Quantitative Metazoan MetaBarcoding

John Wares, University of Georgia – Ecology & Genetics

coin flip

Barcoding

When we use a standardized genetic marker to identify species from environmental samples, it is called "genetic barcoding"

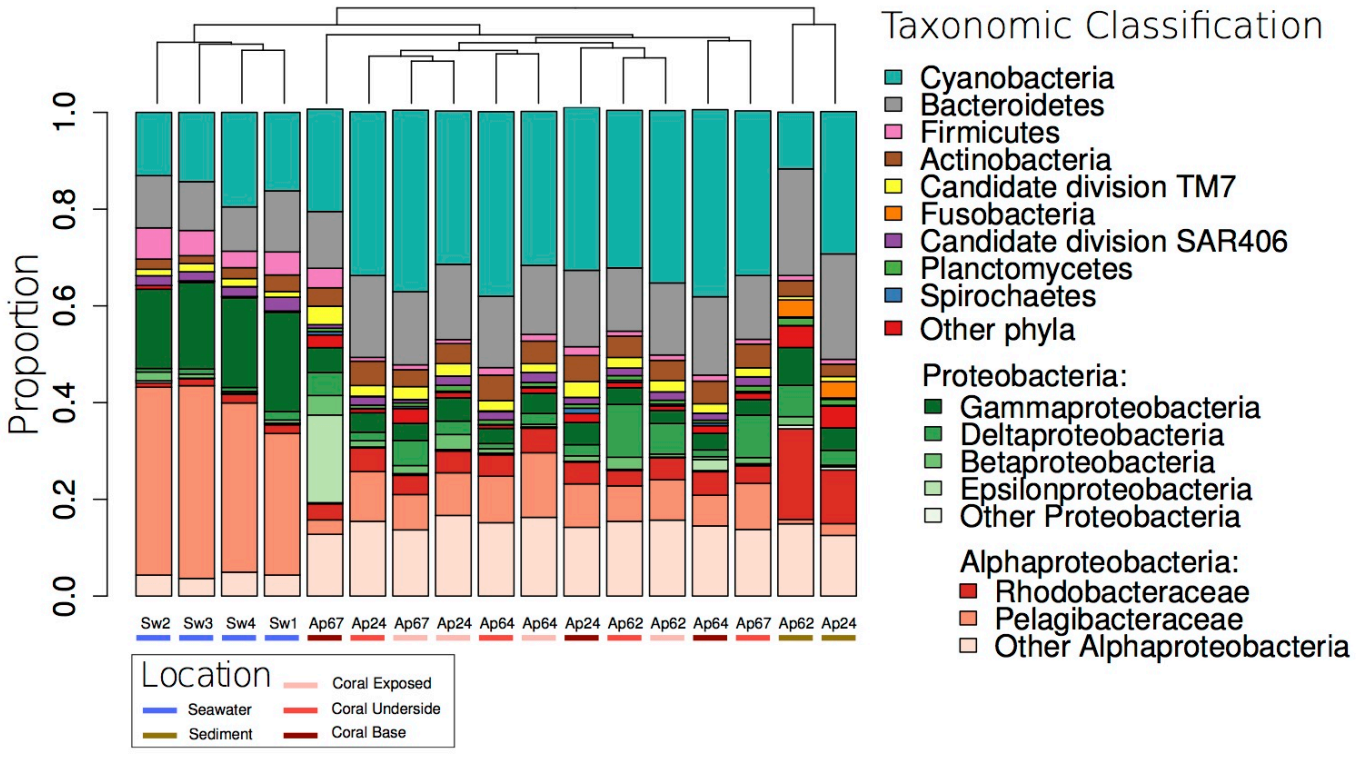
- mitochondrial COI and metazoans
- 16S ribosomal and microbes
- FORF and bivalves

Typically…

Individual tissue/sample harvested, preserved, DNA isolated, PCR, sequencing, match via BLAST or similar algorithm. Requires a database of 'knowns' already, study only as good as prior knowledge can make it!

Problems?

- Imagine a microbe (bacterium). They are probably all the same size, more or less…?
- They may (individual) have varying copy number of the 16S gene (see picante.R), and substitutions in the primer site affect PCR efficiency…!
- But we use such data routinely (e.g. QIIME) to represent the relative abundance of microbial taxa!



