

There once was a grid at ol' Carkeek

First Author<sup>\*1</sup>, Second Author<sup>1,2</sup>, and Third Author<sup>2</sup>

<sup>1</sup>Department of Computer Science, L<sup>A</sup>T<sub>E</sub>X University

<sup>2</sup>Department of Mechanical Engineering, Superfabulous University

April 14, 2016

## 1 Keywords

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

## 3 Abstract

4 This is the text of the abstract.

## 5 Introduction

6 Biodiversity surveillance is being revolutionized by DNA-based detection of organisms from en-  
7 vironmental samples. ?(specifically speed and scope of ecological studies). Many researchers are  
8 justifiably cautious about the ?(adoption) of this new form of data. Their apprehension is rooted  
9 in the premise that traditional survey approaches are more accurate because the chain of inference  
10 between observation and ecological data is usually short: A researcher sees two swans in Lake Hopat-  
11 cong and infers the lake is occupied by at least 2 swans. DNA based surveys, on the other hand,  
12 consist of a longer chain of inference: DNA sequences are reported by a sequencing machine, the  
13 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 PLOS ONE). 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 microbial communities in marine environments because they are subject to dynamic physical forces (CITE). If environmental circulation In general, the relationship between community dissimilarity (0 = identical; 1 = completely different) and spatial distance is expected to be asymptotic, because communities nearer to each other tend to be more similar than those farther apart. The intercept is expected to be near 0, because samples taken at the same place should be very similar. Deviation in the intercept from 0 indicates heterogeneous community composition/structure over fine scales. A flat relationship between dissimilarity and distance indicates that heterogeneity is not assorted spatially, and can be interpreted in different ways, depending on the mean. If the mean is near 1, the spatial heterogeneity has overwhelmed the spatial scale of sampling. If the mean is near 0, there is no community heterogeneity over the scale sampled.

## 72 **Methods**

### 73 **Environmental Sampling**

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

### 77 **Laboratory Methods**

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

83 Pooled libraries were sequenced on the Illumina NextSeq platform at the Stanford Center for  
84 Functional Genomics (machine ID: NS50061; run ID: 115; flowcell ID: H3LFLAFX). Raw sequence  
85 data in fastq format is publicly available (see Data Availability).

### 86 **Data Preparation (Bioinformatics)**

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

89 We calculated rates of cross-library contamination by counting occurrences of primer sequences:  
90 12S primer sequences appearing in a 16S library (and vice versa) indicate an error in the preparation  
91 or sequencing procedures.

92 We assessed PCR contamination by evaluating the dissimilarity of replicate PCRs of the same  
93 DNA sample, and removed one sample for which the Bray-Curtis dissimilarities between itself and  
94 the other replicates exceeded 0.1 (lib\_B\_tag\_GCGCTC).

95 To scale the OTU counts, we calculated the minimum number of OTU-assigned reads (as op-  
96 posed to raw number of reads) found in these samples (130402), multiplied this by within-sample  
97 proportional abundance of each OTU, and finally rounded these numbers.

## 98 Community Analysis

99 We simultaneously assessed the existence of distinct community types and the membership of sam-  
100 ples to those community types using a partitioning around medoids algorithm (CITE PAM, some-  
101 times referred to as k-medoids clustering), as implemented in the R package fpc (CITE fpc). The  
102 classification of samples to communities was made on the basis of their pairwise Bray-Curtis dis-  
103 similarity, calculated using the function vegdist in the R package vegan (CITE VEGAN).

104 We calculated the great circle distance between points using the Haversine method as imple-  
105 mented by the R package geosphere (CITE geosphere).

## 106 Spatial Model Formulation

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

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

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

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

$$\lambda_{ijk} = \beta_0 + \beta_i + \eta_{ijk} + \epsilon_{ijk} \quad (2)$$

118 Where  $\beta_0$  is a general intercept across all taxa,  $\beta_i$  is a fixed effect accounting for the variance  
119 associated with taxon  $i$ , and  $\eta_{ijk}$  and  $\epsilon_{ijk}$  are random effects of variance resulting from the processes  
120 associated with PCR and spatial location, respectively.

