

# Spatial distribution of environmental DNA in a nearshore marine habitat

James L. O'Donnell<sup>\*1</sup>, Ryan P. Kelly<sup>1</sup>, Andrew O. Shelton<sup>2</sup>, Jameal F. Samhouri<sup>2</sup>, Natalie C. Lowell<sup>1,3</sup>, and Gregory D. Williams<sup>4</sup>

<sup>1</sup>School of Marine and Environmental Affairs, University of Washington, 3707 Brooklyn Ave NE, Seattle, Washington 98105, USA

<sup>2</sup>Northwest Fisheries Science Center, NOAA Fisheries, 2725 Montlake Blvd E, Seattle, Washington 98112, USA

<sup>3</sup>School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat St, Seattle, Washington 98105, USA

<sup>4</sup>Pacific States Marine Fisheries Commission, Under contract to the Northwest Fisheries Science Center, NOAA Fisheries, 2725 Montlake Blvd E, Seattle, WA 98112

November 15, 2016

## 1 **Keywords**

2 metagenomics, metabarcoding, environmental monitoring, molecular ecology, marine, estuarine

## 3 **Abstract**

4 In the face of increasing threats to biodiversity, the advancement of methods for surveying biological  
5 communities is a major priority for ecologists. Recent advances in molecular biological technologies  
6 have made it possible to detect and sequence DNA from environmental samples (environmental DNA

---

<sup>\*</sup>jodonnellbio@gmail.com

or eDNA); however, eDNA techniques have not yet seen widespread adoption as a routine method for biological surveillance primarily due to gaps in our understanding of the dynamics of eDNA in space and time. In order to identify the effective spatial scale of this approach in a dynamic marine environment, we collected marine surface water samples from transects ranging from the intertidal zone to 4 kilometers from shore. Using massively parallel sequencing of 16S amplicons, we identified a diverse community of metazoans and quantified their spatial patterns using a variety of statistical tools. We find evidence for multiple, discrete eDNA communities in this habitat, and show that these communities decrease in similarity as they become further apart. Offshore communities tend to be richer but less even than those inshore, though diversity was not spatially autocorrelated. Taxon-specific relative abundance coincided with our expectations of spatial distribution in taxa lacking a microscopic, pelagic life-history stage, though most of the taxa detected do not meet these criteria. Finally, we use carefully replicated laboratory procedures to show that laboratory treatments were remarkably similar in most cases, while allowing us to detect a faulty replicate, emphasizing the importance of replication to metabarcoding studies. While there is much work to be done before eDNA techniques can be confidently deployed as a standard method for ecological monitoring, this study serves as a first analysis of diversity at the fine spatial scales relevant to marine ecologists and confirms the promise of eDNA in dynamic environments.

## 24 Introduction

No existing biodiversity survey method completely censuses all of the organisms in a given area. For example, towed fishing nets can efficiently sample organisms larger than the mesh and slower than the boat; but overlook viruses and have undesirable effects on charismatic air-breathing species. From a boat or aircraft, scientists can count whales by sight, but not the krill on which they feed. However, DNA-based surveys show great promise as an efficient technique for detecting a previously unthinkable breadth of organisms from a single sample.

Microbiologists have used nucleic acid sequencing to quantify the composition and function of microbial communities in a wide variety of habitats (Handelsman et al., 1998; Tyson et al., 2004; Venter et al., 2004; Iverson et al., 2012). To do so, a sample of environmental medium (e.g. water) containing microorganisms is collected, their DNA or RNA is isolated and sequenced, and the

identity and abundance of sequences is considered to reflect the community of organisms contained in the sample, which indirectly estimates the quantity of organisms in an area.

Macroorganisms shed DNA-containing cells into the environment (environmental DNA or eDNA) that can be sampled in the same way (Ficetola et al., 2008; Thomsen et al., 2012). Potentially, eDNA methods allow a broad swath of macroorganisms to be surveyed from basic environmental samples. However, the accuracy and reliability of indirect estimates of macroorganismal abundance has been debated because the entire organisms are not contained within the sample (Coward et al., 2015). Concern surrounding eDNA methods is rooted in uncertainty about the attributes of eDNA in the environment relative to actual organisms (Shelton et al., 2016; Evans et al., 2016). Basic questions such as how long DNA can persist in that environment and how far DNA can travel remain largely unknown (but see Klymus et al. (2015); Turner et al. (2015); Strickler et al. (2015); Deiner and Altermatt (2014)) and impede inference about local organismal presence from an environmental sample. As a result, estimating the spatial and temporal resolution of eDNA studies in the field is a key step in making these methods practical.

The relationship between local organismal abundance and eDNA is further complicated in habitats where the environmental medium itself may transport eDNA away from its source. We know that genetic material moves away from its source precisely because organisms can be detected indirectly without being present in the sample (Kelly et al., 2016). One might reasonably expect eDNA to travel farther in a highly dynamic fluid such as the open ocean or flowing river than it would through the sediment at the bottom of a stagnant pond (Deiner and Altermatt, 2014; Shogren et al., 2016). Yet even studies of extremely dynamic habitats such as coastlines with high wave energy have found remarkable evidence that eDNA transport is limited enough that DNA methods can detect differences among communities separated by less than 100 meters (Port et al., 2016).

While rigorous laboratory studies have investigated the effects of some environmental factors on eDNA persistence (Klymus et al., 2015; Barnes et al., 2014; Sassoubre et al., 2016) and the transport of eDNA in specific contexts (Deiner and Altermatt, 2014), we suggest that field studies comparing the spatial distribution of communities of eDNA with expectations based on prior knowledge of organisms' distributions are also critical to developing a working understanding of eDNA in the real world.

Species and the communities they comprise are not homogeneously distributed in space; describ-

ing and predicting the spatial heterogeneity of ecosystems is of great interest to ecologists. There is a large and active literature focused on understanding the patterns and causes of community variability in space (Hubbell, 2001; Anderson et al., 2011). One consistently observed pattern of community spatial heterogeneity is that communities close to one another tend to be more similar than those that are farther apart (Nekola and White, 1999). This decrease in community similarity with increasing spatial separation is called distance decay and has been reported from communities of tropical trees (Condit, 2002; Chust et al., 2006), ectomycorrhizal fungi (Bahram et al., 2013), salt marsh plants (Guo et al., 2015), and microorganisms (Martiny et al., 2011; Chust et al., 2013; Wetzel et al., 2012; Bell, 2010). Typically, this relationship is assessed by regressing a measure of community similarity against a measure of spatial separation for a set of sites at which a set of species' abundances (or presences) is calculated.