Now - think of small metazoans, eukaryotes…

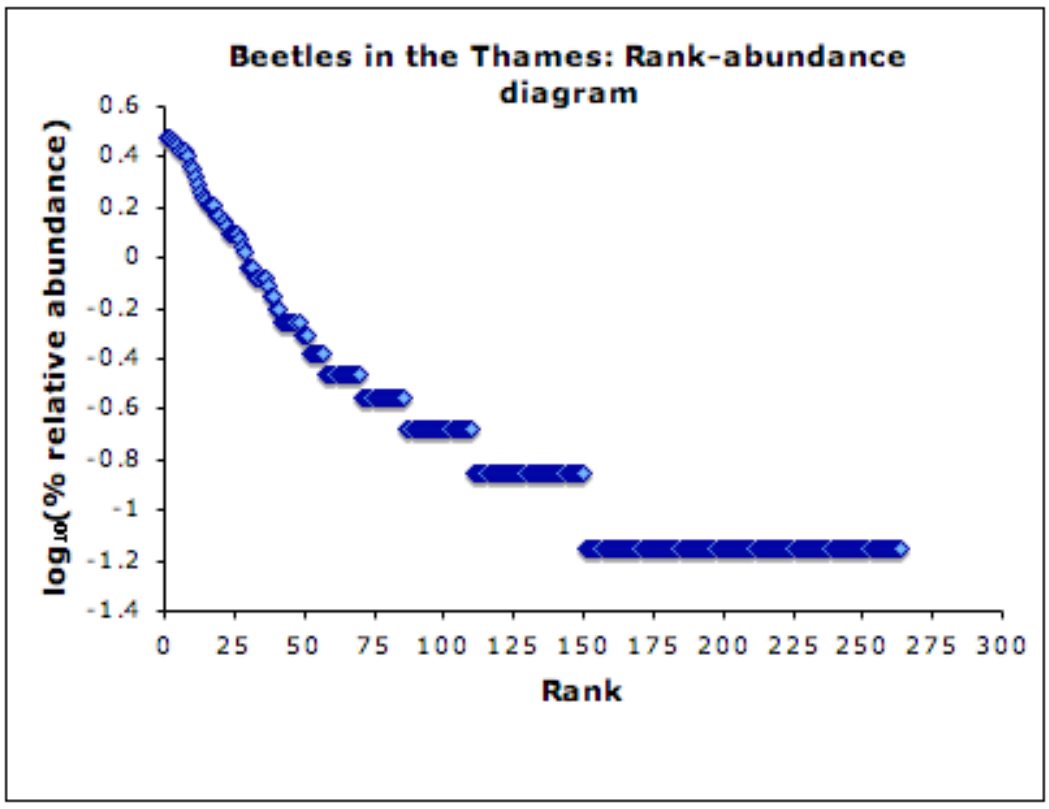
Potentially massive error. Almost useless other than presence/absence?

- variable volume of individual (2 equal abundant species, one 2x the volume…)
- variable PCR efficiency (can cause orders of magnitude error)
- we are left with qPCR approaches to identify 'presence'…
- or mixed environmental samples where abundance is unreliable!

Quantitative Barcoding (QuBar)

Goal: environmental sample of organisms (pile of amphipods from an estuary; jar of cladocerans from Carolina Bay; gut contents of a fish), use standard barcode marker, identify what species are present…

…and their relative abundance, with reliability **without using frequency of amplicon sequences**



Why?

Our ability to explore and interpret biodiversity - in space and time - is limited by throughput. Methods that have transformed microbial ecology are valuable for systems with eukaryotes but not yet as directly applied.

- community assembly and coexistence
- compositional change : environmental tracking
- cryptic diversity (within species) as well as cryptic diversity (hard to observe)

What information?

- may be able to gauge/estimate total # indivs (volume, count) as **upper limit**
- within related taxa, PCR efficiency and/or volume may not vary as much (so frequency not always useless)
- **summary statistics from the sequence data?!**

Summary statistics

Collect related DNA sequence data, align them to each other, that is a lot of information

Ways to summarize:

- how many unique haplotypes? haplotype diversity?
- how many variable/segregating sites in the alignment? (S)
- average pairwise differences (π)
- S, π are estimators of the almighty θ
  the 'population mutation rate'

Example from

Evaluating a sample of sequences from the barnacle , where 19 individuals were haphazardly sampled from the data of Laughlin et al. [-@Laugh12], we see that these data would traditionally report haplotype diversity of 0.7309942, from 9 observed haplotypes (most dominant haplotype at frequency 0.5263158), and in this instance the Gini-Simpson index is equal to 0.6925208. 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 ${}^2D$ is 3.2522523, the effective number of haplotypes in the system. The number of segregating sites is 12.

n.b. all numbers on previous slide automatically generated. change data set, numbers will automatically respond. thanks R!

n.b. number 2

The data on previous slide are from a small sample, an environmental sample of
. In order to apply this method, we are using about
(or whatever population/species) to reverse-engineer coalescent maths.