## 121 **Results**

### 122 **Data Quality (Bioinformatics)**

123 All value ranges are reported as (mean  $\pm$  standard deviation).

124 There was a very low frequency of cross-contamination from other libraries into those reported here  
125 ( $5\text{e-}05 \pm 8\text{e-}05$ ; max 0.00034)

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

### 132 **Community Analysis**

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

### 135 **Spatial Model Output**

## 136 **Discussion**

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

## 138 **Acknowledgements**

139 We wish to thank all of the little people.

## 140 **Funding**

141 This study was funded by our super-rich uncle.

## 142 **Author Contributions**

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

## 147 **Data Availability**

148 All sequence files and metadata are available from EMBL:

149 <http://www.ebi.ac.uk/ena/data/view/XXXXXXXXX>

150 All analyses were performed using scripts available from the project repository on GitHub:

151 [https://github.com/jimmyodonnell/Carkeek\\_eDNA\\_grid](https://github.com/jimmyodonnell/Carkeek_eDNA_grid)

152

## 153 **Figures**

## 154 **Supplemental Material**

## 155 **Bioinformatic Methods**

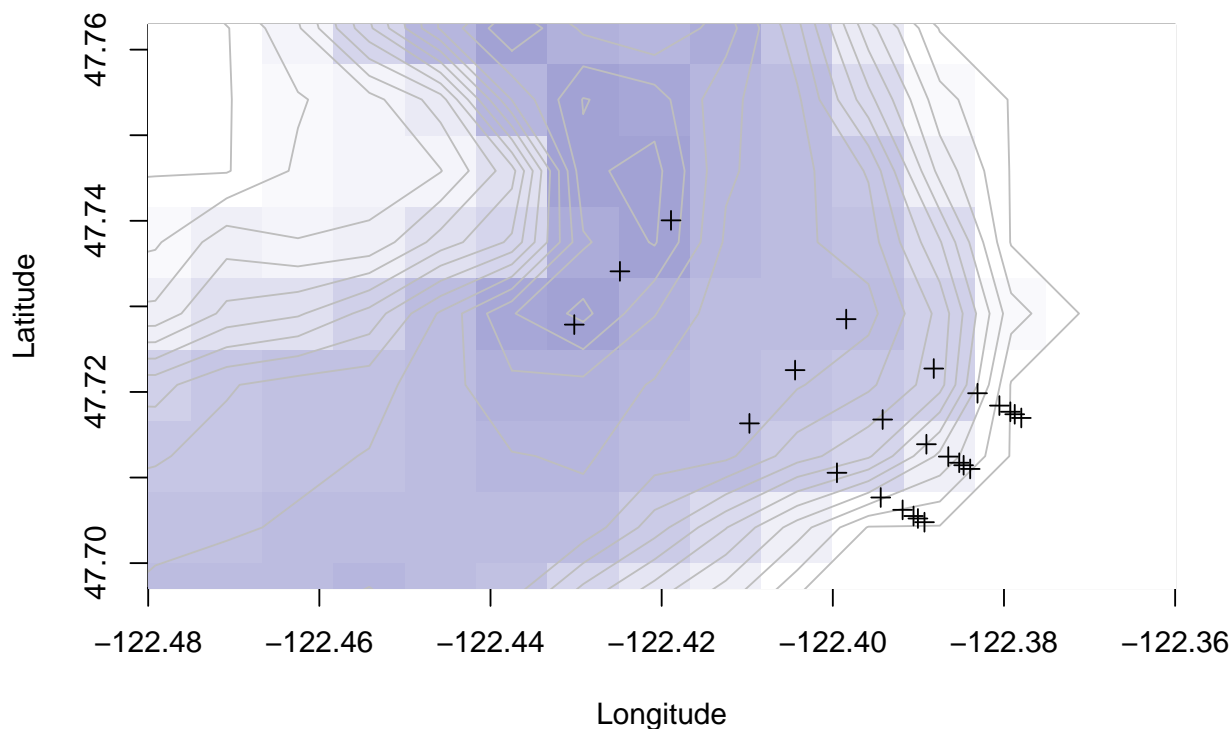


Figure 1: TODO: Plot with GEBCO 30-second data or remove grid coloring and color by isobath. Looking into filling by contour. Geographic position of collected samples. Lines give XXX meter isobaths.

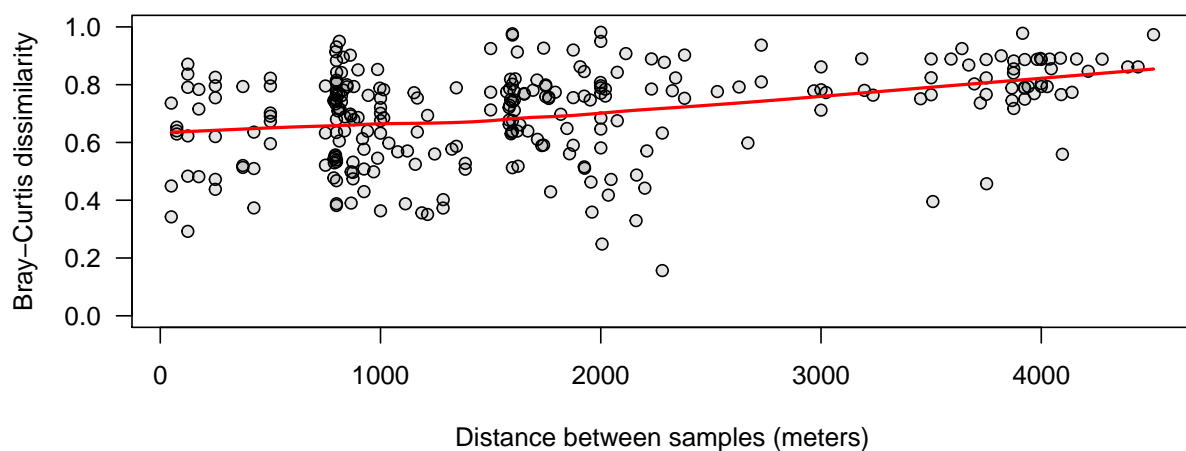


Figure 2: Pairwise Bray-Curtis dissimilarity of eDNA communities plotted against pairwise spatial distance. Line represents prediction of Gaussian LOESS (degree = 1; span = 2/3). Restricting comparison to within-transect has no qualitative difference in the outcome (see 'diss\_by\_dist\_by\_transect.pdf').



