

There once was a grid at ol' Carkeek

First Author^{*1}, Second Author^{1,2}, and Third Author²

¹Department of Computer Science, L^AT_EX University

²Department of Mechanical Engineering, Superfabulous University

March 29, 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

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

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

38 Shelton et al. (CITE Shelton 2016) outlined an approach for structuring statistical models
39 of DNA surveys that address these issues. This framework improved on alternative statistical
40 techniques by explicitly accounting for the ?(hierarchical | nested | multilevel) structure of the
41 study design, which allows error and uncertainty at each level to be ?(explicitly accounted for|
42 modeled | propagated throughout the model). That study demonstrated an improvement in the
43 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.

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 Methods

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

71 reported by (CITE PORT 2015).

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

75 Data Preparation (Bioinformatics)

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

78 We calculated rates of cross-library contamination by counting occurrences of primer sequences:
79 12S primer sequences appearing in a 16S library (and vice versa) indicate an error in the preparation
80 or sequencing procedures.

81 Community Analysis

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

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

89 Spatial Model Formulation

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

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

97 values are discrete counts, we use the natural exponent, e^λ . Thus,

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

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

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

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

104 Results

105 Data Quality (Bioinformatics)

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

107 There was a very low frequency of cross-contamination from other libraries into those reported here
108 ($5\text{e-}05 \pm 8\text{e-}05$; max 0.00034)

