

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 22, 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 envi-  
7 ronmental samples. (specifically speed and scope of ecological studies). While this approach has  
8 been used for decades to survey micro-organisms (CITE VENTER), it has more recently become  
9 used as a tool for surveying macro-organisms (CITE EARLY EDNA). The technique is founded  
10 on the premises that: 1) organisms shed DNA-containing material into the environment, 2) the  
11 DNA contained in this material breaks down over time, which leads to the conclusion that: (1) the  
12 concentration of DNA originating from an organism should decrease with distance from its source.

---

\*first.author@funstuff.com

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

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

31 Many researchers are justifiably cautious about the ?(adoption) of this new form of data. Their  
32 apprehension is rooted in the premise that traditional survey approaches are more accurate because  
33 the chain of inference between observation and ecological data is usually short: A researcher sees two  
34 swans in Lake Hopatcong and infers the lake is occupied by at least 2 swans. DNA based surveys, on  
35 the other hand, consist of a longer chain of inference: DNA sequences are reported by a sequencing  
36 machine, the machine identifies the sequence of products of a polymerase chain reaction (PCR),  
37 PCR amplifies pieces of DNA from a purified genomic DNA sample, DNA is purified (extracted)  
38 from an environmental sample, environmental samples contain DNA from organisms present, the  
39 organisms present are representative of the biological community about which we wish to make  
40 inference. ?( reverse order? tie to concrete example (swans of Lake Hopatcong)). Clearly, this  
41 process is more complex than visual surveys, as the relationship between several steps is complex or  
42 unknown. But consider that the processes ?(behind | underlying) other more widely-used ecological

43 survey techniques are similarly complex, such as bird surveys based on song, or visual identification  
44 of fungal spores. When alternate survey approaches are impossible or inefficient, we are more  
45 willing to accept any available survey data, regardless of the complexity or uncertainty underlying  
46 it. (microbiologists have enthusiastically relied on DNA-based surveys for years for this reason,  
47 (though yes, they also do not have the problem of disconnect between individual and cell)).

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

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

70 Here, we apply this framework to a DNA survey of (nearshore | coastal) marine habitat. (TODO  
71 add commentary on current dogma surrounding distribution of DNA in well-mixed (marine) habi-  
72 tats). 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 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. Likewise, dissimilarity cannot exceed 1. Deviation 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 close to 1, there is high spatial heterogeneity over the spatial scale of sampling. If the mean is 0, all samples are identical, 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.

## 98 **Methods**

### 99 **Environmental Sampling**

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

### 103 **Laboratory Methods**

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

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

### 112 **Data Preparation (Bioinformatics)**

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

115 We calculated rates of cross-library contamination by counting occurrences of primer sequences:  
116 12S primer sequences appearing in a 16S library (and vice versa) indicate an error in the preparation  
117 or sequencing procedures.

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

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

124 We compiled life history data for taxonomic groups at the family level and higher. For groups  
125 with sparse data at the family level (e.g. Nemertea), we used the data from the higher level group.

## 126 **Community Analysis**

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

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

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

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

$$com \sim V_{max}[d]/K_m[d] \quad (1)$$

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

## 151 Spatial Model Formulation

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

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

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

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

$$\lambda_{ijk} = \beta_0 + \beta_i + \eta_{ijk} + \epsilon_{ijk} \quad (3)$$

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

## 166 Results

### 167 Data Quality (Bioinformatics)

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

169 There was a very low frequency of cross-contamination from other libraries into those reported here  
170 (5e-05 $\pm$ 8e-05; max 0.00034)

171 We assessed the consistency of PCR by conducting 4 replicate PCRs for each environmental  
172 sample and calculating the mean pairwise Bray-Curtis dissimilarity of the resulting communities

173 (scaled to minimum read depth per sample). 92 of the 96 amplicon samples had mean Bray-Curtis  
174 dissimilarity  $\leq 0.052$ ; 1 sample had a value of 0.341, which elevates the value of the other replicates.  
175 After removal of this sample, the highest mean Bray-Curtis dissimilarity among replicates within  
176 an environmental sample was 0.034.