We can apply these methods derived from community ecology to understand spatial patterns and patchiness of eDNA. The underlying mechanism thought to drive the slope of this relationship is the rate of movement of individuals among sites, which may be driven by underlying processes such as habitat suitability. Because eDNA is shed and transported away from its source, the increased movement of eDNA particles should homogenize community similarity, and thus erode the distance decay relationship of eDNA communities.

Puget Sound is a deep, narrow fjord in Washington, USA, where a narrow band of shallow bottom hugs the shoreline and abruptly gives way to a central depth of up to 300 meters. This form allows the juxtaposition of communities associated with distinctly different habitats: shallow, intertidal benthos, and euphotic pelagic (Burns, 1985). At the upper reaches of the intertidal, the shoreline substrate varies from soft, fine sediment to cobble and boulder rubble. Soft intertidal sediments are inhabited by burrowing bivalves (Bivalvia), segmented worms (Annelida), and acorn worms (Enteropneusta), and in some lower intertidal and high subtidal ranges by eelgrass (*Zostera marina*) (Kozloff, 1973; Dethier, 2010). Eelgrass meadows harbor epifaunal and infaunal biota, and attract transient species which use the meadows for shelter and to feed on resident organisms. Hard intertidal surfaces support a well-documented biota including barnacles (Sessilia), mussels (Bivalvia:Mytilidae), anemones (Actinaria), sea stars (Asteroidea), urchins (Echinoidea), Bryzoans (Ectoprocta), crustaceans (Decapoda), and a variety of algae (Dethier, 2010). Hard bottoms of the lower intertidal and high subtidal are home to macroalgae such as Laminariales and Desmarestiales

95 which provides habitat for a distinct community of fish and invertebrates. The upper pelagic is  
96 home to a diverse assemblage of microscopic plankton including diatoms and larvae (Strickland,  
97 1983), as well as transitory fish and marine mammals.

98 We took advantage of this setting to explore the spatial variation and distribution of marine  
99 eDNA communities. Using PCR-based methods and massively parallel sequencing, we surveyed 16S  
100 mtDNA from a suite of marine animals in water samples collected over a grid of sites extending  
101 from the shoreline out to 4 kilometers offshore in Puget Sound, Washington, USA. We leverage  
102 this sampling design to perform the first explicitly spatial analysis of eDNA-derived community  
103 similarity. We investigate two primary objectives. First we examine the spatial patterning of  
104 eDNA and determine the degree to which eDNA community similarity can be predicted by physical  
105 proximity. We expect that physical proximity will be a strong predictor of community similarity,  
106 and that community differences can be detected over small distances. Second, we examine the  
107 distribution of diversity from eDNA data, and compare it to our expectations based on distributions  
108 of macrobial communities. We expect that distinct eDNA communities exist in this setting, and  
109 that their spatial distribution coincides with that of adult macrobial organisms. Because of the  
110 vastly different communities of benthic macrobial metazoans as a function of distance from shore,  
111 we expect that more than one eDNA community is present across our 4 kilometer sampling grid,  
112 and that communities change as a function of distance from shore. For this reason, we examine two  
113 diversity measures of eDNA communities that have been widely used to reveal broad scale patterns  
114 based on macrobiota in many ecological systems. Finally, we identify the taxa represented in the  
115 eDNA communities, which span a range of life-history characteristics, and we expect that the spatial  
116 distribution of eDNA will most closely resemble the distribution of adults in taxa with low dispersal  
117 potential.

## 118 **Methods**

119 There are seven discrete steps to our methodology: (1) Environmental sample collection, (2) isolation  
120 of particulates from water via filtration, (3) isolation of DNA from filter membrane, (4) amplification  
121 of target locus via PCR, (5) sequencing of amplicons, (6) bioinformatic translation of raw sequence  
122 data into tables of sequence abundance among samples, and (7) community ecological analyses of

123 eDNA. We provide brief overviews of these steps here, and encourage the reader to review the fully  
124 detailed methods presented in the supplementary material (Supplemental Material).

## 125 **Environmental Sampling**

126 Starting from lower-intertidal patches of *Zostera marina*, we collected water samples at 1 meter  
127 depth from 8 points (0, 75, 125, 250, 500, 1000, 2000, and 4000 meters) along three parallel transects  
128 separated by 1000 meters (24 sample locations total; Figure 1). To destroy residual DNA on  
129 equipment used for field sampling and filtration, we washed with a 1:10 solution of household bleach  
130 (8.25% sodium hypochlorite; 7.25% available chlorine) and deionized water, followed by thorough  
131 rinsing with deionized water. Each environmental sample was collected in a clean 1 liter high-density  
132 polyethylene bottle, the opening of which was covered with 500 micrometer nylon mesh to prevent  
133 entry of larger particles. Immediately after collecting the sample, the mesh was replaced with a  
134 clean lid and the sample was held on ice until filtering.

## 135 **Filtration**

136 One liter from each water sample was filtered in the lab on a clean polysulfone vacuum filter  
137 holder fitted with a 47 millimeter diameter cellulose acetate membrane with 0.45 micrometer pores.  
138 Filter membranes were moved into 900 microliters of Longmire buffer (Longmire et al., 1997) using  
139 clean forceps and stored at room temperature (Renshaw et al., 2014). To test for the extent of  
140 contamination attributable to laboratory procedures, we filtered three replicate 1 liter samples of  
141 deionized water. These samples were treated identically to the environmental samples throughout  
142 the remaining protocols.