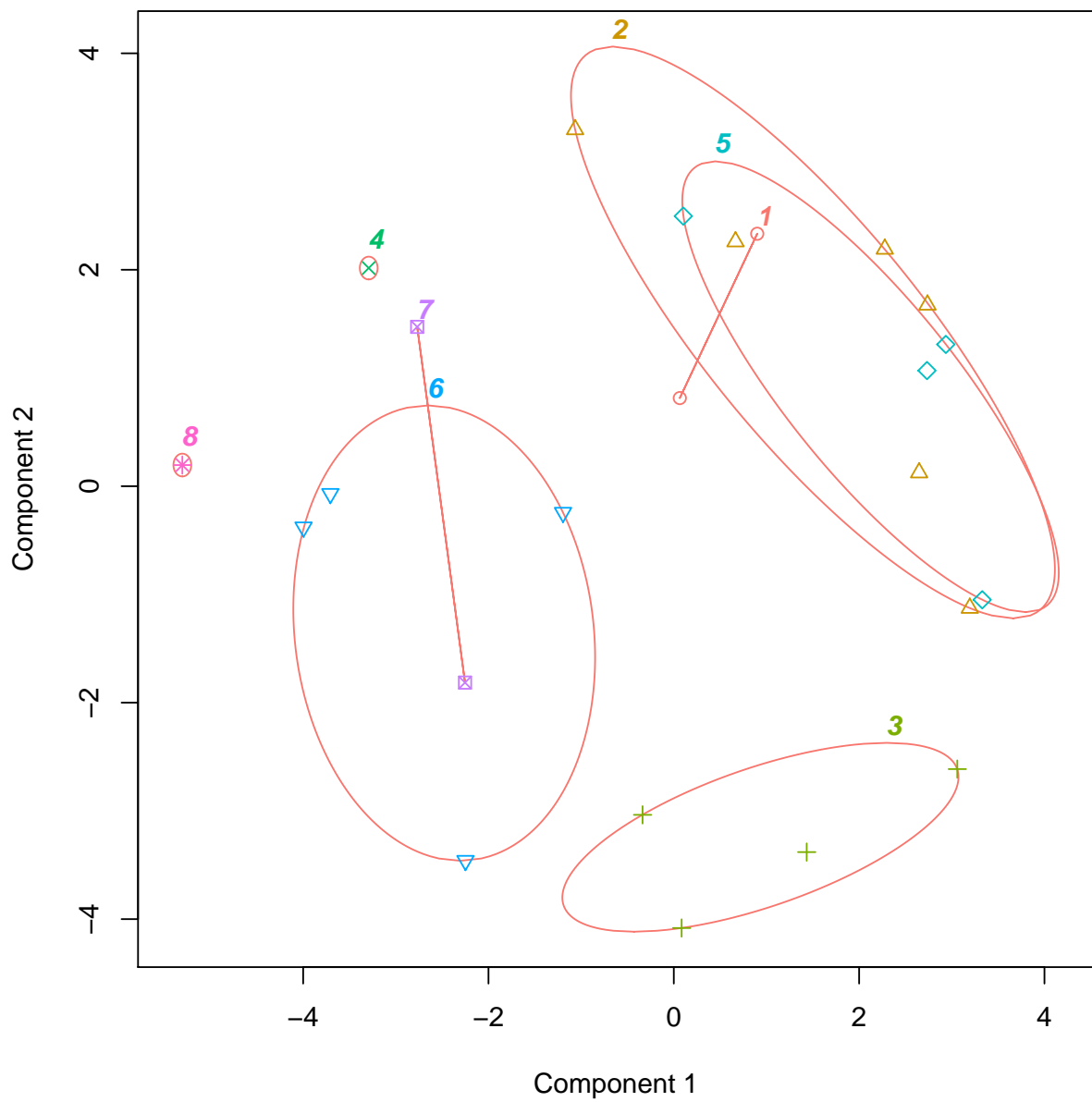


Figure 3: TODO figure out color of ellipses; I can't even plot them gray without Plot of partitioning around medoids (PAM) analysis of OTU sequence abundance from 4 replicate PCRs at each of 24 sampling points. Points represent communities of OTUs; color and shape indicate cluster membership as determined by PAM analysis. Ellipses indicate the smallest area of a cluster that contains all of its members.

