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

May 3, 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). While this approach has been used for decades to survey micro-organisms (CITE VENTER), it has more recently become used as a tool for surveying macro-organisms (CITE EARLY EDNA). The technique is founded on the premises that: 1) organisms shed DNA-containing material into the environment, 2) the DNA contained in this material breaks down over time, which leads to the conclusion that: (1) the concentration of DNA originating from an organism should decrease with distance from its source.

*first.author@funstuff.com

For an effectively immobile organism (e.g. fungal hyphae) in an effectively static environmental medium (e.g. soil), the relationship between DNA concentration and distance appears straightforward (CITE ???). One might expect the relationship to break down in a dynamic environment: An immobile tree may shed cells into the air, which may be carried by the wind for some distance. This has been a cause for concern for ecologists interested in surveying aquatic organisms, because the robustness of eDNA surveys in aquatic environments depends on the distance that DNA travels from a source. This distance will be affected by the nature of the environment: Material is less likely to be transported (quickly over) great distances in an ice-covered lake than along a turbulent shoreline or a fast-moving stream. One expects more DNA downstream of a fish in a river than upstream, and

In marine environments, this relationship is further complicated by the fact that many organisms have dispersive early life history stages (gametes and larvae) that can travel great distances but are difficult to track. These individuals will obscure the expected relationship between DNA concentration and adult abundance because they shed genetic material into the environment and they are small enough to be captured directly by environmental sampling protocols. For example, while a traditional survey of a sessile, intertidal bivalve would identify the most biomass along the shore, there may be a greater concentration of their genetic material at some distance away during a spawning event, or in an ocean current carrying many larvae.

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 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.

We would love to estimate the minimum distance over which eDNA community differences can be detected.

Some authors have cautioned against the use of DNA-based macrobial communities in marine environments because they are subject to dynamic physical forces (CITE).

Samples collected of ecological communities may vary in dissimilarity from 0 (completely identical) to 1 (completely different). For samples collected from multiple locations, the relationship between their spatial distance and community dissimilarity is of interest because it reflects the amount of community heterogeneity—changes in abundance and composition—over the spatial scale sampled. The intercept is expected to be 0, because only within-sample comparisons can have 0 spatial separation, and communities have no dissimilarity within a sample if sampling method is repeatable. Likewise, dissimilarity cannot exceed 1, and reaches 1 only when multiple discrete community types are sampled. The trend is expected to be asymptotic if communities within a habitat are spatially structured, where the value of the asymptote and the rate of increase provide insight about community turnover. A flat relationship indicates that heterogeneity is not assorted spatially, and can be interpreted in different ways, depending on the asymptote. If the asymptote is close to 1, there is high spatial heterogeneity over the spatial scale of sampling. If the asymptote is close 0, all samples are similar, and we infer there is complete community homogeneity over the scale sampled. The rate at which community dissimilarity approaches the mean gives an indication of the rate of community turnover.

Methods

Environmental Sampling

Starting from lower-intertidal patches of *Zostera marina*, we collected water samples at 1 meter depth from 8 points (0, 75, 125, 250, 500, 1000, 2000, and 4000 meters) along three parallel transects separated by 1000 meters (Figure 1).

Laboratory Procedures

Samples were randomly assigned to PCR primer and library adapter index sequences. The sequencing run consisted of 14 samples ('libraries') prepared using different index sequences ligated during library preparation. Of these libraries, ten comprised of amplicons prepared using the 16S protocol reported above, and four comprised of amplicons prepared using a 12S protocol similar to that reported by (CITE PORT 2015).

Pooled libraries were sequenced on the Illumina NextSeq platform at the Stanford Center for Functional Genomics (machine ID: NS50061; run ID: 115; flowcell ID: H3LFLAFXX). Raw sequence data in fastq format is publicly available (see Data Availability).

Data Preparation (Bioinformatics)

Detailed bioinformatic methods are provided in the supplemental material, and scripts used from raw sequencer output onward can be found in the project directory on GitHub (see Data Availability).

We calculated rates of cross-library contamination by counting occurrences of primer sequences: 12S primer sequences appearing in a 16S library (and vice versa) indicate an error in the preparation or sequencing procedures.

We checked for experimental error by evaluating the Bray-Curtis dissimilarity of proportional read abundance among replicate PCRs of the same DNA sample (0.033±0.063), and excluded one PCR replicate for which the dissimilarities between itself and the other replicates exceeded 1 SD (lib_B_tag_GCGCTC).

To account for variation in the number of sequencing reads (sequencing depth) recovered per sample, we multiplied the within-sample proportional abundance of each OTU by the minimum sequencing depth (130402), and rounded to the nearest integer.