109 Community Analysis

110 Spatial Model Output

111 Discussion

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

113 Acknowledgements

114 We wish to thank all of the little people.

115 **Funding**

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

117 **Author Contributions**

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

122 **Data Availability**

123 All sequence files and metadata are available from EMBL:

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

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

126 https://github.com/jimmyodonnell/Carkeek_eDNA_grid

127

128 **Figures**

129 **Supplemental Material**

130 **Bioinformatic Methods**

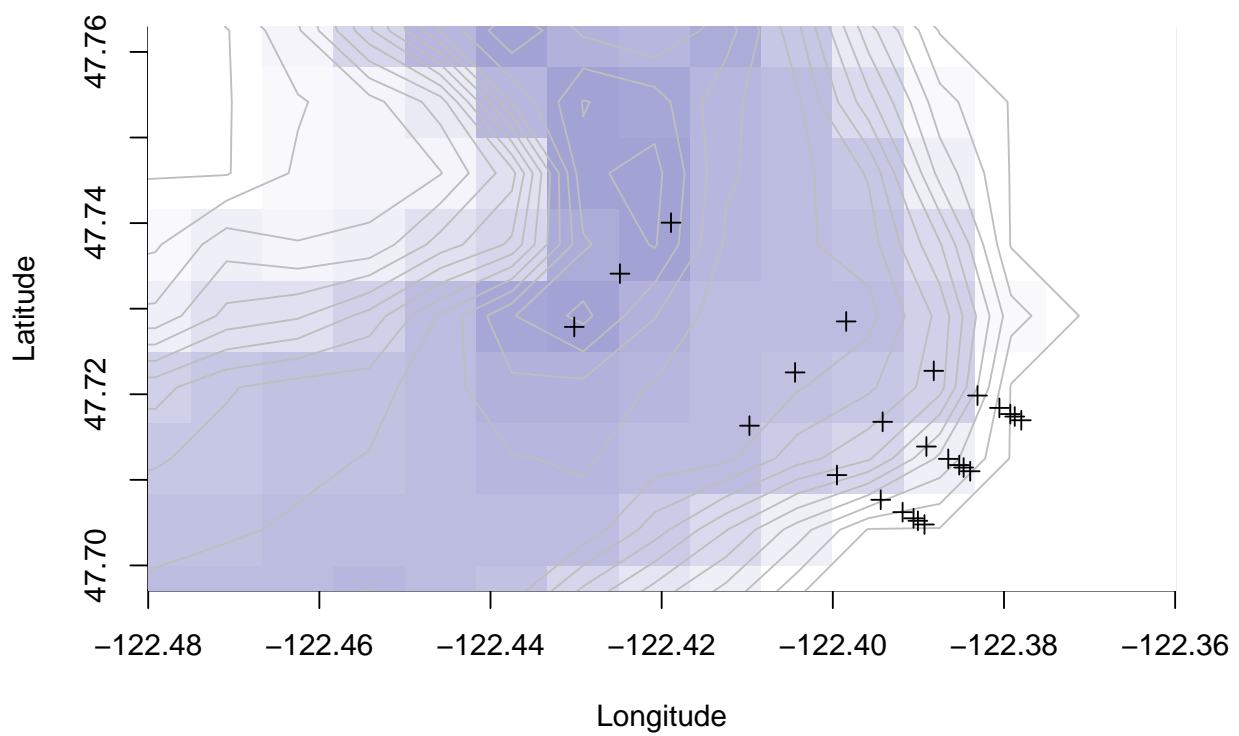


Figure 1: TODO: Plot with GEBCO 30-second data or remove grid coloring and color by isobath. Looking into filling by contour. Geographic position of collected samples. Lines give XXX meter isobaths.