## 177 **Community Analysis**

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

180 The estimated asymptote of community dissimilarity as a function of spatial distance ( $V_m$ ) was  
181 0.72 ( $p \ll 0.05$ ), and the distance at which half this dissimilarity was accumulated ( $K_m$ ) is 23.8  
182 kilometers ( $p = 0.006$ ). Residual standard error of the fit of the model is 0.1563 on 274 degrees of  
183 freedom.

## 184 **Spatial Model Output**

## 185 **Discussion**

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

## 187 **Acknowledgements**

188 We wish to thank all of the little people.

## 189 **Funding**

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

## 191 **Author Contributions**

192 Conceived and designed the experiments: James L. O'Donnell, Ryan P. Kelly, A. Ole Shelton.

193 Collected the data: James L. O'Donnell, Greg Williams, Natalie C. Lowell, Ryan P. Kelly, A. Ole



194 Shelton, Jameal F. Samhouri. Conducted the analyses: . Wrote the first draft: . Edited the  
195 manuscript: .

## 196 **Data Availability**

197 All sequence files and metadata are available from EMBL:

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

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

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

201

## 202 **Figures**

## 203 **Supplemental Material**

## 204 **Bioinformatic Methods**

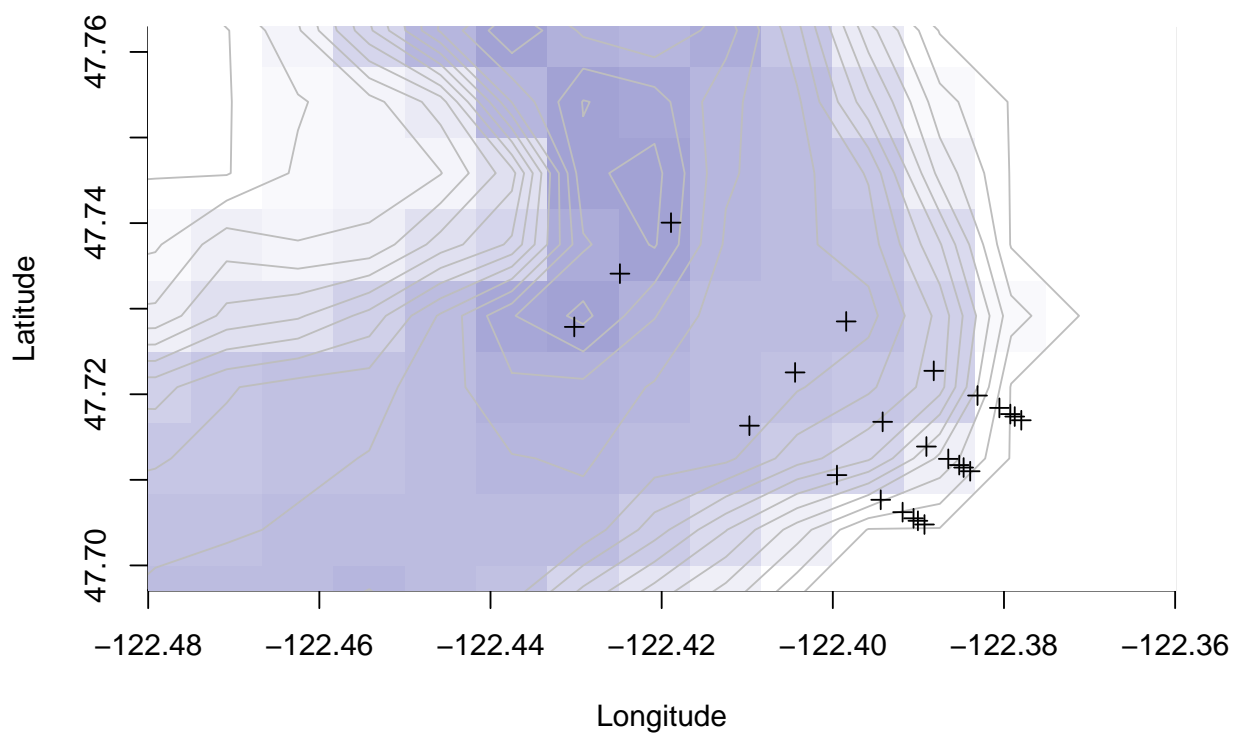


Figure 1: TODO: Plot with GEBCO 30-second data or remove grid coloring and color by isobath. Looking into filling by contour. Geographic position of collected samples. Lines give XXX meter isobaths.