Because each step in the massively parallel sequencing workflow is sensitive to contamination, it is possible that some sequences are the result of contamination during field sampling, filtration, DNA extraction, PCR, fragment size selection, quantitation, sequencing adapter ligation, or the sequencing process itself. Some authors have argued that these risks could bias sequence abundance, making those data meaningless and prohibiting quantitative estimates, yet convert count data to binary presence absence data on the basis of the sequence abundance greater than some arbitrary threshold). Recent work has shown that this binary treatment of data can (?falsely?) overestimate taxon richness and falsely elevate the estimate of taxon turnover among samples (CITE LERAY FORTHCOMING). We take the view that contaminants are unlikely to manifest as sequences in the final dataset in consistent abundance across replicates; indeed, our data show that the process from PCR onward is remarkably consistent. Thus, we calculated from our data the maximum number of sequence counts (after scaling to correct for sequencing depth variation) for which there is turnover in presence-absence among PCR replicates within an environmental sample. We use this number to determine a conservative minimum threshold above which we can be confident that counts are consistent among replicates and not of (?spurious | dubious?) origin, and exclude from further analysis observations where the mean abundance across PCR replicates within samples does not reach this threshold.

Ecological Analyses

We subset the data in a variety of ways and conducted each analysis on all subsets. We report the subset used with each analysis, and report results on alternative subsets in the supplemental material. For all analyses beyond the assessment of PCR consistency, we use the mean taxon abundance across PCR replicates from each of the 24 environmental samples. Our subsetting methods were (1) exclude rare taxa ?(threshold)?, (2) exclude abundant taxa ?(threshold)?, (3) subsampling of taxa randomly, (4) subsampling of taxa proportional to their abundance, (5) subsampling of taxa inversely proportional to their abundance, (6) exclude taxa found in only one environmental sample (spatially invariant), (7) exclude non-marine taxa (e.g. humans, pigs), (8) exclude taxa whose known individual range (including gametes and larvae) exceeds the spatial scale of our study. We also tested a variety of transformations of the mean scaled abundance data, including (1) log ($log_e x$), and (2) binary ($1 = x > 1$; $0 = x < 1$).

We simultaneously assessed the existence of distinct community types and the membership of samples to those community types using a partitioning around mediods algorithm (CITE PAM, sometimes referred to as k-medoids clustering), as implemented in the R package fpc (CITE fpc). The classification of samples to communities was made on the basis of their pairwise Bray-Curtis dissimilarity, calculated using the function vegdist in the R package vegan (CITE VEGAN).

We calculated the great circle distance between points using the Haversine method as implemented by the R package geosphere (CITE geosphere).

To estimate the maximum dissimilarity and the rate of community turnover in space, we modeled community dissimilarity as a function of distance from shore following a Michaelis-Menten relationship:

$$com \sim V_{max}[d]/K_m[d] \tag{1}$$

where *com* is community dissimilarity, d is spatial distance, and where the asymptote is given by V_{max}, and the distance at which half the asymptote has been reached is given by K_m. Model fit was assessed using the function nls in R (CITE R).

Organismal Life History Data

We collated coarse-scale data on life history characteristic for each of the major taxonomic groups recovered, including dispersal range of the gametes, larvae, and adults, adult habitat type and selectivity, and adult body size. Dispersal range was given as an order-of-magnitude approximation of the scale of dispersal: for example, internally fertilized species were assigned a gamete range of 0 km, while broadcast spawners were assigned a gamete range of 10 km. Similarly, adult range size was approximated as 0 km (sessile), 1 km (motile but not pelagic), or 10 km (highly mobile, pelagic). Variables were specified as 'multiple' for groups known to span more than 1 magnitude of range size. For groups to which sequences were annotated with high confidence, but for which life history strategy is diverse or poorly known (e.g. families in the phylum Nemertea), we used conservative, coarse approximations at a higher taxonomic rank. We assessed whether or not marine invertebrate taxa are thought to be present in Puget Sound by checking for their presence in a comprehensive checklist of invertebrates of Puget Sound CITE KOZLOFF. These data are available as part of the REFERENCE SUPPLEMENTAL DATA.

Spatial Model Formulation

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,

$$Z_{ijk} \sim Poisson(e^{\lambda_{ijk}}) \tag{2}$$

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}):

$$\lambda_{ijk} = \beta_0 + \beta_i + \eta_{ijk} + \epsilon_{ijk} \tag{3}$$

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

Data Quality (Bioinformatics)

All value ranges are reported as (mean ± standard deviation).

There was a very low frequency of cross-contamination from other libraries into those reported here (5e-05±8e-05; max 0.00034)

We assessed the consistency of PCR by conducting 4 replicate PCRs for each environmental sample and calculating the mean pairwise Bray-Curtis dissimilarity of the resulting communities (scaled to minimum read depth per sample). 92 of the 96 amplicon samples had mean Bray-Curtis dissimilarity ≤ 0.052; 1 sample had a value of 0.341, which elevates the value of the other replicates. After removal of this sample, the highest mean Bray-Curtis dissimilarity among replicates within an environmental sample was 0.034.