## 143 **DNA Purification**

144 DNA was purified from the membrane following a phenol:chloroform:isoamyl alcohol protocol follow-  
145 ing Renshaw (Renshaw et al., 2014). Preserved membranes were incubated at 65C for 30 minutes  
146 before adding 900 microliters of phenol:chloroform:isoamyl alcohol and shaking vigorously for 60  
147 seconds. We conducted two consecutive chloroform washes by centrifuging at 14,000 rpm for 5  
148 minutes, transferring the aqueous layer to 700 microliters chloroform, and shaking vigorously for 60  
149 seconds. After a third centrifugation, 500 microliters of the aqueous layer was transferred to tubes

150 containing 20 microliters 5 molar NaCl and 500 microliters 100% isopropanol, and frozen at -20C  
151 for approximately 15 hours. Finally, all liquid was removed by centrifuging at 14000 rpm for 10  
152 minutes, pouring off or pipetting out any remaining liquid, and drying in a vacuum centrifuge at  
153 45C for 15 minutes. DNA was resuspended in 200 microliters of ultrapure water. Genomic DNA  
154 extracted from tissue of a species absent from the sampled environment (*Oreochromis niloticus*)  
155 served as positive control for the remaining protocols.

## 156 **PCR Amplification**

157 From each DNA sample, we amplified an approximately 115 base pair (bp) region of the mito-  
158 chondrial gene encoding 16S RNA using a two-step polymerase chain reaction (PCR) protocol  
159 described by O'Donnell et al. (2016). In the first set of reactions, primers were identical in ev-  
160 ery reaction (forward: AGTTACYYTAGGGATAACAGCG; reverse: CCGGTCTGAACTCAGAT-  
161 CAYGT); primers in the second set of reactions included these same sequences but with 3 variable  
162 nucleotides (NNN) and an index sequence on the 5' end (see Sequencing Metadata). We used the  
163 program OligoTag (Coissac, 2012) to generate 30 unique 6-nucleotide index sequences differing by  
164 a minimum Hamming distance of 3 (see Sequencing Metadata). Indexed primers were assigned to  
165 samples randomly, with the identical index sequence on the forward and reverse primer to avoid  
166 errors associated with dual-indexed multiplexing (Schnell et al., 2015). In a UV-sterilized hood,  
167 we prepared 25 microliter reactions containing 18.375 microliters ultrapure water, 2.5 microliters  
168 10x buffer, 0.625 microliters deoxynucleotide solution (8 millimolar), 1 microliter each forward and  
169 reverse primer (10 micromolar, obtained lyophilized from Integrated DNA Technologies (Coralville,  
170 IA, USA)), 0.25 microliters Qiagen HotStar Taq polymerase, and 1.25 microliter genomic eDNA  
171 template at 1:100 dilution in ultrapure water. PCR thermal profiles began with an initialization  
172 step (95C; 15 min) followed by cycles (40 and 20 for the first and second reaction, respectively) of  
173 denaturation (95C; 15 sec), annealing (61C; 30 sec), and extension (72C; 30 sec). 20 identical PCRs  
174 were conducted from each DNA extract using non-indexed primers; these were pooled into 4 groups  
175 of 5 in order to ensure ample template for the subsequent PCR with indexed primers. In order to  
176 isolate the fragment of interest from primer dimer and other spurious fragments generated in the  
177 first PCR, we used the AxyPrep Mag FragmentSelect-I kit with solid-phase reversible immobiliza-  
178 tion (SPRI) paramagnetic beads at 2.5x the volume of PCR product (Axygen BioSciences, Corning,

179 NY, USA). A 1:5 dilution in ultrapure water of the product was used as template for the second  
180 reaction. PCR products of the second reaction were purified using the Qiagen MinElute PCR Pu-  
181 rification Kit (Qiagen, Hilden, Germany). Ultrapure water was used in place of template DNA and  
182 run along with each batch of PCRs to serve as a negative control for PCR; none of these produced  
183 visible bands on an agarose gel. In total, four separate replicates from each of 31 DNA samples  
184 were carried through the two-step PCR process for a total of 124 sequenced PCR products. These  
185 were combined with additional samples from other projects, totaling 345 samples for sequencing.

## 186 **DNA Sequencing**

187 PCR products were pooled according to their primer index in equal concentration, and 150 ng was  
188 prepared for library sequencing using the KAPA high-throughput library prep kit with real-time  
189 library amplification protocol (KAPA Biosystems, Wilmington, MA, USA). Each of these ligated  
190 sequencing adapter including an additional 6 base pair index sequence (NEXTflex DNA barcodes;  
191 BIOO Scientific, Austin, TX, USA). Thus, each PCR product was identifiable via its unique com-  
192 bination of index sequences in the sequencing adapters and primers. Fragment size distribution  
193 and concentration of each library was quantified using an Agilent 2100 BioAnalyzer. Libraries were  
194 pooled in equal concentrations and sequenced for 150 base pairs in both directions (PE150) using  
195 an Illumina NextSeq at the Stanford Functional Genomics Facility (machine NS500615, run 115,  
196 flowcell H3LFLAFXX), where 20% PhiX Control v3 was added to act as a sequencing control and  
197 to enhance sequencing depth. Raw sequence data in fastq format is publicly available (see Data  
198 Availability).

## 199 **Sequence Data Processing (Bioinformatics)**

200 Detailed bioinformatic methods are provided in the supplemental material, and analysis scripts  
201 used from raw sequencer output onward can be found in the public project directory (see Analysis  
202 Scripts). Briefly, we performed five steps to process the sequence data: (1) Merge paired-end  
203 reads, (2) eliminate low-quality reads, (3) eliminate PCR artifacts (chimeras), (4) cluster reads by  
204 similarity into operational taxonomic units (OTUs), and (5) match observed sequences to taxon  
205 names. Additionally, we checked for consistency among PCR replicates, excluded extremely rare  
206 sequences, and rescaled (rarefied) the data to account for differences in sequencing depth. The data



for input to further analyses are a contingency table of the mean count of unique sequences, OTUs, or taxa present in each environmental sample.

## Ecological Analyses

After gathering the data, we use the eDNA community observed at each location to make inferences about the spatial patterning of eDNA communities. We use statistical tools from community ecology to assess the spatial structure of eDNA communities. We report similarity (1- dissimilarity) rather than dissimilarity in all cases for ease of interpretation.