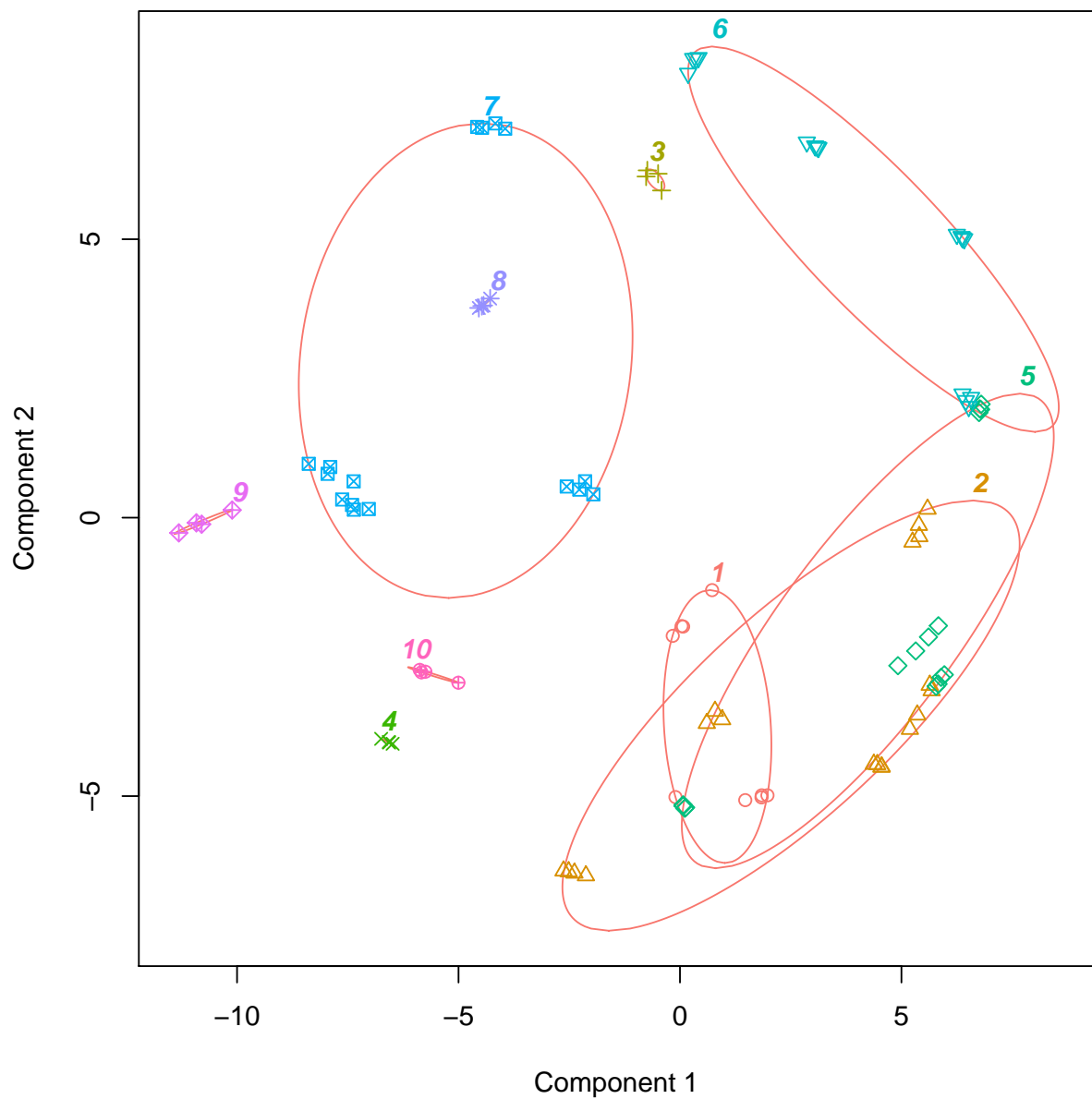


Figure 2: TODO figure out color of ellipses; I can't even plot them gray without Plot of partitioning around medoids (PAM) analysis of OTU sequence abundance from 4 replicate PCRs at each of 24 sampling points. Points represent communities of OTUs; color and shape indicate cluster membership as determined by PAM analysis. Ellipses indicate the smallest area of a cluster that contains all of its members.

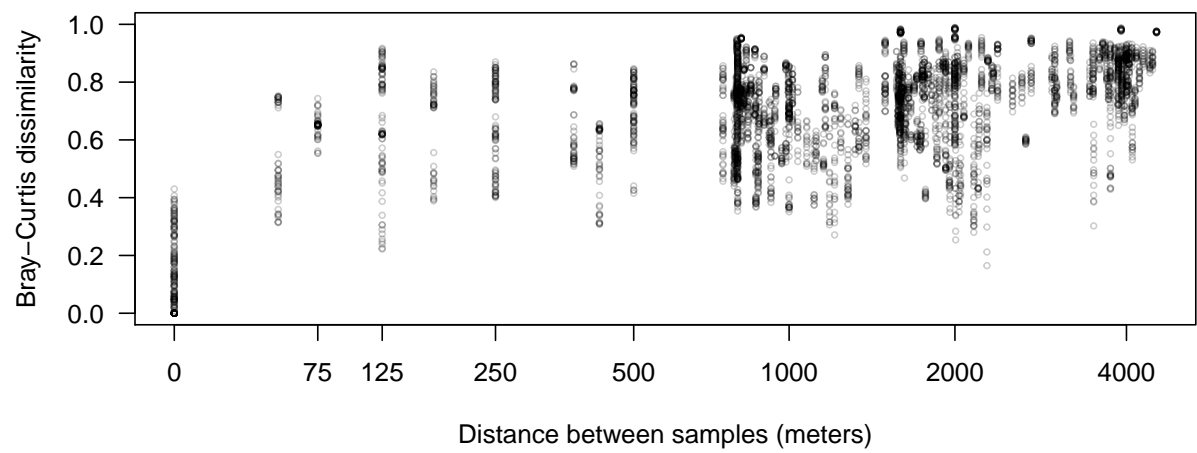


Figure 3: Pairwise Bray-Curtis dissimilarity of eDNA communities plotted against pairwise spatial distance.

../../../../Figures/slope_plots.pdf

Figure 4: 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).

../../../../Figures/var_boxplots.pdf

Figure 5: Box plots of estimates of variance associated with each level of the multilevel model, corresponding to stages of the eDNA sampling protocol.