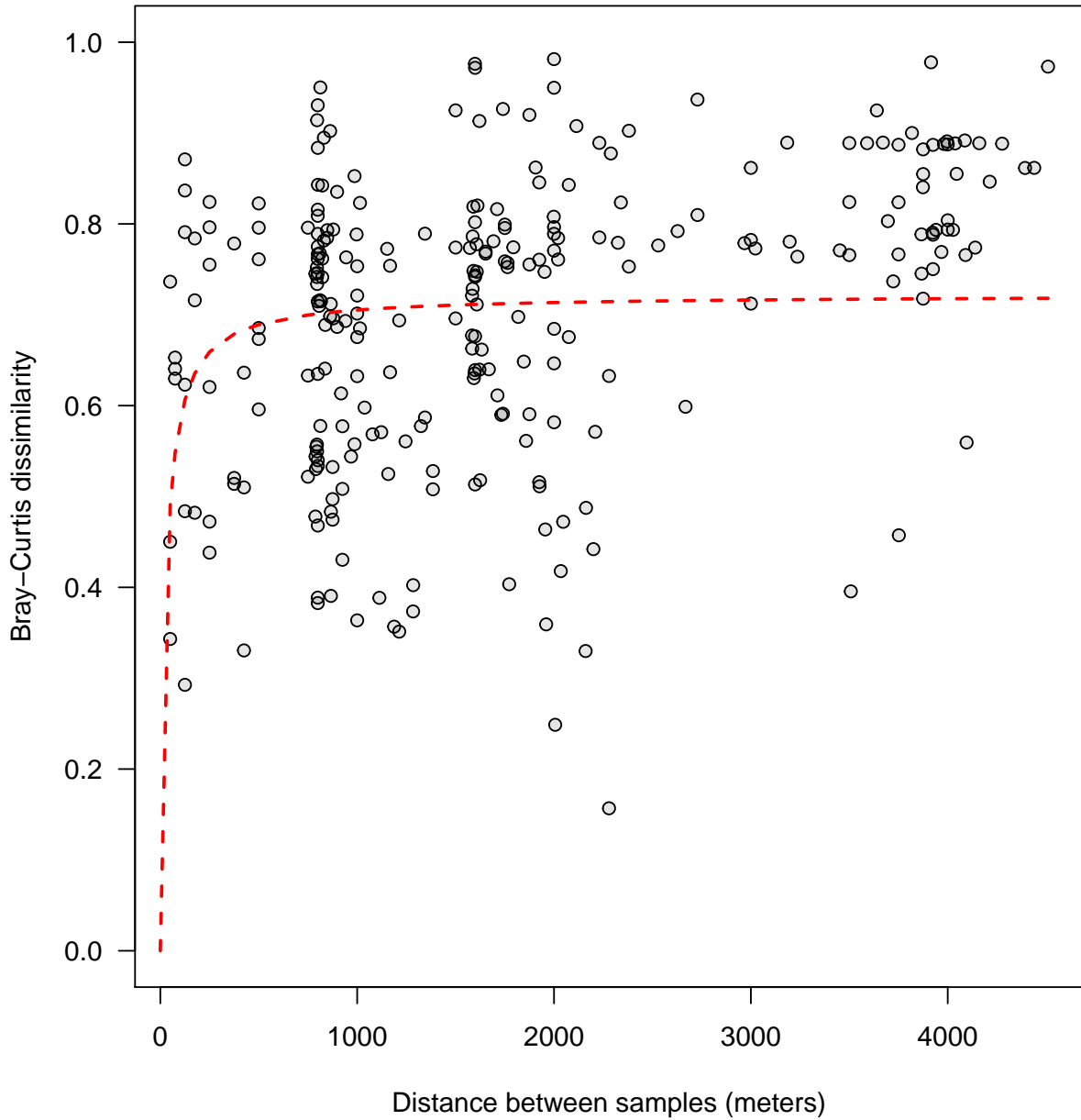


Figure 2: Pairwise Bray-Curtis dissimilarity of eDNA communities plotted against pairwise spatial distance. Line represents prediction of the Non-linear Least Squares regression to a Michaelis-Menten model ( $V_m = 0.72$ ,  $p \ll 0.05$ ;  $K_m = 23.8$  kilometers,  $p = 0.006$ ;  $RSE = 0.1563$ ;  $df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