### Objective 1: Community similarity as a function of distance

#### Distance Decay

To address our first objective and determine whether or not nearby samples are more similar than distant ones, we fit a nonlinear model to represent decreasing community similarity with distance. We calculated the pairwise Bray-Curtis similarity (1 - Bray-Curtis dissimilarity) between eDNA communities using the R package *vegan* (Oksanen et al., 2016) and the great circle distance between sampling points using the Haversine method as implemented by the R package *geosphere* (Hijmans, 2016). This model is similar to the Michaelis-Menten function, but with an asymptote fixed at 0:

$$y_{ij} = \frac{AB}{B + x_{ij}} \quad (1)$$

Where the relationship between community similarity ( $y_{ij}$ ) and spatial distance ( $x_{ij}$ ) between observations  $i$  and  $j$  is determined by the similarity of samples at distance 0 ( $A$ ), and the distance at which half the total change in similarity is achieved ( $B$ ). This allows for a samples collected very close together (near 0) to have similarity significantly less than one. We assessed model fit using the R function *nls* (R Core Team, 2016), using the *nl2sol* algorithm from the *Port* library to solve separable nonlinear least squares using analytically computed derivatives (<http://netlib.org/port/nsg.f>). We set bounds of 0 and 1 for the intercept parameter and a lower bound of 0 for the distance at half similarity; starting values of these parameters were 0.5 and  $x_{max}/2$ , respectively. We calculated a 95% confidence interval for the parameters and the predicted values using a first-order Taylor expansion approach implemented by the function *predictNLS* in the R package *propagate* (Spiess,

232 2014).

233 There are other conceptually reasonable forms to expect the space-by-similarity relationship  
234 to take; we present these in the supplemental material along with alternative data subsets and  
235 similarity indices (see Supplemental Material).

## 236 **Objective 2: Spatial distribution of diversity**

### 237 **Community Classification**

238 To determine the spatial distribution and variation of eDNA communities (objective 2), we used  
239 multivariate classification algorithms. We simultaneously assessed the existence of distinct com-  
240 munity types and the membership of samples to those community types using an unsupervised  
241 classification algorithm known as partitioning around medoids (PAM; sometimes referred to as k-  
242 medoids clustering) (Kaufman and Rousseeuw, 1990), as implemented in the R package cluster  
243 (Maechler et al., 2016). The classification of samples to communities was made on the basis of  
244 their pairwise Bray-Curtis similarity, calculated using the function `vegdist` in the R package `vegan`  
245 (Oksanen et al., 2016). Other distance metrics were evaluated but had no appreciable effect on the  
246 outcome of the analysis (Figure 8). In order to choose an optimal number of clusters ( $K$ ), we evalu-  
247 ated the distribution of silhouette widths, a measure of the similarity between each sample and its  
248 cluster compared to its similarity to other clusters. We repeated the analysis using fuzzy clustering  
249 (FANNY, (Kaufman and Rousseeuw, 1990)); however, the results were qualitatively similar to the  
250 results using PAM so we omit them here.

### 251 **Aggregate Measures of Diversity**

252 We calculated two measures of diversity (richness and 1-Simpson's Index) to ask if aggregate metrics  
253 of the eDNA community showed evidence of spatial patterning. Richness is a measure of the number  
254 of distinct types of organisms present and so ranges from 1 (only one taxon observed) to  $R$ , the  
255 number of taxa observed across all samples. 1-Simpson's is a measure of the evenness of the  
256 distribution of abundance of taxa in a sample and ranges from 0 to 1, with the value interpreted as  
257 the probability that two sequences randomly selected from the sample will belong to different taxa.  
258 Thus larger values of the index indicate more evenly divided communities. We calculated Moran's

I for both diversity metrics to test for spatial autocorrelation. We also tested for a linear effect of log-transformed distance from shore on each measure of diversity to ask how diversity changes over this strong environmental gradient.

## **Taxon and Life History Patterns**

After assigning taxon names to the abundance data, we plotted the distribution in space of a selection of taxa to compare with our expectations on the basis of adult distributions (objective 2). Our aim was to understand where each taxon occurred in the greatest proportional abundance, and its distribution in space relative to that maximum. Thus, we rescaled each sample to proportional abundance, extracted the data from a single taxon, and scaled those values between 0 and 1. We collated life history characteristics 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 (see Life History Data).

## **Results**

### **Sequence Data Processing (Bioinformatics)**

Preliminary sequence analysis strongly suggested that the observed variation among environmental samples reflects true variation in the environment, rather than variability due to lab protocols, for the following reasons (note that all value ranges are reported as mean plus and minus one standard deviation). First, all libraries passed the FastQC per-base sequence quality filter, generating a total of 371,576,190 reads passing filter generated in each direction. Second, samples in this study were represented by an adequate number of reads ( $333,537.9 \pm 112,200.5$ ), with no individual sample

receiving fewer than 130,402 reads. Third, there was a very low frequency of cross-contamination from other libraries into those reported here ( $5\text{e-}05 \pm 8\text{e-}05$ ; max proportion 0.00034). Fourth, after scaling all samples to the same sequencing depth, OTUs with abundance greater than 178 reads (0.14% of a sample's reads) experienced no turnover among PCR replicates within a sample. Fifth, sequence abundances among PCR replicates within water samples were remarkably consistent. A single sample had low similarity among PCR replicates (0.659) after removing this outlier, the lowest mean similarity among replicates within a sample was 0.966. Overall similarities among PCR replicates within a sample were extremely high ( $0.976 \pm 0.013$ ), and far higher than that of than among samples ( $0.3 \pm 0.16$ ).

## Ecological Analyses

### Distance Decay

Physical proximity is a good predictor of eDNA community similarity: Similarity decreased from 0.40 (95%CI = 0.36, 0.45) to half that amount at 4500 meters (95%CI = 2900, 7500) (Figure 2).

### Community Classification

Despite a clear trend in community similarity as a function of spatial separation, the results from our classification analysis are difficult to interpret. The silhouette analysis indicated the presence of 8 distinct communities; however, the gain in mean silhouette width from 2 was small (0.1), and lacked a distinctive peak (Figure 4), indicating substantial uncertainty in the clustering algorithm. Thus, we present the results of cluster assignment for both  $K = 2$  and  $K = 8$  to illustrate the range of results (Figure 3). Excluding taxa which occur in only one site had no discernible effect on the outcome of the PAM analysis (number of clusters, assignment to clusters). While there was no distinct spatial divide indicating the presence of an inshore versus an offshore community, one of the two communities (at  $K = 2$ ) occurred in only 2 out of 18 samples inside 1000 meters from shore, and never occurred within 125 meters of shore, suggesting the presence of an inshore and offshore community.