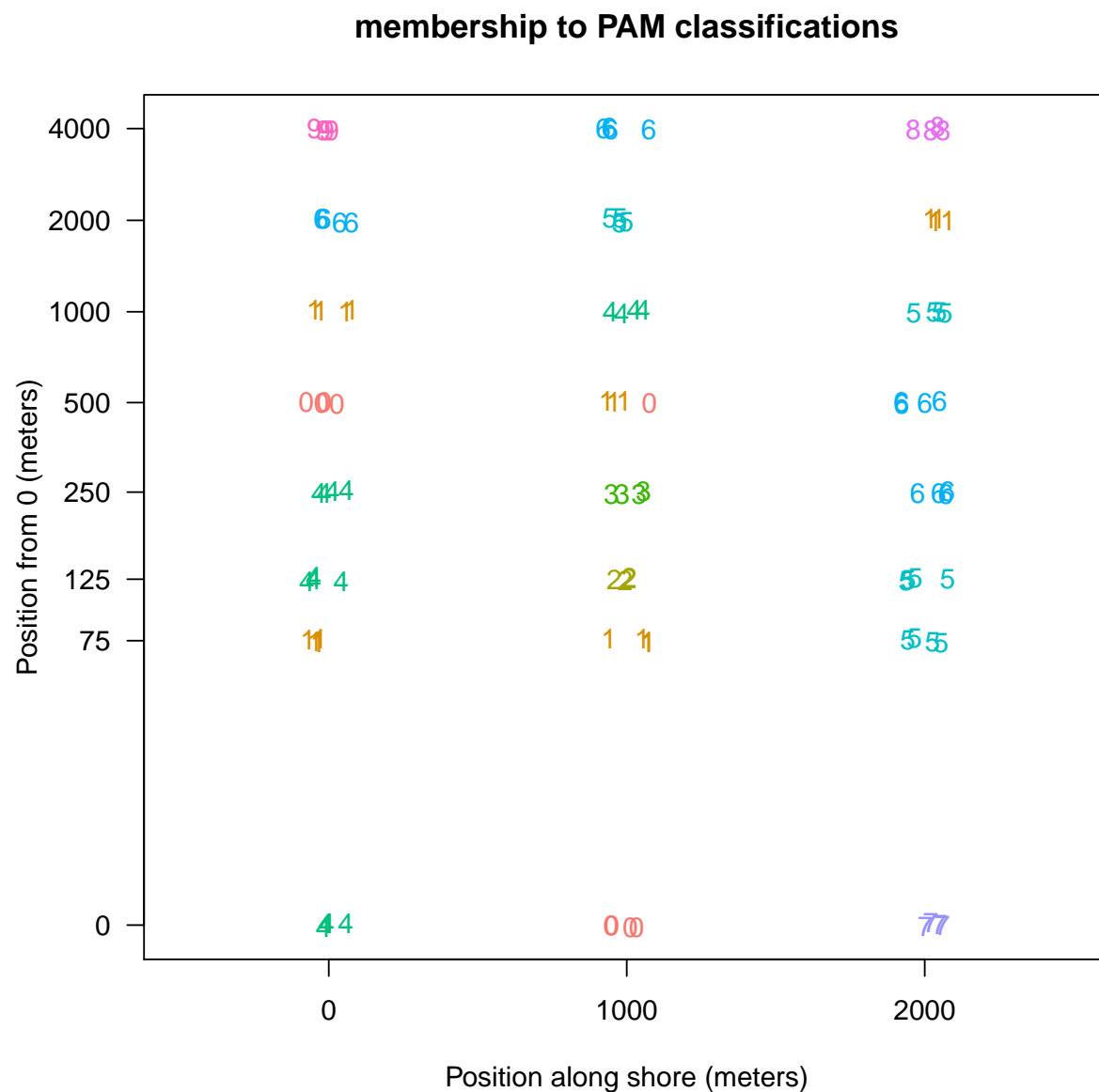


Figure 6: Geographic position of collected samples, colored by membership to clusters identified by partitioning around medoids algorithm. Points are jittered in both horizontal and vertical dimension to distinguish among four replicate PCR products sequenced from each environmental sample.