Community Analysis

Excluding spatially-invariant taxa ?(taxa which occur in only one spatial location) had no discernible effect on the outcome of the PAM analysis (number of clusters, assignment to clusters).

The estimated asymptote of community dissimilarity as a function of spatial distance (V_m) was 0.72 ($p <<< 0.05$), and the distance at which half this dissimilarity was accumulated (K_m) is 23.8 kilometers ($p = 0.006$). Residual standard error of the fit of the model is 0.1563 on 274 degrees of freedom.

Spatial Model Output

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

All sequence files and metadata are available from EMBL:
http://www.ebi.ac.uk/ena/data/view/XXXXXXXXX
All analyses were performed using scripts available from the project repository on GitHub:
https://github.com/jimmyodonnell/Carkeek_eDNA_grid

Figures

Supplemental Material

Bioinformatic Methods

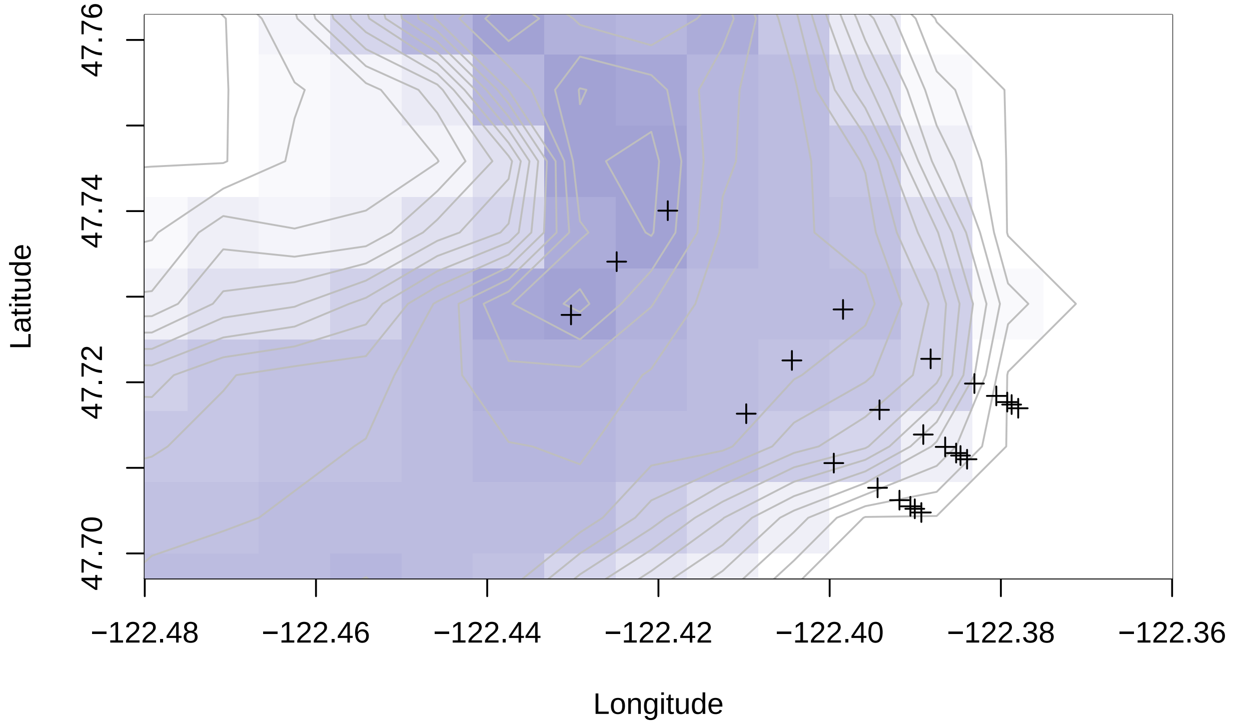

Figure 1: TODO: Plot with GEBCO 30-second data or remove grid coloring and color by isobath. Looking into filling by contour. Geographic position of collected samples. Lines give XXX meter isobaths.