## 311 Diversity in Space

312 Sites offshore tend to be less diverse (richness) and more even (1-Simpson's D) than those in-  
313 shore (Figure 6). Mean OTU richness declined by 1.42 per 1000 meters from a mean of 17.6 taxa  
314 (95%CI = 2.15) inshore to 11.9 taxa (95%CI = 4.31) at offshore locations ( $p = 0.0415$ ; Fig. fig-  
315 ures/diversity.pdf). 1-Simpson's diversity, the probability that two reads chosen at random from a  
316 sample belong to different species, increased by .0666 per 1000 meters from 0.225 (95%CI = 0.0558)  
317 to 0.491 (95%CI =  $\pm 0.112$ ), indicating that sequence reads were less evenly distributed among  
318 taxa in offshore samples ( $p \ll 0.05$ ; Figure 6). There was no evidence for spatial autocorrelation  
319 for any of the diversity metrics (Moran's I,  $p > 0.05$ ; Figure 5).

## 320 Taxon and Life History Patterns

321 We were able to assign a taxon name with confidence to 136 of 146 OTU sequences. The vast ma-  
322 jority of sequences (97.6%) and OTUs (96.9%) were matched to organisms that have high potential  
323 for dispersal at either the gamete, larval, or adult stage, making it impossible to determine whether  
324 the source of that DNA was adults with well-documented spatial patterns (e.g. sessile nearshore  
325 specialists) or highly mobile early life history stages. Of the 6 OTUs for which dispersal is limited  
326 during all life history stages, only 2 occurred in more than two samples, precluding a quantita-  
327 tive comparison of spatial dispersion based on life history characteristics. These were assigned to  
328 *Cymatogaster aggregata*, a viviparous nearshore fish with internal fertilization, and *Cupolaconcha*  
329 *meroclista*, a sessile Vermetid gastropod with presumed internal fertilization and short larval dis-  
330 persal (Strathmann and Strathmann, 2006; Phillips and Shima, 2010; Calvo and Templado, 2004).  
331 *Cymatogaster aggregata* was distinctly more abundant close to shore, with no sequences occurring in  
332 any sample beyond 250 meters (Figure 7). *Cupolaconcha meroclista* showed no such distinct spatial  
333 trend, occurring in nearly equal abundance at three sites, 75, 500, and 2000 meters from shore. An  
334 additional species that was highly abundant in the sequence data, the krill *Thysanoessa raschii*,  
335 has pelagic adults, highly seasonal reproduction, and sinking eggs; their distribution was consistent  
336 with our expectations based on a tendency of adults to aggregate offshore. Finally, the two most  
337 abundant taxa in the dataset were the mussel genus *Mytilus* and the Barnacle order Sessilia; the  
338 adults of both taxa are sessile and occur exclusively on hard intertidal substrata but have highly

339 motile larvae.

## 340 Discussion

341 Indirect surveys of organismal presence are a key development in ecosystem monitoring in the face  
342 of increased anthropogenic pressure and dwindling resources for ecological research. Monitoring of  
343 organisms using environmental DNA is an especially promising method, given the rapid pace of  
344 advancement in technological innovation and cost efficiency in the field of DNA sequencing and  
345 quantification. For the first time in a marine environment, we document four key patterns: (1)  
346 communities far from one another tend to be less similar than those that are nearby, (2) distinct  
347 eDNA communities exist and are distributed in a non-random fashion, (3) diversity declines with  
348 distance from shore, and (4) spatial patterning of eDNA is associated with taxon-specific life history  
349 characteristics.

### 350 (1) Communities far from one another tend to be less similar than those that are 351 nearby

352 We demonstrate that more distant locations have less similar eDNA communities than more proxi-  
353 mate locations in Puget Sound, a dynamic marine environment. Our finding is in line with observa-  
354 tions based on traditional surveys of terrestrial plants and fungi (Nekola and White, 1999; Bahram  
355 et al., 2013; Condit, 2002; Chust et al., 2006) and of microorganisms in freshwater (Wetzel et al.,  
356 2012), marine (Chust et al., 2013), and estuarine (Martiny et al., 2011) environments. To our knowl-  
357 edge, it is the first to report such a pattern using massively parallel sequencing of environmental  
358 DNA in the marine environment, and the first using any technique to describe this pattern from  
359 microbial metazoans. We note that the theoretical expectation is that samples at very close distance  
360 be nearly completely similar, while our samples separated by the 50 meters were only 40% similar.  
361 We interpret this to reflect the highly dynamic nature of this environment, which could cause DNA  
362 to be distributed quickly from its source, eroding the rise in similarity at small distances. At the  
363 same time, community similarity decreased to very low levels at larger scales, indicating that DNA  
364 distribution is not completely unpredictable. This finding implies that the effectively sampled area  
365 of individual water samples for eDNA analysis is likely to be quite small (<100m) in this nearshore

environment. Our estimated distance-decay relationship does indicate that proximate samples are more similar than distant samples, but we suggest this pattern is partially obscured by other factors, including signal from mobile, microscopic life-stages.

## **(2) Distinct eDNA communities exist and are distributed in a non-random fashion**

We demonstrate strong evidence for distinct community types and the non-random spatial patterning of those communities. While the spatial distributions of communities is surprising if one were concerned only with the macroscopic life stages of metazoans, it indeed does align with the broader view that even offshore pelagic communities are comprised of and influenced by nearshore organisms. While there was no distinct break in community types between onshore and offshore sites, there was some clustering of community types that may be explained by oceanographic features such as nearshore eddies generated by strong tidal exchange in a steep bathymetric setting (Yang and Khangaonkar, 2010). It would be useful to better understand such features during the period of sampling, by way of oceanographic monitoring devices.

## **(3) Richness declines and evenness increases with distance from shore**

We detected a general pattern of declining richness and increasing evenness with increasing distance offshore. Such a pattern is consistent with many other ecosystems which show strong clines in diversity metrics over environmental gradients. The coastal ocean is a highly productive and diverse ecosystem (Ray, 1988). However, our study is novel in that it corroborates a cline well-known on macroscales for macrobiota on a much smaller spatial scale for microscopic animals, suggesting that there may be a self-similarity across scales in diversity patterning (Levin, 1992). Intriguingly, the cline in diversity from inshore to offshore was not determined by shared changes in communities as one moved offshore; the classification analysis suggested a fair amount of differences among communities at a given offshore distance (Figure 3). Furthermore, the uncertainty in identification of the number of distinct clusters to best characterize the community underlines the difficulty of identifying community patterns with the number of taxonomic groups considered here. We suspect that the signature of eDNA from microscopic life-stages may explain our inability to easily detect spatial community level patterns that align with our initial expectations.