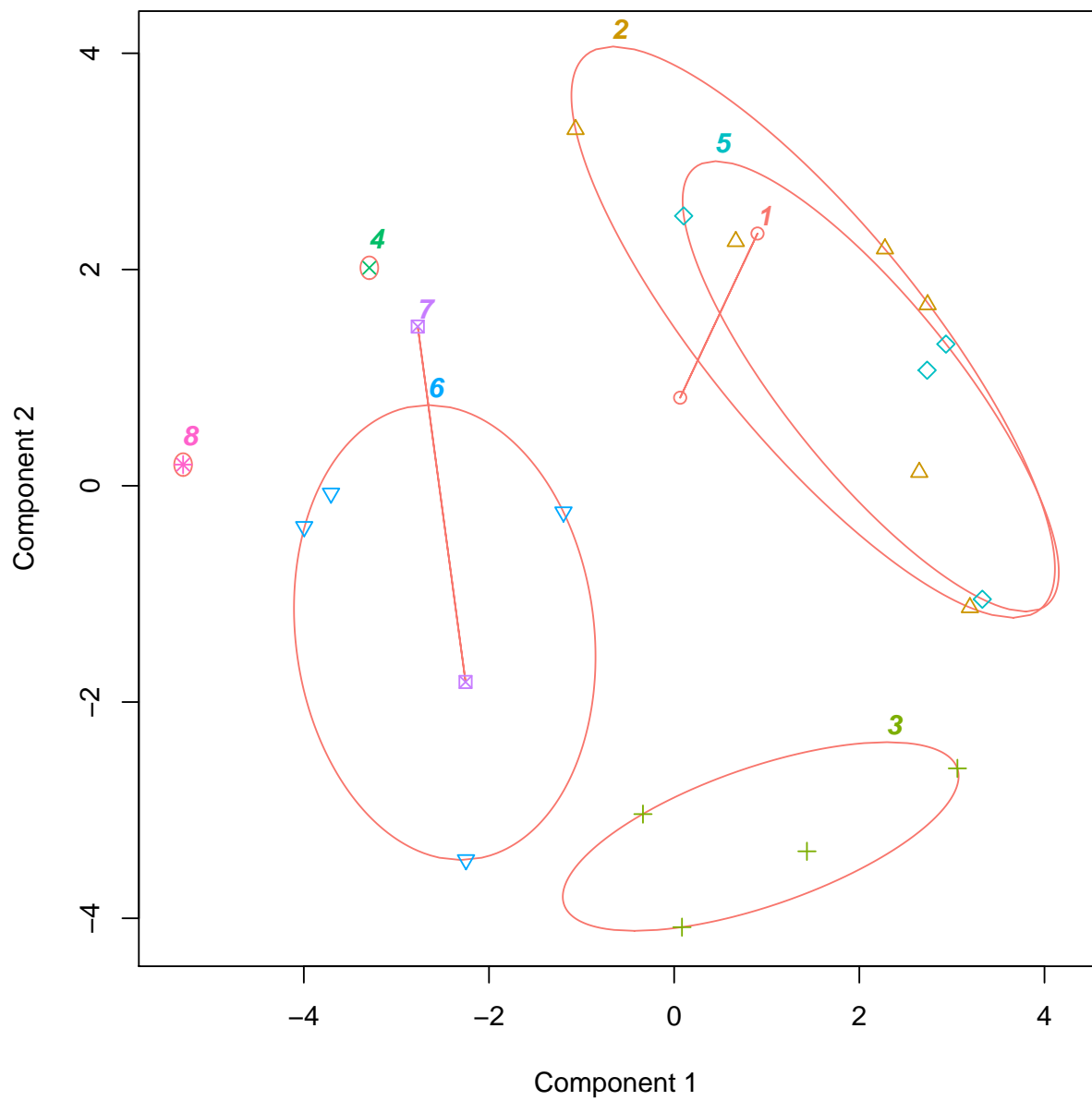


Figure 3: TODO figure out color of ellipses; I can't even plot them gray without Plot of partitioning around medoids (PAM) analysis of OTU sequence abundance from 4 replicate PCRs at each of 24 sampling points. Points represent communities of OTUs; color and shape indicate cluster membership as determined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

