There once was a grid at ol' Carkeek

First Author*[1], Second Author[1,2], and Third Author[2]

[1]Department of Computer Science, LaTeX University

[2]Department of Mechanical Engineering, Superfabulous University

March 11, 2016

Keywords

Stuff, things, neat, cool, wow, instafun, tags4likes, etc

Abstract

This is the text of the abstract.

Introduction

Biodiversity surveillance is being revolutionized by DNA-based detection of organisms from environmental samples. ?(specifically speed and scope of ecological studies). Many researchers are justifiably cautious about the ?(adoption) of this new form of data. Their apprehension is rooted in the premise that traditional survey approaches are more accurate because the chain of inference between observation and ecological data is usually short: A researcher sees two swans in Lake Hopatcong and infers the lake is occupied by at least 2 swans. DNA based surveys, on the other hand, consist of a longer chain of inference: DNA sequences are reported by a sequencing machine, the machine identifies the sequence of products of a polymerase chain reaction (PCR), PCR amplifies

*first.author@funstuff.com

pieces of DNA from a purified genomic DNA sample, DNA is purified (extracted) from an environmental sample, environmental samples contain DNA from organisms present, the organisms present are representative of the biological community about which we wish to make inference. ?(reverse order? tie to concrete example (swans of Lake Hopatcong)). Clearly, this process is more complex than visual surveys, as the relationship between several steps is complex or unknown. But consider that the processes ?(behind | underlying) other more widely-used ecological survey techniques are similarly complex, such as bird surveys based on song, or visual identification of fungal spores. When alternate survey approaches are impossible or inefficient, we are more willing to accept any available survey data, regardless of the complexity or uncertainty underlying it. (microbiologists have enthusiastically relied on DNA-based surveys for years for this reason, (though yes, they also do not have the problem of disconnect between individual and cell)).

The ability of DNA surveys to make quantitative inference about communities has been touted by some (CITE new fish quantitation paper) and doubted by others (CITE european eelgrass PLOSONE). For example, a study linking (blah blah blah) concluded that "metabarcoding is powerful, yet blind" (CITE european eelgrass). Conversely, others have reported strong quantitative and intuitive links between DNA-based and traditional survey methods (CITE Port 2016 MOLECO). These studies usually rely on simple statistical models to link DNA quantity to some measurable ecosystem property like biomass (but see CITE). When confronted with data collected in ?(complex ways/studies/whatever), simple models ?(may | often) fail to detect relationships when they exist, or vice versa ?(they are prone to inflated risk of BOTH type I and type II error) (CITE, see Woltman 2012). For example, (CITE, look for that Gelman paper) have demonstrated that when data are structured in a hierarchical fashion (e.g. test scores of students in schools belonging to districts belonging to states), a low number of replicates at the first level of hierarchy (SEE THE PAPER). Similarly, (describe hospital/school problems).

Shelton et al. (CITE Shelton 2016) outlined an approach for structuring statistical models of DNA surveys that address these issues. This framework improved on alternative statistical techniques by explicitly accounting for the ?(hierarchical | nested | multilevel) structure of the study design, which allows error and uncertainty at each level to be ?(explicitly accounted for| modeled | propagated throughout the model). That study demonstrated an improvement in the estimate of higher-level (e.g. ecological community) quantities when the processes linking them to the data are specified. As an example, it was shown that incorporation of data about the mismatch between primer and template DNA sequence can improve the estimate of the relative abundance of unique DNA templates input to a PCR.

Here, we apply this framework to a DNA survey of ?(nearshore | coastal) marine habitat. (TODO add commentary on current dogma surrounding distribution of DNA in well-mixed (marine) habitats). We document the variability associated with lab based ?(procedures | replication | treatment; i.e. filter+DNA+PCR+seq), and the spatial scale over which DNA communities vary in this habitat. We ?(show that | tested whether) a taxon's spatial distribution predicts (the slope of the relationship between distance from shore and DNA abundance or to what degree DNA abundance is explained by distance from shore for each taxon). We focus partly on species with known life histories that define their spatial distribution (e.g. shallow water livebearing fishes or sessile intertidal organisms with ?(motile/planktonic/pelagic) larvae or gametes). For these taxa whose spatial distribution is well-documented and restricted, we calculate the rate of change in space and compare this rate among taxa with similar spatial distributions. In turn, the distribution of rate of change serves as an estimate of the spatial distribution of DNA in this habitat.

Methods

We use the general framework outlined by Shelton et al (CITE). That study outlined the structure for estimation of the proportional biomass of a taxon (B_i) given the proportional counts of sequences recovered from a parallel sequencing run (Z_i).

We modeled the counts of DNA sequences (Z) from each of a given taxon i, in each replicate PCR j, from each replicate of a given location k (hence, Z_{ijk}), as though they are ?(proportional to/drawn from)? a Poisson distribution. A Poisson distribution is described by one and only one parameter, λ, which is equal to both the mean and variance. Because in this case our modeled values are discrete counts, we use the natural exponent, e^{λ}. Thus,

$$Z_{ijk} \sim Poisson(e^{\lambda_{ijk}}) \tag{1}$$

In turn, we further assume this parameter λ is linearly proportional to a suite of taxon-, pcr-, and site- specific parameters describing the variance associated with each sub-process linking the amount of DNA (Y) of a given taxon i at a given location k in a DNA extract (hence Y_{ik}):

$$\lambda_{ijk} = \beta_0 + \beta_i + \eta_{ijk} + \epsilon_{ijk} \tag{2}$$

Where β_0 is a general intercept across all taxa, β_i is a fixed effect accounting for the variance associated with taxon i, and η_{ijk} and ϵ_{ijk} are random effects of variance resulting from the processes associated with PCR and spatial location, respectively.

Results

We found that if you have two apples, and someone gives you another two apples, you have four apples.

Discussion

Boy those results sure are neat. Now, the pressing question becomes: How do you like them apples?

Acknowledgements

We wish to thank all of the little people.

Funding

This study was funded by our super-rich uncle.

Author Contributions

Conceived and designed the experiments: James L. O'Donnell, Ryan P. Kelly, A. Ole Shelton. Collected the data: James L. O'Donnell, Greg Williams, Natalie C. Lowell, Ryan P. Kelly, A. Ole Shelton, Jameal F. Samhouri. Conducted the analyses: . Wrote the first draft: . Edited the manuscript: .

Data Availablity

The data and code used to generate our results can be found at the following url:

Figures