393 **(4) Spatial patterning of eDNA is associated with taxon-specific life history character-**  
394 **istics.**

395 In contrast to our expectations, other taxa including species with sessile adult stages restricted  
396 to benthic hard substrates (e.g. barnacles, mussels) are among the most abundant taxa at sites  
397 furthest from shore. However, the larvae and gametes of these taxa are abundant, pelagic, and  
398 can be transported long distances by water movement (Strathmann, 1987). This indicates that we  
399 likely detected DNA of their pelagic phase gametes and larvae. It is always possible that DNA  
400 of adults was advected over long distances and detected offshore but in light of our results with  
401 krill and surfperch, we view this as unlikely. We interpret our results as evidence that the chaotic  
402 spatial distribution of eDNA communities (Figure 3) results from our primers' affinity for many  
403 species which at some point exist as microscopic pelagic gametes or larvae. Our results emphasize  
404 that expected results based on easily visually observed individuals or detectable with traditional  
405 sampling gear such as nets may be very different from results using eDNA. This does caution that  
406 eDNA surveys may have different purposes and may not be directly comparable to existing surveys  
407 (Shelton et al., 2016).

408 We acknowledge that sampling artifacts may have affected our results. For example if entire  
409 multicellular individuals were captured in our samples, their DNA could be in much greater density  
410 than eDNA, affecting the observed community. Our sampling bottles excluded particles larger than  
411 500 micrometers, but gametes and very small larvae could have gained entry. It is possible that  
412 even a single small individual, containing many thousand mitochondria, would overwhelm the signal  
413 of another species from which hundreds of cells had been sloughed from many, larger individuals.  
414 Data on larval size distribution at the time of sampling from each species in our data set would  
415 allow us to estimate the frequency of such events. Nevertheless, it is precisely the sensitivity to  
416 small particles that makes the eDNA approach powerful, so we are reluctant to recommend that  
417 aquatic eDNA sampling use finer pre-filtering. Instead, we emphasize the importance of designing  
418 and selecting primer sets that selectively amplify target organisms. In the case of the present study,  
419 in order to recover patterns matching our expectations, this would be non-transient, benthic marine  
420 organisms lacking any pelagic life stage.

421 Our results also highlight the need for curated life-history databases. As technological advances



422 increase the speed and throughput of DNA sequencing and sequence processing, making sense of  
423 these data in a timely manner requires that natural history data be stored in standard formats in  
424 centralized repositories. The rate at which we can make sense of high-throughput survey methods  
425 will be limited by our ability to collate auxiliary data. Databases such as Global Biodiversity  
426 Information Facility (GBIF), Encyclopedia of Life (EOL), and FishBase (Parr et al., 2014; Froese  
427 and Pauly, 2016) contain records of taxonomy, occurrence, and other rudimentary data types, but  
428 there is no centralized, standardized repository for even basic natural history data such as body  
429 size. As NCBI's nucleotide and protein sequence database (GenBank) has facilitated a multitude of  
430 transformative studies in diverse fields, an ecological analog would be a huge boon for biodiversity  
431 science.

432 Surveys based on eDNA are intensely scrutinized because of the danger that the final data are  
433 subject to complicated laboratory and bioinformatic procedures. Finding virtually no variability  
434 among lab and bioinformatic treatments from the point of PCR onward, we were confident our  
435 results represented actual field-based differences among samples. However, we note that one PCR  
436 replicate had a clear signal of contamination in that the sequence community was extremely similar  
437 to those from a different environmental sample. The source of this error is difficult to identify, but  
438 seems most likely to be an error during PCR preparation, either in assignment or pipetting during  
439 preparation of indexed primers. While the remainder of our results would be largely unchanged  
440 had we sequenced a single replicate per environmental sample, we believe the sequencing of PCR  
441 replicates is critical for ensuring data quality in eDNA sequencing studies.

442 While there is much work to be done before eDNA techniques can be confidently deployed as a  
443 standard method for ecological monitoring, this study serves as a first analysis of diversity at the  
444 fine spatial scales that are likely to be relevant to eDNA work in the field across a range of study  
445 systems.

## 446 **Acknowledgements**

447 We wish to thank Robert Morris, E. Virginia Armbrust, and James Kralj.

## 448 **Funding**

449 This work was supported by a grant from the David and Lucile Packard Foundation to RPK (grant  
450 2014-39827). The funders had no role in study design, data collection and analysis, decision to  
451 publish, or preparation of the manuscript.

## 452 **Author Contributions**

453 Conceived and designed the experiments: JL O'Donnell, RP Kelly, AO Shelton; Collected the data:  
454 JL O'Donnell, NC Lowell, GD Williams, RP Kelly, AO Shelton, JF Samhouri; Conducted the  
455 analyses: JL O'Donnell; Wrote the first draft: JL O'Donnell; Edited the manuscript: JL O'Donnell,  
456 AO Shelton, RP Kelly, JF Samhouri, GD Williams, NC Lowell

## 457 **Ethics Statement**

458 The authors declare no conflict of interest. No permits were required to do any of the research  
459 described here.