For example, I know from much work on this barnacle that θ is about 10. I also have prior information with which to estimate .

Population mutation rate and DIVERSITY

- π and other measures based on (polymorphic) site spectrum change little as sample size increases >10, and frequency is the problem we are trying to get around

- the number of segregating sites, however, is **sample-size dependent** and (it is just a number)

Probability: k, n, given θ

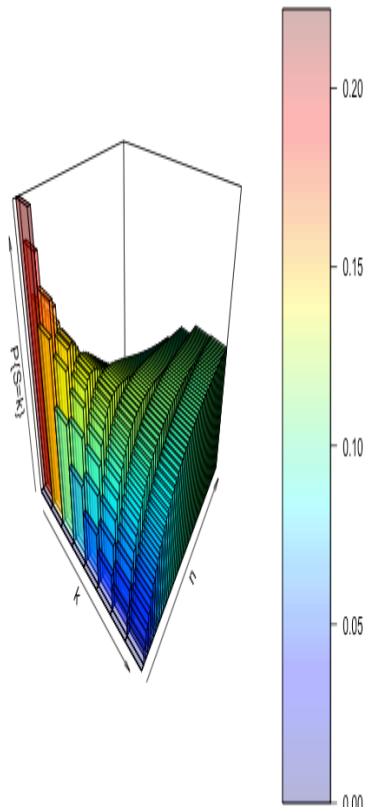
theta:

2

segregating sites:

4 20

1 5 9 17



Hap Diversity

Two ways to evaluate this, both have information in them.

A population can be described with 'haplotype diversity'

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

but note some frequency information in there…
(however at a , number of haplotypes is minimum number of individuals)

How to handle haplotype diversity?

Have been experimenting with a gamma distribution but feels very , e.g. what are parameters for gamma?

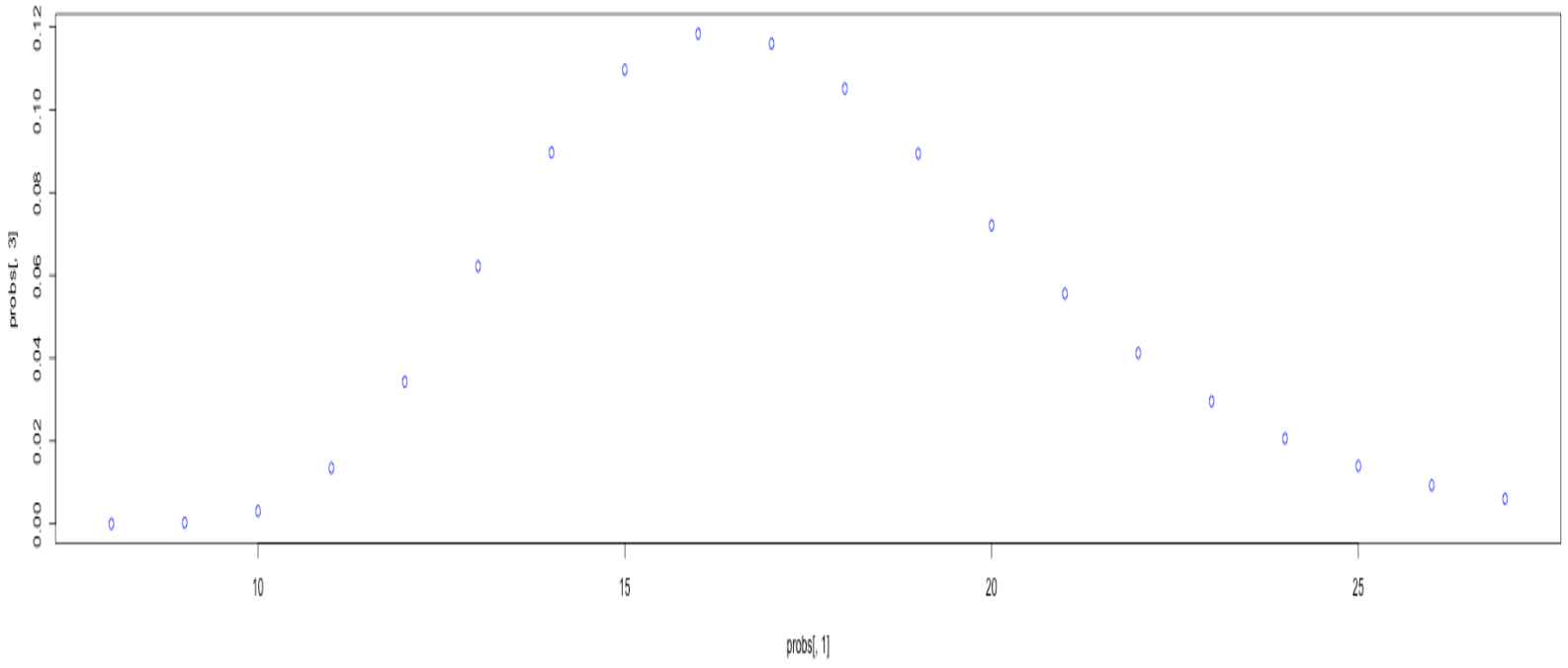
numhaps/shape param:

8

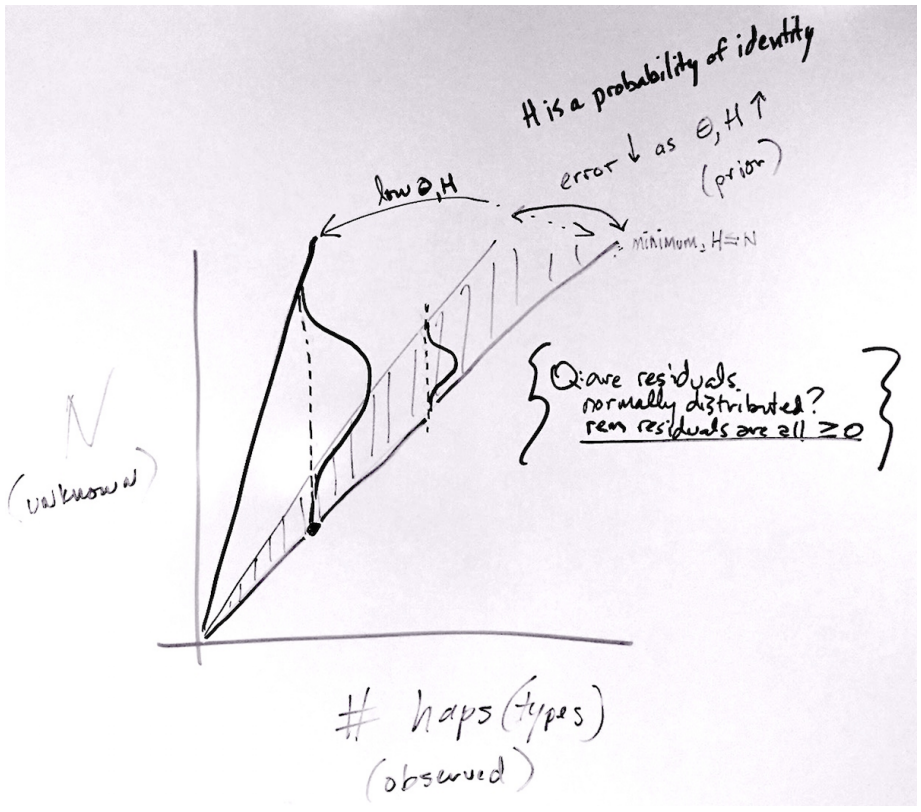
hap diversity/scale param:

0 0.8

0 0.4 0.8



Doesn't matter which



We are still dealing with how to handle, probabilistically, the number of individuals that went into our observation

the prior information about system

Other ways to handle diversity?

"True Diversity" http://en.wikipedia.org/wiki/Diversity_index (http://en.wikipedia.org/wiki/Diversity_index) generalizable form of diversity indices, including Shannon diversity which is closely related to haplotype diversity.