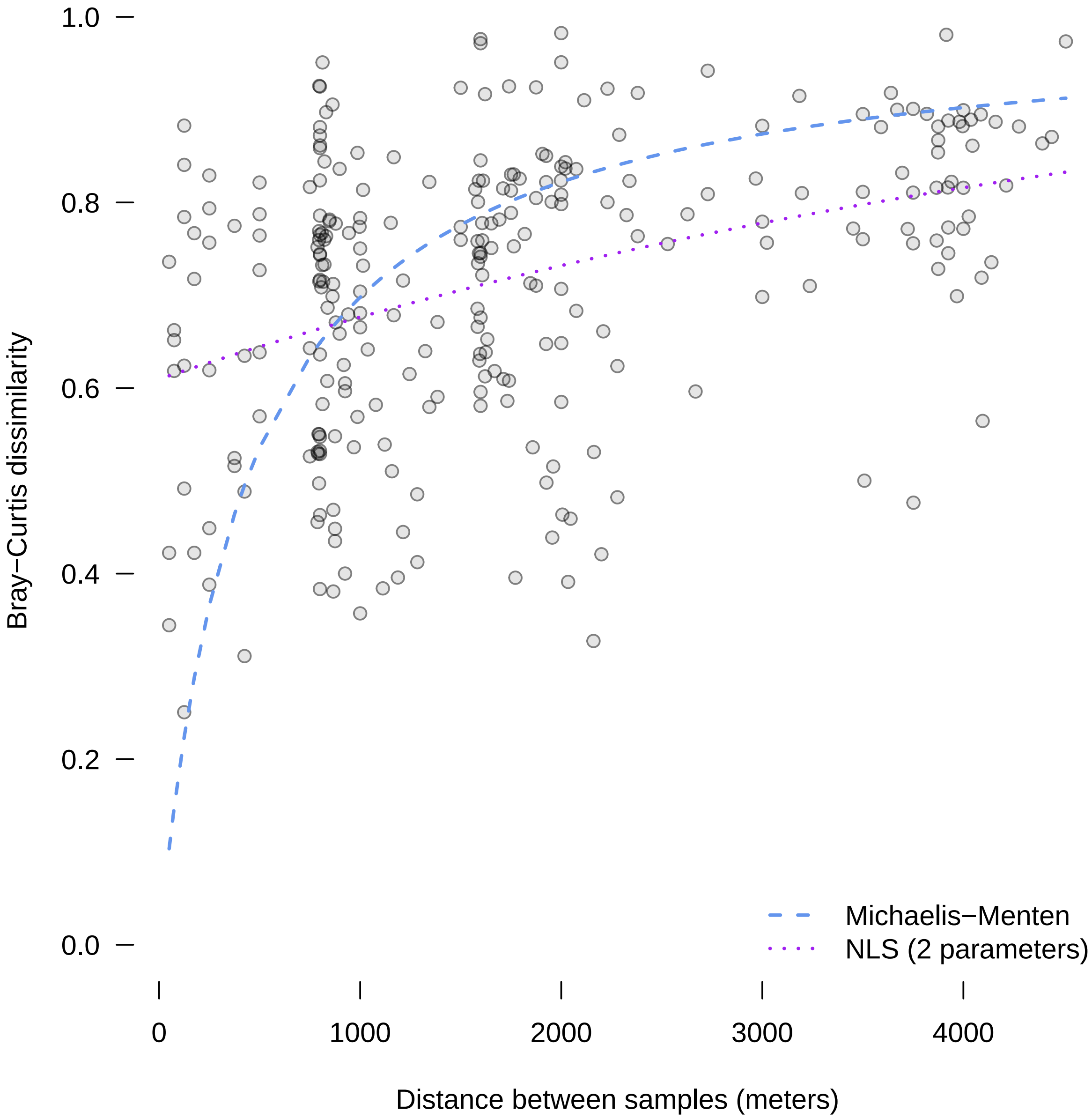

Figure 2: Pairwise Bray-Curtis dissimilarity of eDNA communities plotted against pairwise spatial distance. Line represents prediction of the Non-linear Least Squares regression to a Michaelis-Menten model ($V_m = 0.72$, $p <<< 0.05$; $K_m = 23.8$ kilometers, $p = 0.006$; RSE = 0.1563; df = 274.). Restricting comparison to within-transect has no qualitative difference in the outcome (see 'diss_by_dist_by_transect.pdf').