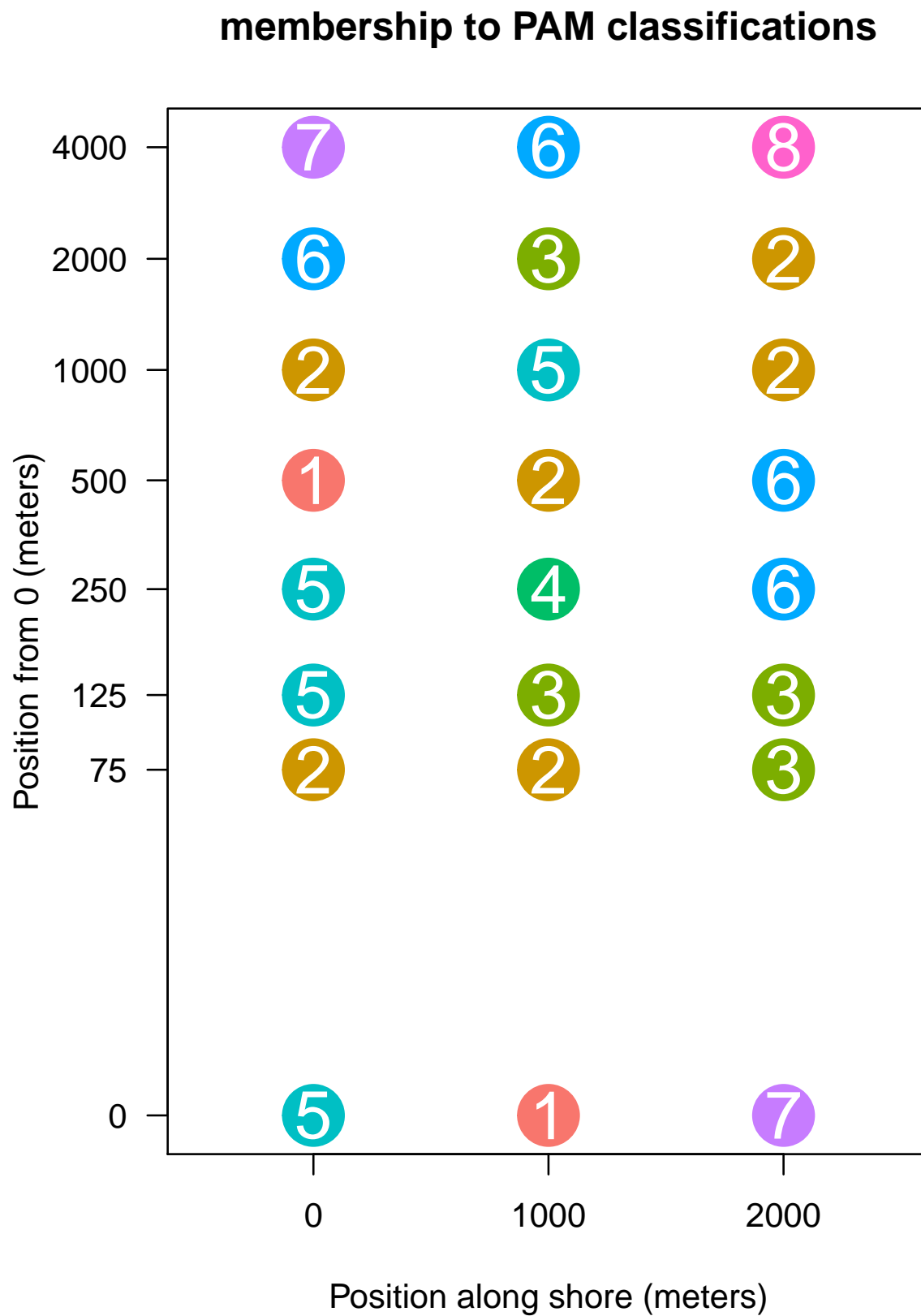


Figure 4: Geographic position of collected samples, colored by membership to clusters identified by partitioning around medoids algorithm.