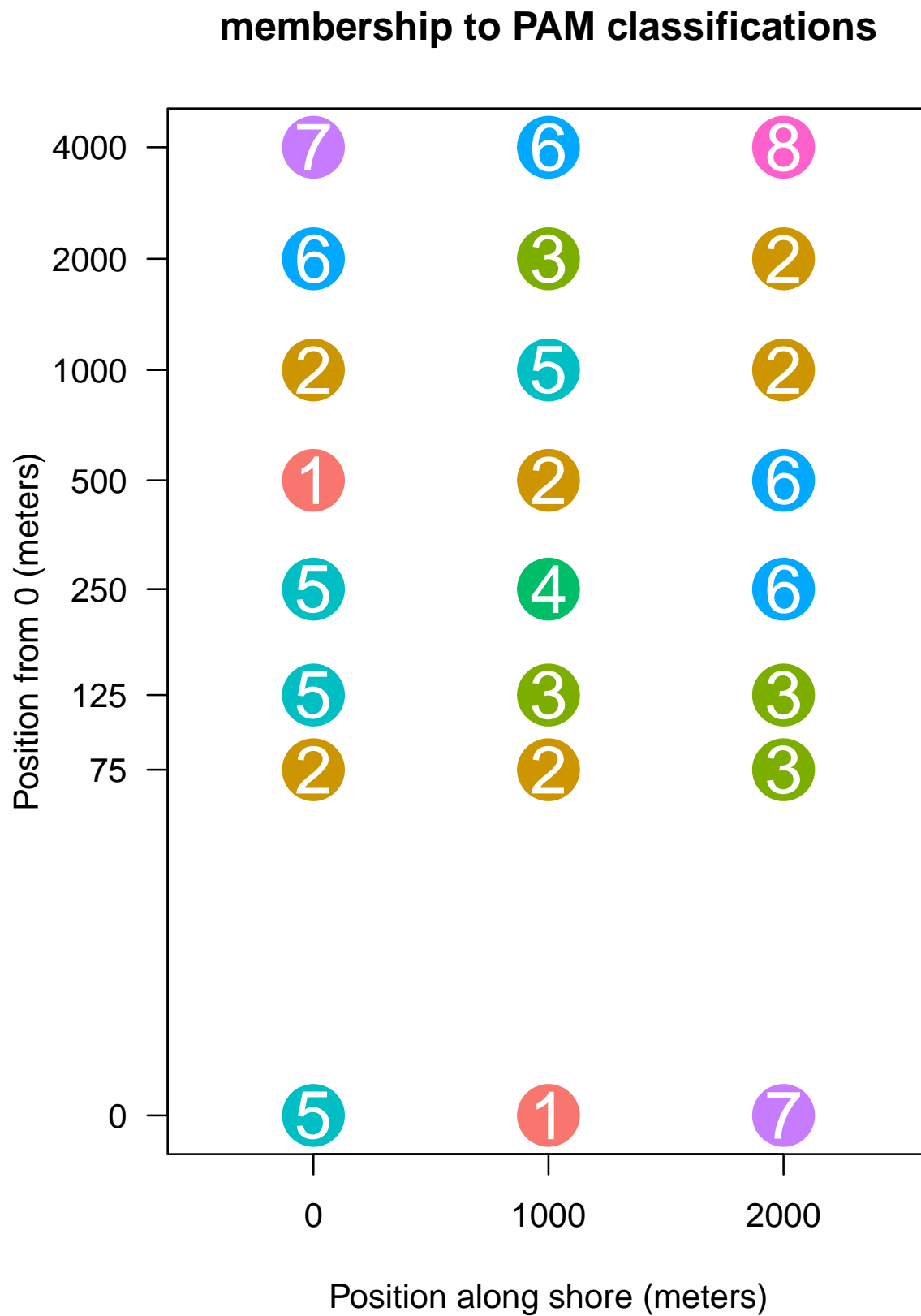


Figure 4: Geographic position of collected samples, colored by membership to clusters identified by partitioning around medoids algorithm.

../../../../Figures/slope\_plots.pdf

Figure 5: Fit lines of DNA sequence counts as a function of distance from shore for a selection of taxa for which we have strong preconceived expectations (left). Box plots of the estimates of the slopes for taxa (100 most abundant), grouped by life history traits (right).



Figure 6: Box plots of estimates of variance associated with each level of the multilevel model, corresponding to stages of the eDNA sampling protocol.