## 460 **Data Availability**

### 461 **0.1 Sequence Data**

462 All sequence files and metadata are available from EMBL:

463 <http://www.ebi.ac.uk/ena/data/view/TODO>

464 (TODO SUBMIT DATA AND METADATA; ADD URL)

### 465 **0.2 Project Repository**

466 The following components are available from the project repository on GitHub:

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

468 (TODO INSERT DOI FROM FINAL COMMIT BEFORE PUBLICATION)

### 469 0.2.1 Sequencing Metadata

470 Sequencing metadata is available in: `Data/metadata_spatial.csv`

### 471 0.2.2 Life History Data

472 Life history data is available in: `Data/life_history.csv`

### 473 0.2.3 Analysis Scripts

474 All analyses were performed using scripts available in the Analysis subdirectory.

## 475 References

- 476 Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., Sanders,  
477 N. J., Cornell, H. V., Comita, L. S., Davies, K. F., Harrison, S. P., Kraft, N. J. B., Stegen, J. C.,  
478 and Swenson, N. G. (2011). Navigating the multiple meanings of beta diversity: A roadmap for  
479 the practicing ecologist. *Ecology Letters*, 14(1):19–28.
- 480 Bahram, M., Kõljalg, U., Courty, P. E., Diédhiou, A. G., Kjølner, R., Pölme, S., Ryberg, M., Veldre,  
481 V., and Tedersoo, L. (2013). The distance decay of similarity in communities of ectomycorrhizal  
482 fungi in different ecosystems and scales. *Journal of Ecology*, 101(5):1335–1344.
- 483 Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., and Lodge, D. M.  
484 (2014). Environmental conditions influence eDNA persistence in aquatic systems. *Environmental*  
485 *Science and Technology*, 48(3):1819–1827.
- 486 Bell, T. (2010). Experimental tests of the bacterial distance–decay relationship. *The ISME Journal*,  
487 4(11):1357–1365.
- 488 Burns, R. E. (1985). *The shape and form of Puget Sound*. Washington Sea Grant, Seattle, 1 edition.
- 489 Calvo, M. and Templado, J. (2004). Reproduction and development in a vermetid gastropod,  
490 *Vermetus triquetrus*. *Invertebrate Biology*, 123(4):289–303.
- 491 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L.  
492 (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421.

493 Chamberlain, S. a. and Szöcs, E. (2013). taxize: taxonomic search and retrieval in R. *F1000Research*,  
494 2(0):191.

495 Chamberlain, S. A., Szöcs, E., Boettiger, C., Ram, K., Bartomeus, I., Foster, Z., and O'Donnell,  
496 J. L. (2016). taxize: Taxonomic information from around the web. R package.

497 Chust, G., Chave, J., Condit, R., Aguilar, S., Lao, S., and Perez, R. (2006). Determinants and  
498 spatial modeling of tree beta-diversity in a tropical forest landscape in Panama. *Journal of*  
499 *Vegetation Science*, 17(1):83–92.

500 Chust, G., Irigoien, X., Chave, J., and Harris, R. P. (2013). Latitudinal phytoplankton distribution  
501 and the neutral theory of biodiversity. *Global Ecology and Biogeography*, 22(5):531–543.

502 Coissac, E. (2012). OligoTag: A Program for Designing Sets of Tags for Next-Generation Sequencing  
503 of Multiplexed Samples. In Pompanon, F. and Bonin, A., editors, *Data Production and Analysis in*  
504 *Population Genomics SE - 2*, volume 888 of *Methods in Molecular Biology*, pages 13–31. Humana  
505 Press.

506 Condit, R. (2002). Beta-Diversity in Tropical Forest Trees. *Science*, 295(5555):666–669.

507 Cowart, D. a., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., and Arnaud-Haond, S.  
508 (2015). Metabarcoding Is Powerful yet Still Blind: A Comparative Analysis of Morphological and  
509 Molecular Surveys of Seagrass Communities. *Plos One*, 10(2):e0117562.

510 Deiner, K. and Altermatt, F. (2014). Transport distance of invertebrate environmental DNA in a  
511 natural river. *PLoS ONE*, 9(2).

512 Dethier, M. N. (2010). Overview of the ecology of Puget Sound beaches. In Shipman, H., Dethier,  
513 M. N., Gelfenbaum, G., Fresh, K. L., and Dinicola, R. S., editors, *Puget Sound Shorelines and*  
514 *the Impacts of Armoring—Proceedings of a State of the Science Workshop*, page 262.

515 Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*,  
516 26(19):2460–2461.

517 Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y., Jerde, C. L., Mahon, A. R.,  
518 Pfrender, M. E., Lamberti, G. A., and Lodge, D. M. (2016). Quantification of mesocosm fish and

519 amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*,  
520 16(1):29–41.

521 Ficetola, G. F., Miaud, C., Pompanon, F., and Taberlet, P. (2008). Species detection using envi-  
522 ronmental DNA from water samples. *Biology letters*, 4(4):423–425.

523 Froese, R. and Pauly, D. (2016). FishBase.

524 Guo, H., Chamberlain, S. A., Elhaik, E., Jalli, I., Lynes, A. R., Marczak, L., Sabath, N., Vargas,  
525 A., Więski, K., Zelig, E. M., and Pennings, S. C. (2015). Geographic variation in plant commu-  
526 nity structure of salt marshes: Species, functional and phylogenetic perspectives. *PLoS ONE*,  
527 10(5):e0127781.

528 Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular  
529 biological access to the chemistry of unknown soil microbes: a new frontier for natural products.  
530 *Chemistry & Biology*, 5(10):R245–R249.

531 Hijmans, R. J. (2016). *geosphere: Spherical Trigonometry*.

532 Hubbell, S. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography.*, volume 32.

533 Iverson, V., Morris, R. M., Frazar, C. D., Berthiaume, C. T., Morales, R. L., and Armbrust,  
534 E. V. (2012). Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine  
535 Euryarchaeota. *Science*, 335(6068):587–590.

536 Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster*  
537 *Analysis*.

538 Kelly, R. P., O’Donnell, J. L., Lowell, N. C., Shelton, A. O., Samhour, J. F., Hennessey, S. M.,  
539 Feist, B. E., and Williams, G. D. (2016). Genetic signatures of ecological diversity along an  
540 urbanization gradient. *PeerJ*, 4:e2444.

541 Klymus, K. E., Richter, C. A., Chapman, D. C., and Paukert, C. (2015). Quantification of eDNA  
542 shedding rates from invasive bighead carp *Hypophthalmichthys nobilis* and silver carp *Hypoph-*  
543 *thalmichthys molitrix*. *Biological Conservation*, 183:77–84.

544 Kozloff, E. N. (1973). *Seashore life of Puget Sound, the Strait of Georgia, and the San Juan*  
545 *Archipelago*. University of Washington Press, Seattle.

546 Levin, S. A. (1992). The problem of pattern and scale in ecology. *Ecology*, 73(6):1943–1967.

547 Longmire, J. L., Maltbie, M., and Baker, R. J. (1997). Use of lysis buffer in DNA isolation and its  
548 implication for museum collections. *Museum of Texas Tech University*, 163.

