Quantitative Metabarcoding

John Wares and Paula Pappalardo

February 17, 2015

Documenting the distribution and abundance of biodiversity 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, there are still glaring deficiencies in our ability to efficiently catalog the diversity of even related guilds at a location. Next generation sequencing has transformed the study of microbial and viral diversity, but eukaryotic diversity remains a more difficult prospect.

The reasons for this include the very diversity of these organisms. Though it is also known that "barcode" loci in microbes may vary in abundance among taxa and populations leading to inexact measures of relative abundance (picante ref), such a problem may be exacerbated when individuals of different life stages and different species vary in size and thus cellular quantity - a typical metazoan barcode locus, mitochondrial COI, would then over-represent the abundance of larger individuals.

Additionally, differential amplification of barcode genes, with efficiency declining with every base pair mismatch between genomic and primer sites, leads to further uncontrolled biases in enumerating the relative abundance of taxa (DOI: 10.1111/nph.12923) DOI: 10.1111/1755-0998.12355. Given the goal of including diversity that may be distantly related (e.g. the zooplankton from freshwater pools in South Carolina incorporate greater than 400 million years of phylogenetic diversity), we are often left with presence-absence data at best unless a skilled taxonomist can individually key these organisms out.

The approach in barcoding individuals of any taxon has typically followed a pattern of exploring the sequence divergence of samples from a catalog of known diversity. Either individuals are sequenced, thus maintaining both identity and recognition of abundance - a method that has not changed in over 20 years - or samples of many individuals are sequenced, and the match of individual reads from NGS data are either used to infer relative abundance (in the case of microbial diversity) or simply used to verify the presence of particular taxa. What has not been sufficiently considered is the baseline information we may have about a population or species that is measurable with these same data, in terms of genetic diversity.

The summary statistics for DNA sequence diversity are well established and generally recognize the population mutation rate THETA 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 (Wares 2010). Nevertheless, these summary statistics - including Watterson's theta, a sample-normalized estimator of THETA using the number of segregating sites in a sample - may provide information necessary to generate abundance patterns from NGS data. However, this information also has its limits: nucleotide diversity (pi) 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 apply the method to data from a well-characterized but low-diversity system of intertidal crustaceans. This serves as a pilot for far more diverse systems.

1. You have a bunch of FASTA files of sequence data, each file is a population. Read them in with loop (or WHILE), so each has an index number, each population has a short name.

2. Each population run summary statistics and attach those to the short name (array)

3. All sequences in 'unknown' file are BLASTed to all populations. Max hit (best match) is counted for each; if best hit is still x% divergent, they are put into 'unknown' category, otherwise they are pooled into category of best hit.

4. unknowns distance matrix, cluster and break into K populations somehow appropriate, perhaps by x'% divergent, label as unknown1...unknownZ

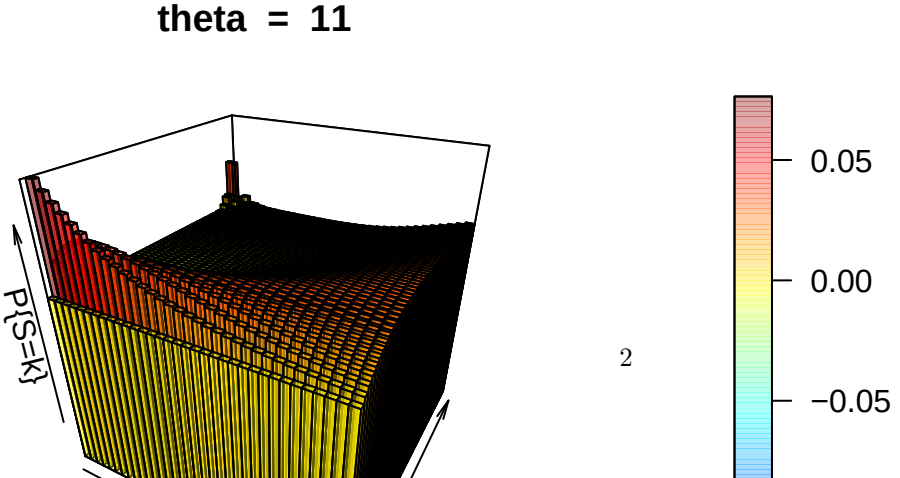
5. unknowns BLAST to known, net nucleotide divergence to nearest known is recorded along with that KNOWN