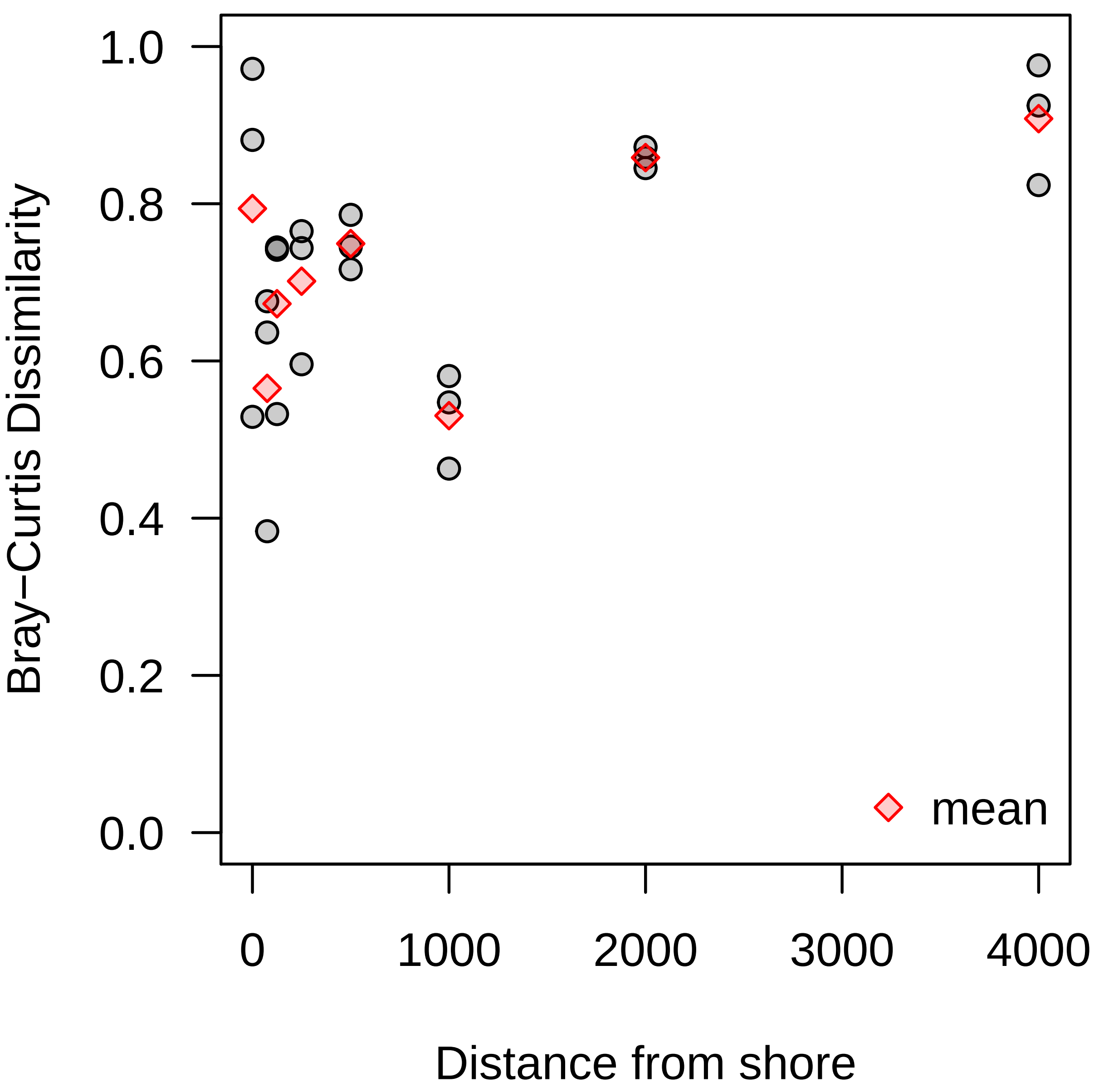

Figure 3: Pairwise dissimilarity (Bray-Curtis) across transects plotted against distance from shore. A linear model determined no effect of distance from shore on dissimilarity.

