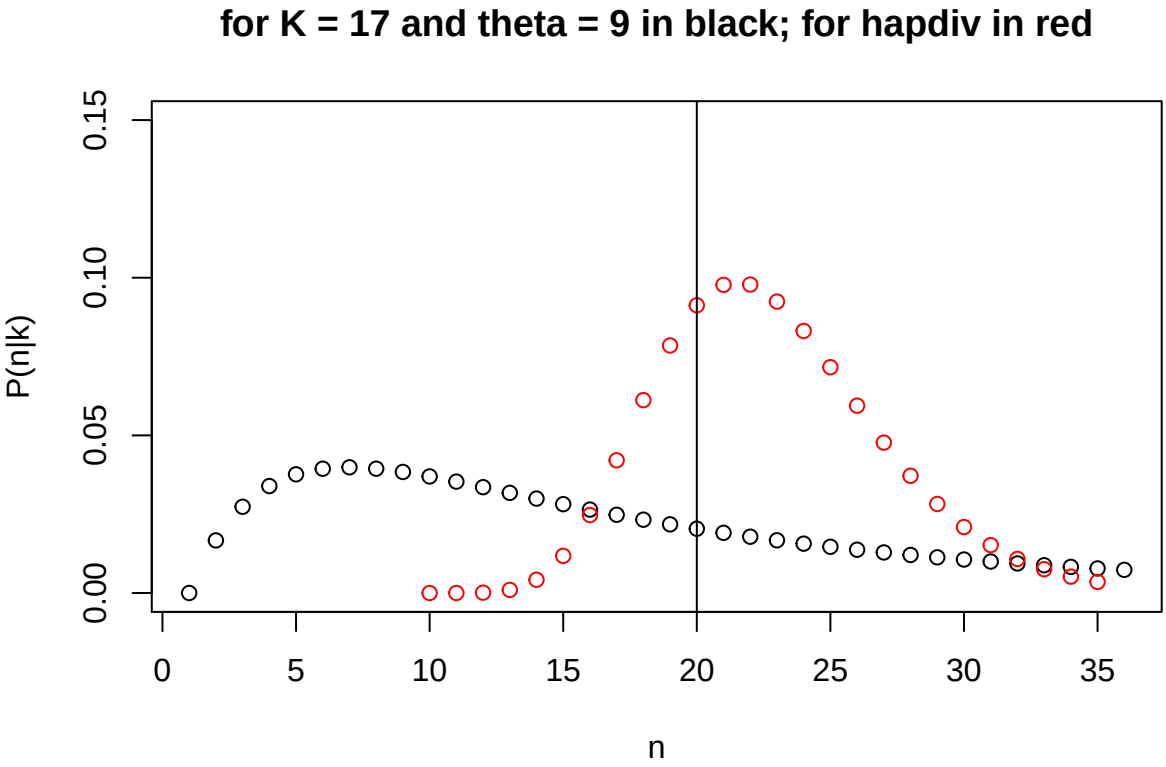
If it IS a probability that a newly sampled individual has a new haplotype, then observing X haplotypes should be able to give us back this distribution in a sense that ignores that the observed haplotypes have frequencies. So perhaps the distribution is bounded by NUMHAPS, but the distribution is shaped by SIMP such that if the effective number haplotypes is much lower than the observed number, that in itself says something about how likelihood changes as N > NUMHAPS (but then both the scale and shape are being influenced by same component??)



Then bring in what you can do with theta and the number of segregating sites.



Results



Remember that the input data for this single taxon included 20 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??

Discussion

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.

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.

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.

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.

Laughlin, K., C. Ewers, and J. P. Wares. 2012. "Mitochondrial Lineages in Notochthamalus Scabrosus as Indicators of Coastal Recruitment and Interactions." Journal Article. *Ecology and Evolution* 2: 1584–92. doi:10.1002/ece3.283.

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.

Sherwin, W. B. 2010. "Entropy and Information Approaches to Genetic Diversity and Its Expression: Genomic Geography." Journal Article. *Entropy* 12 (7): 1765–98. doi:Doi 10.3390/E12071765.

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.