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

seems good because TD = number of types needed for the average proportional abundance of the types to equal that observed in the dataset of interest… the frequency can be removed from consideration in next-gen pooled sample, relative to what is known from prior empirical sample (I hope)

True Haplotype Diversity?

first lets see if 'haplotype diversity'

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

can be transformed into , or the effective number of haplotypes…

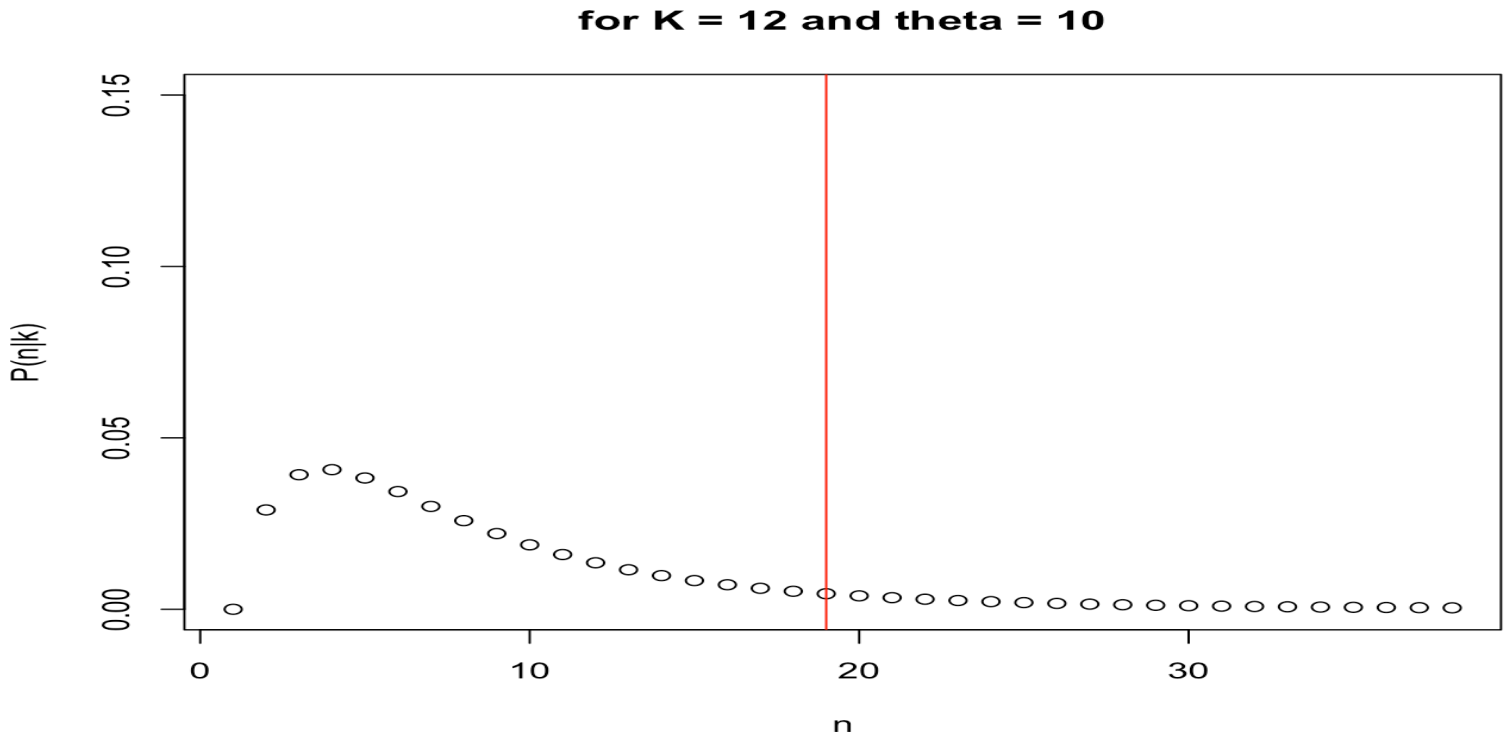
$$^2D = (\sum_{i=1} p_i^2)^{-1} = \frac{1}{\sum p_i^2}$$

close but no, that isn't it. (Shannon diversity been used instead of Haplotype diversity, but they are not exactly equivalent)…do we need to use haplotype diversity or can we simply use True Diversity on haplotype counts?

post hoc

- (note in my slide earlier, we are already using both and ^2D - but doesn't solve the nature of approach)
- also, of course, these are not independent ways of thinking about diversity. so, for example, can figure out which way works better, but cannot easily combine into a likelihood (I think. And Marc Feldman agrees.)