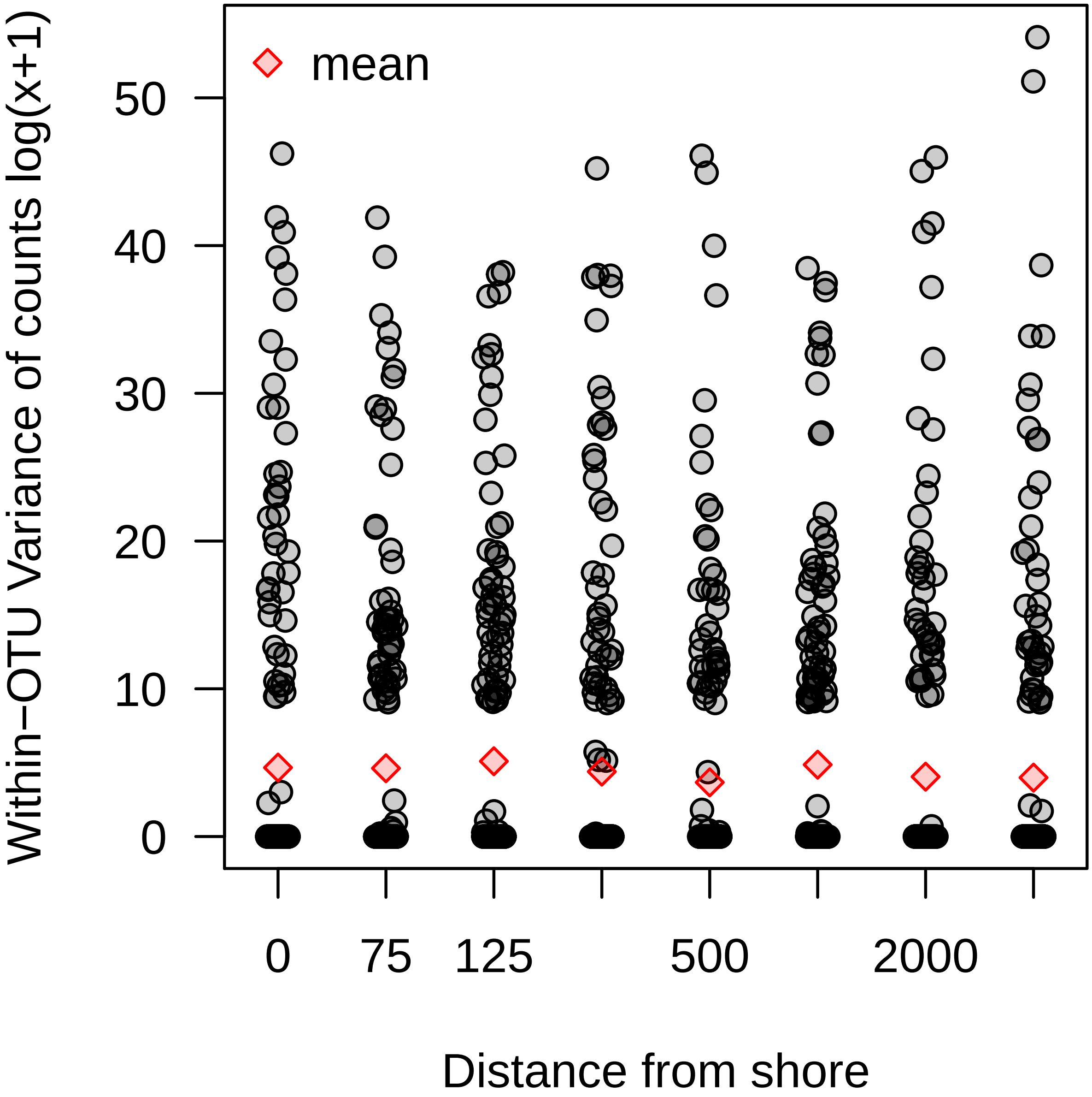

Figure 4: Within-OTU variance across transects plotted against distance from shore.