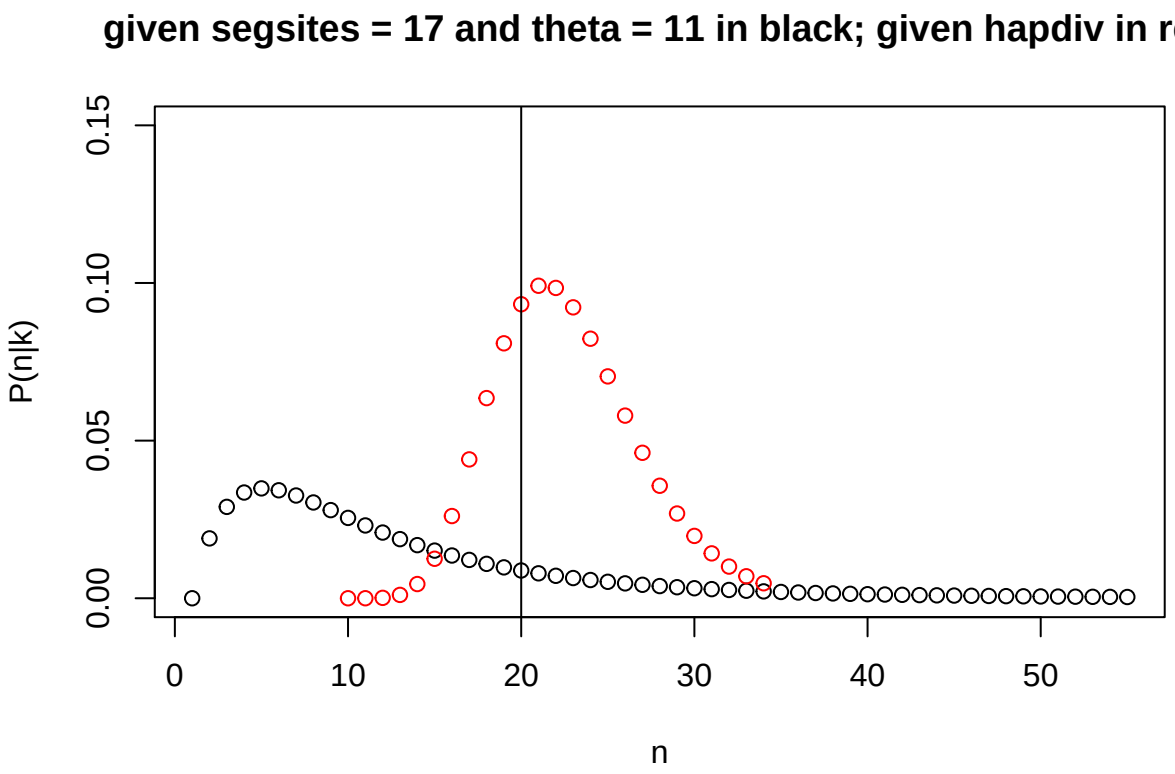
6. now you have all populations characterized and named either by what they BLAST to or what they are closest to (and how distant)

7. summary statistics on all the populations from the sample

do we want to use this script to re-do my old code for grabbing data from Genbank adn getting sumstats? now could use the package in Jehlius/picoroco papers...

8. we use summary statistics from sample populations relative to baseline PRIOR to estimate a POSTERIOR distribution of abundance in that sample.





9. can choose to even update the PRIOR based on this posterior, e.g. unknown 1 is now a population with SOME information on summary statistics (depends on sample size)

10. relative species abundance curve, most common species is in position 1 on X axis and abundance is plotted on Y as boxplot of likely abundance.... all the way down to the populations that you can only tell are present

11. you are now in the lognormal world of Hubbell, flawed though that may be.