Test THETA against DATA

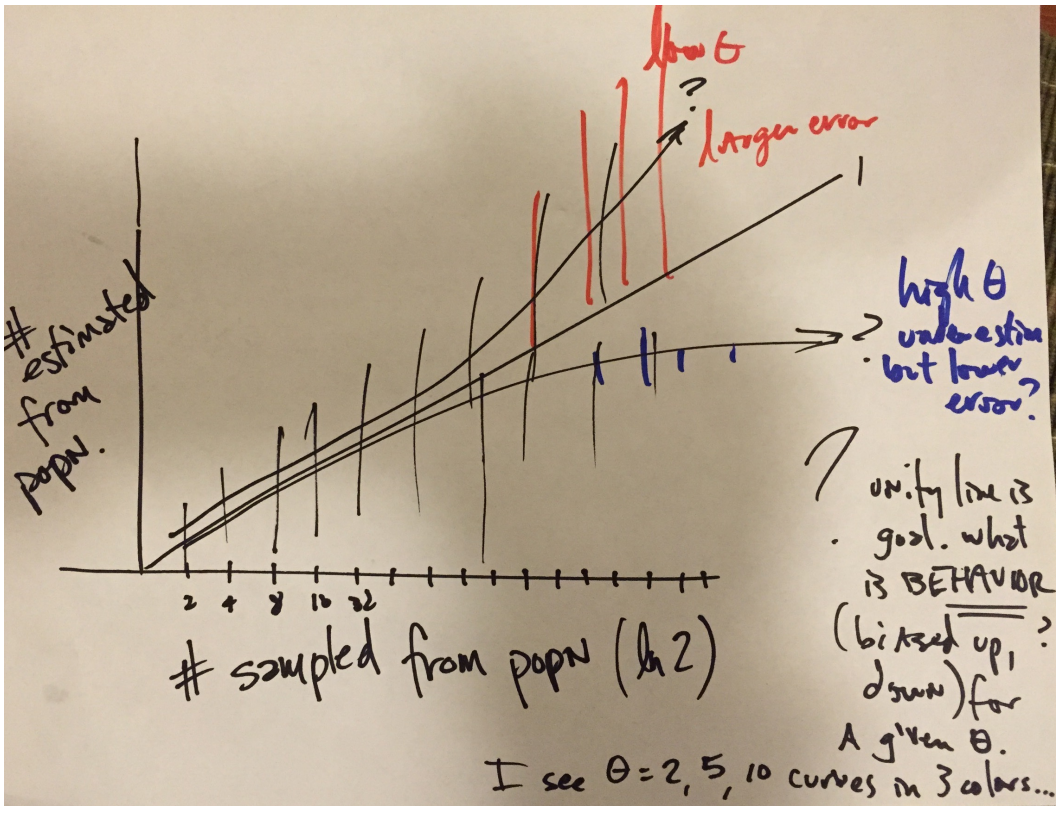


Remember that the input data for this single taxon included 19 individuals, the vertical line in plot above. NB THE DATA YOU INPUT RIGHT NOW MAY BE MIXED FOR NORTH AND SOUTH TYPES! FIX THIS INPUT FILE! More caveats: what about effect of natural populations/barcodes being in behavior???

Caveat selector

- worth remembering: "neutrality is dead" (Matt Hahn, 2015, but we've known for a while…)
- stat tests (e.g. Tajima's D) for standard mtCOI barcode clear evidence of purifying selection (Wares 2010, 2011; Ewers and Wares 2012)
- this analysis must account for diversity AND diversity processes at locus of choice
- hoping for analytical but may need to try something like approximate Bayesian

Simulations



may need to include simulations that influence site frequency spectrum to match input data?

Goals

- find strength/power/potential of which seems to have greatest power when θ large relative to number observed segregating sites
- power/potential of some form of diversity of haplotypes, also appears to have greatest power to estimate when high haplotype diversity (e.g., observed # haps likely closer to # individuals)
- overall high diversity and limited sample size may be best because cap on (if I know ~200 total indivs in sample, n_i must be <200), # haplotypes puts minimum on
- remember that individual barcode/identification of specimens is pricey and inefficient…not sure that this will improve markedly

Big Picture

Simple: I'm lazy. Good taxonomists are expensive. I want to improve our ability to do community surveys. Greater environmental sampling leads to greater environmental knowledge.

When the world is changing so fast, there is an urgency to find solutions and applications.

- Not totally clear that this qualifies as a solution, but a good exploration.

Together with…

Paula Pappalardo! (we are becoming fluent in .Rmd, git, more fun ways to talk like nerds in multiple languages)

thanks!

More Thoughts

the other things you are thinking of, including applications - screen Noto cline, Cfragilis cline - microcrustacean zooplankton (replacing Marcus Zokan!)

more info on Shiny and **ioSlides**

rmarkdown.rstudio.com/ioslides_presentation_format.
bullets

http://cpsievert.github.io/slides/markdown/#/
(http://cpsievert.github.io/slides/markdown/#/)