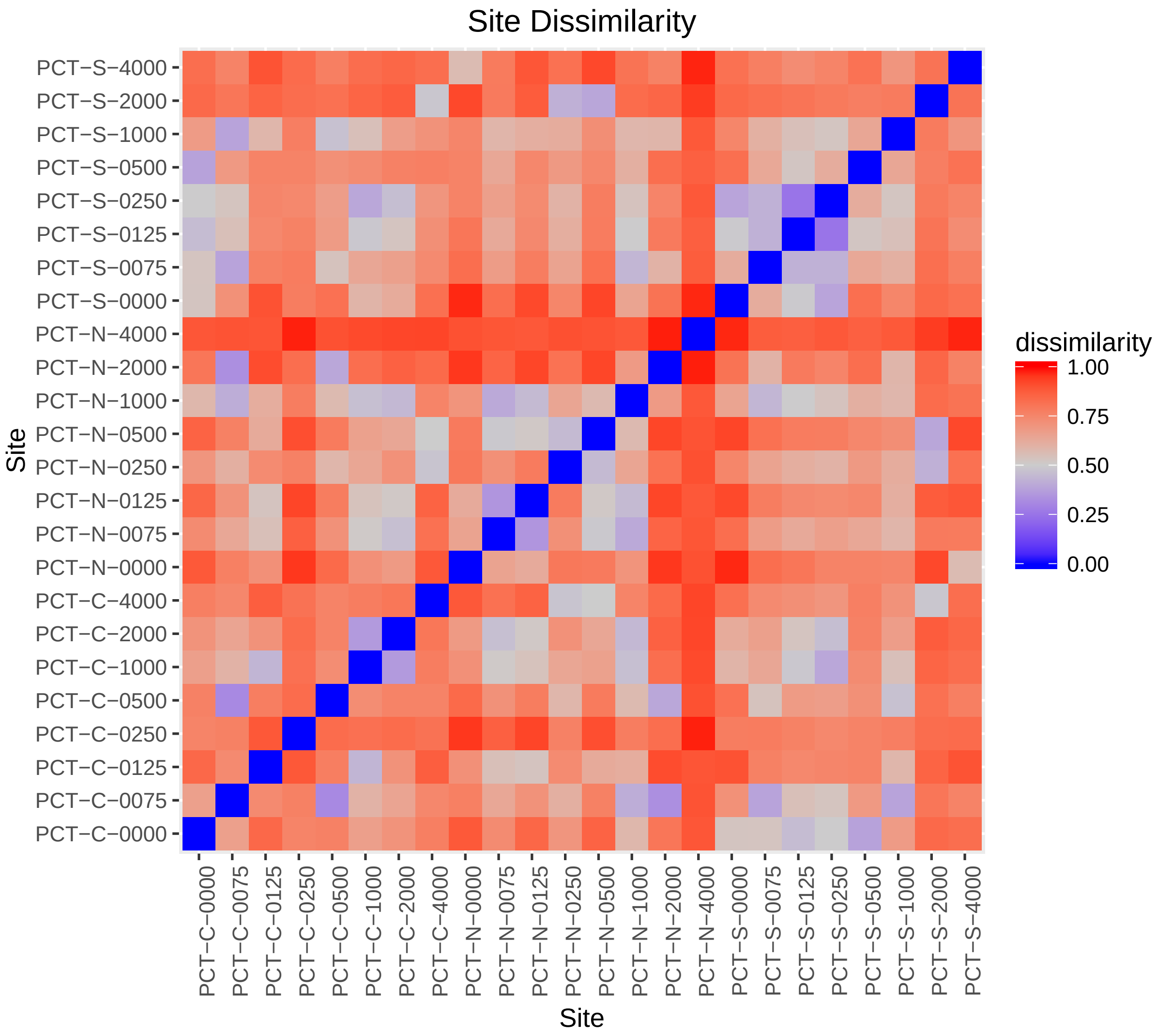

Figure 5: Heatmap of pairwise site dissimilarity (Bray-Curtis).

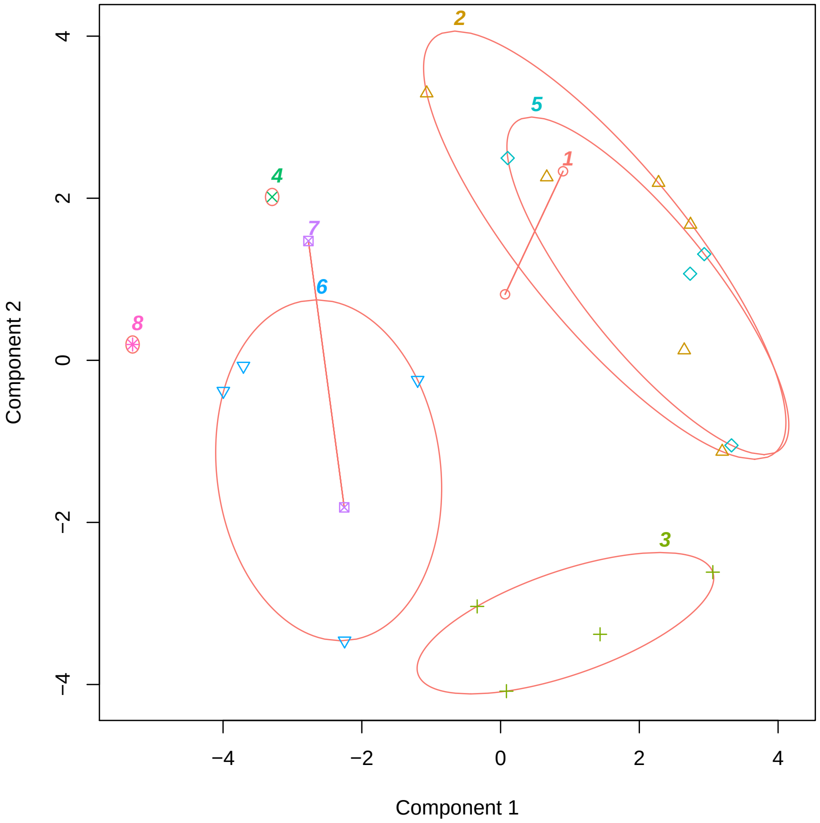

Figure 6: TODO figure out color of ellipses; I can't even plot them gray without Plot of partitioning around medoids (PAM) analysis of OTU sequence abundance from 4 replicate PCRs at each of 24 sampling points. Points represent communities of OTUs; color and shape indicate cluster membership as determined by PAM analysis. Ellipses indicate the smallest area of a cluster that contains all of its members.

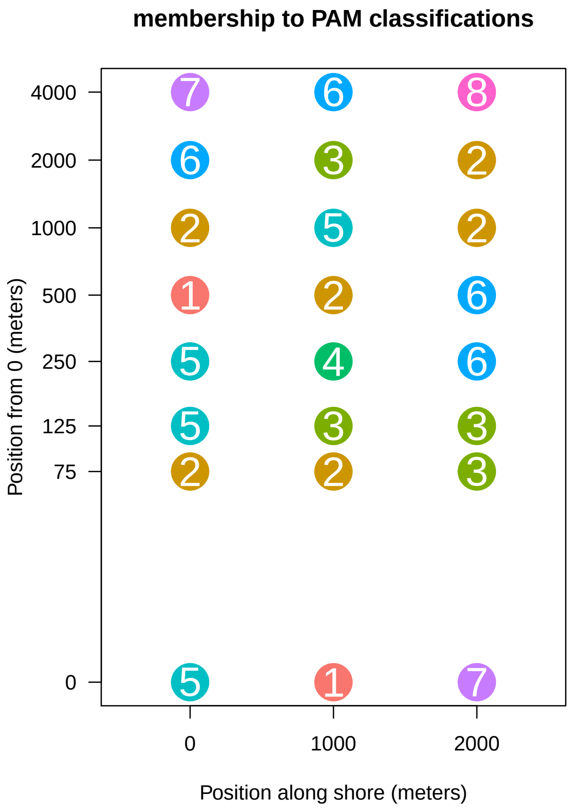

Figure 7: Geographic position of collected samples, colored by membership to clusters identified by partitioning around medoids algorithm.