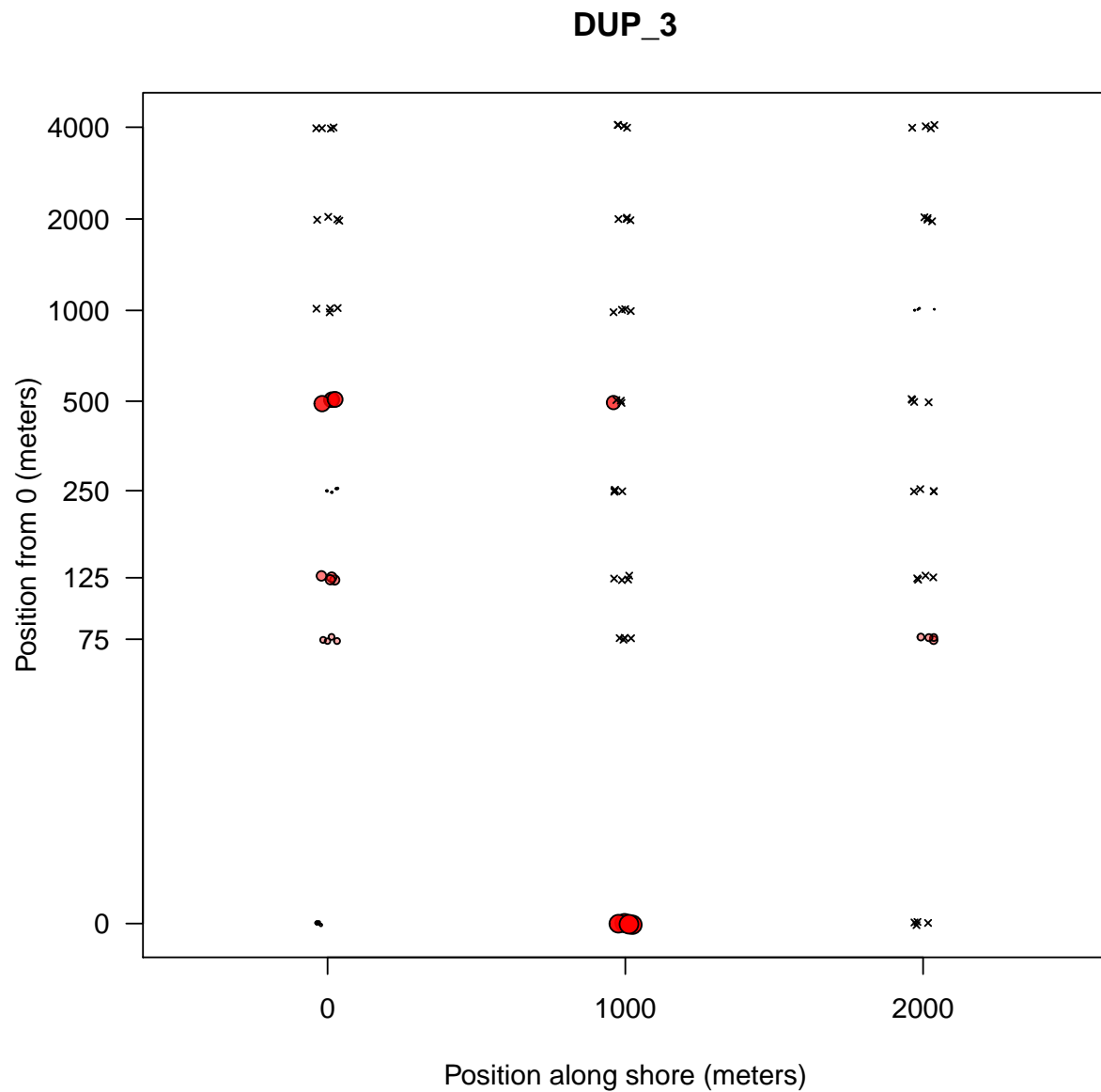


Figure 7: Example of a DNA sequence's spatial distribution. This sequence is annotated to SPECIES X, which is found only in shallow, structured habitats such as patches of *Zostera marina*. Point size and color transparency indicates abundance relative to other DNA sequences from that sample, scaled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