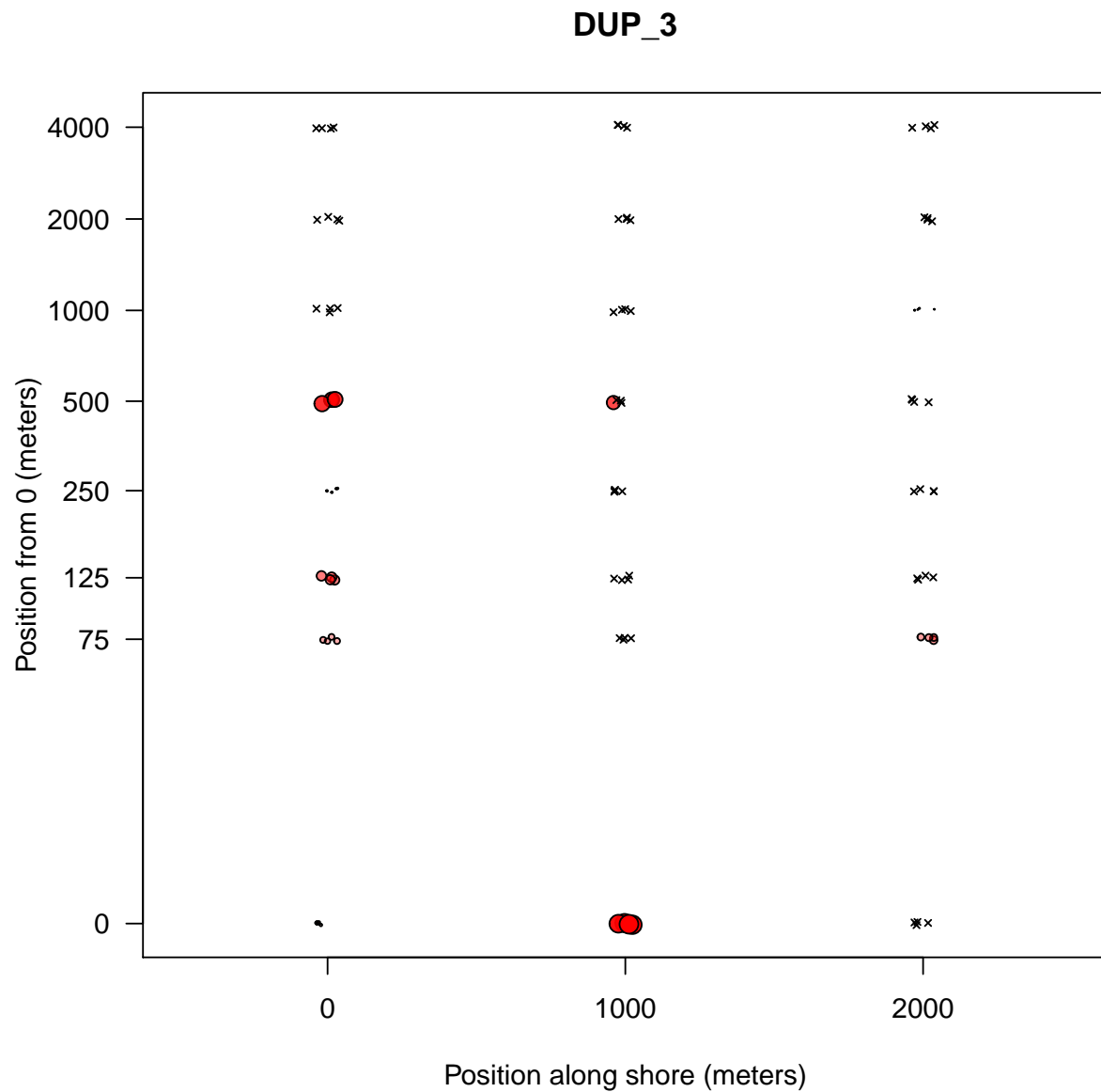


Figure 5: Example of a DNA sequence's spatial distribution. This sequence is annotated to SPECIES X, which is found only in shallow, structured habitats such as patches of *Zostera marina*. Point size and color transparency indicates abundance relative to other DNA sequences from that sample, scaled 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".