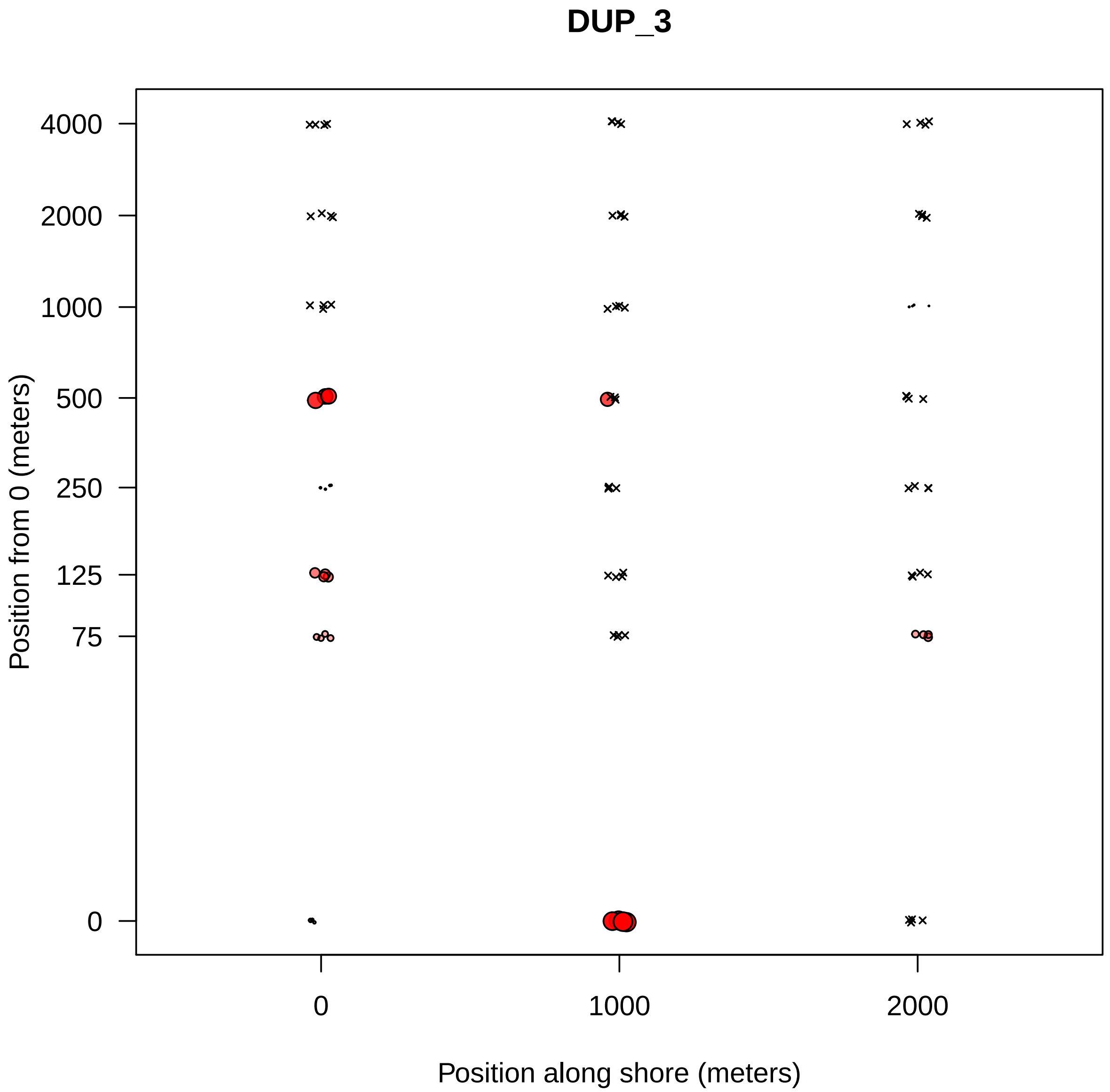

Figure 8: Example of a DNA sequence's spatial distribution. This sequence is annotated to SPECIES X, which is found only in shallow, structured habitats such as patches of *Zostera marina*. Point size and color transparency indicates abundance relative to other DNA sequences from that sample, scaled to the maximum value for this sequence (no fill = 0, full fill = 1). Samples from which this sequence was not recovered are indicated by an "x".