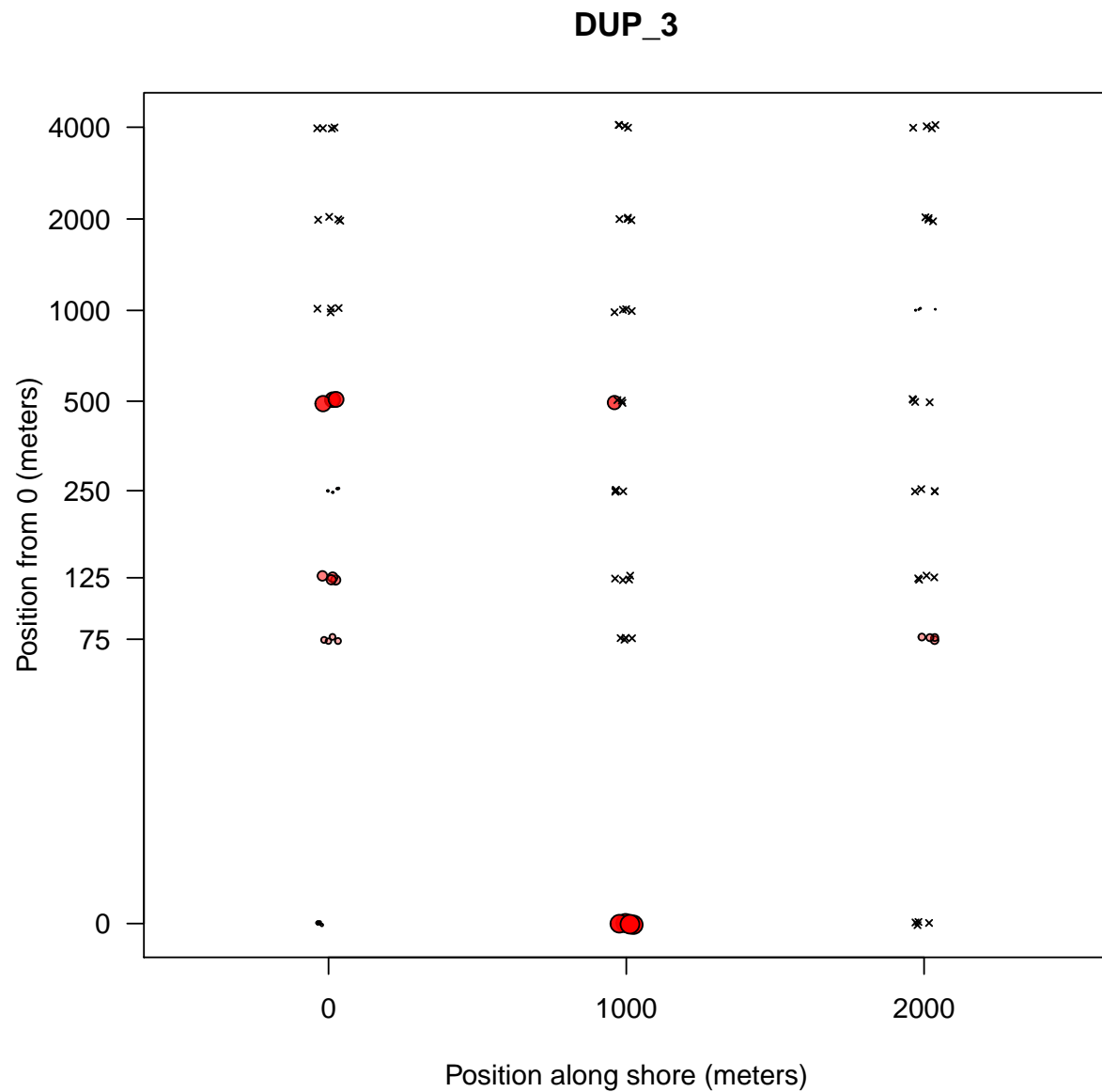


Figure 7: Example of a DNA sequence's spatial distribution. This sequence is annotated to SPECIES X, which is found only in shallow, structured habitats such as patches of *Zostera marina*. Point size and color transparency indicates abundance relative to other DNA sequences from that sample, scaled 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".