549 Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2016). *cluster: Cluster*  
550 *Analysis Basics and Extensions*.

551 Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014). Swarm: robust and  
552 fast clustering method for amplicon-based studies. *PeerJ*, 2:e593.

553 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads.  
554 *EMBnet.journal*, 17(1):10.

555 Martiny, J. B. H., Eisen, J. A., Penn, K., Allison, S. D., and Horner-Devine, M. C. (2011). Drivers of  
556 bacterial beta diversity depend on spatial scale. *Proceedings of the National Academy of Sciences*,  
557 108(19):7850–7854.

558 Nekola, J. C. and White, P. S. (1999). The distance decay of similarity in biogeography and ecology.  
559 *Journal of Biogeography*, 26(4):867–878.

560 O’Donnell, J. L., Kelly, R. P., Lowell, N. C., and Port, J. A. (2016). Indexed PCR Primers Induce  
561 Template-Specific Bias in Large-Scale DNA Sequencing Studies. *PLoS ONE*, 11(3):1–11.

562 Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R.,  
563 O’Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H.  
564 (2016). *vegan: Community Ecology Package*.

565 Parr, C. S., Wilson, N., Leary, P., Schulz, K. S., Lans, K., Walley, L., Hammock, J. A., Goddard,  
566 A., Rice, J., Studer, M., Holmes, J. T. G., and Corrigan, R. J. (2014). The Encyclopedia of  
567 Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodiversity Data Journal*,  
568 2(2):e1079.

569 Phillips, N. E. and Shima, J. S. (2010). Reproduction of the vermetid gastropod *dendropoma*  
570 maximum (Sowerby, 1825) in Moorea, French Polynesia. *Journal of Molluscan Studies*, 76(2):133–  
571 137.

572 Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J.,  
573 Yamahara, K. M., and Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest  
574 ecosystem using environmental DNA. *Molecular Ecology*, 25(2):527–541.

575 R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation  
576 for Statistical Computing, Vienna, Austria.

577 Ray, G. C. (1988). Ecological diversity in coastal zones and oceans. In Wilson, E. O. and Peter,  
578 F. M., editors, *Biodiversity*, chapter 4. National Academies Press (US), Washington, DC.

579 Renshaw, M. A., Olds, B. P., Jerde, C. L., Mcveigh, M. M., and Lodge, D. M. (2014). The room  
580 temperature preservation of filtered environmental DNA samples and assimilation into a phenol-  
581 chloroform-isoamyl alcohol DNA extraction. *Molecular Ecology Resources*.

582 Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open  
583 source tool for metagenomics. *PeerJ*, 4:e2584.

584 Sassoubre, L. M., Yamahara, K. M., Gardner, L. D., Block, B. A., and Boehm, A. B. (2016).  
585 Quantification of Environmental DNA (eDNA) Shedding and Decay Rates for Three Marine  
586 Fish. *Environmental Science & Technology*, 50(19):10456–10464.

587 Schnell, I. B., Bohmann, K., and Gilbert, M. T. P. (2015). Tag jumps illuminated - reducing  
588 sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*,  
589 15(6):1289–1303.

590 Shelton, A. O., O'Donnell, J. L., Samhouri, J. F., Lowell, N., Williams, G. D., and Kelly, R. P.  
591 (2016). A framework for inferring biological communities from environmental DNA. *Ecological*  
592 *Applications*, 26(6):1689–.

593 Shogren, A. J., Tank, J. L., Andruszkiewicz, E. A., Olds, B., Jerde, C., and Bolster, D. (2016).  
594 Modelling the transport of environmental DNA through a porous substrate using continuous  
595 flow-through column experiments. *Journal of The Royal Society Interface*, 13(119):423–425.

596 Spiess, A.-N. (2014). *propagate: Propagation of Uncertainty*.

597 Strathmann, M. F. (1987). *Reproduction and Development of Marine Invertebrates of the Northern*  
598 *Pacific Coast: Data and Methods for the Study of Eggs, Embryos, and Larvae*. University of  
599 Washington Press, Seattle.

600 Strathmann, M. F. and Strathmann, R. R. (2006). A Vermetid Gastropod with Complex Intracap-  
601 sular Cannibalism of Nurse Eggs and Sibling Larvae and a High Potential for Invasion. *Pacific*  
602 *Science*, 60(1):97–108.

603 Strickland, R. M. (1983). *The Fertile Fjord: Plankton in Puget Sound*. University of Washington  
604 Press, Seattle.

605 Strickler, K. M., Fremier, A. K., and Goldberg, C. S. (2015). Quantifying effects of UV-B, temper-  
606 ature, and pH on eDNA degradation in aquatic microcosms. *Biological Conservation*, 183:85–92.

607 Thomsen, P. F., Kielgast, J., Iversen, L. L., Møller, P. R., Rasmussen, M., and Willerslev, E. (2012).  
608 Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples.  
609 *PLoS ONE*, 7(8):1–9.

610 Turner, C. R., Uy, K. L., and Everhart, R. C. (2015). Fish environmental DNA is more concentrated  
611 in aquatic sediments than surface water. *Biological Conservation*, 183:93–102.

612 Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M.,  
613 Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure  
614 and metabolism through reconstruction of microbial genomes from the environment. *Nature*,  
615 428(6978):37–43.

616 Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. a., Wu, D.,  
617 Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W.,  
618 Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch,  
619 C., Rogers, Y.-H., and Smith, H. O. (2004). Environmental genome shotgun sequencing of the  
620 Sargasso Sea. *Science*, 304(5667):66–74.



- 621 Wetzel, C. E., de Bicudo, D. C., Ector, L., Lobo, E. A., Soininen, J., Landeiro, V. L., and Bini,  
622 L. M. (2012). Distance Decay of Similarity in Neotropical Diatom Communities. *PLoS ONE*,  
623 7(9):e45071.
- 624 Yang, Z. and Khangaonkar, T. (2010). Multi-scale modeling of Puget Sound using an unstructured-  
625 grid coastal ocean model: From tide flats to estuaries and coastal waters. *Ocean Dynamics*,  
626 60(6):1621–1637.
- 627 Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: A fast and accurate Illumina  
628 Paired-End reAd mergeR. *Bioinformatics*, 30(5):614–620.

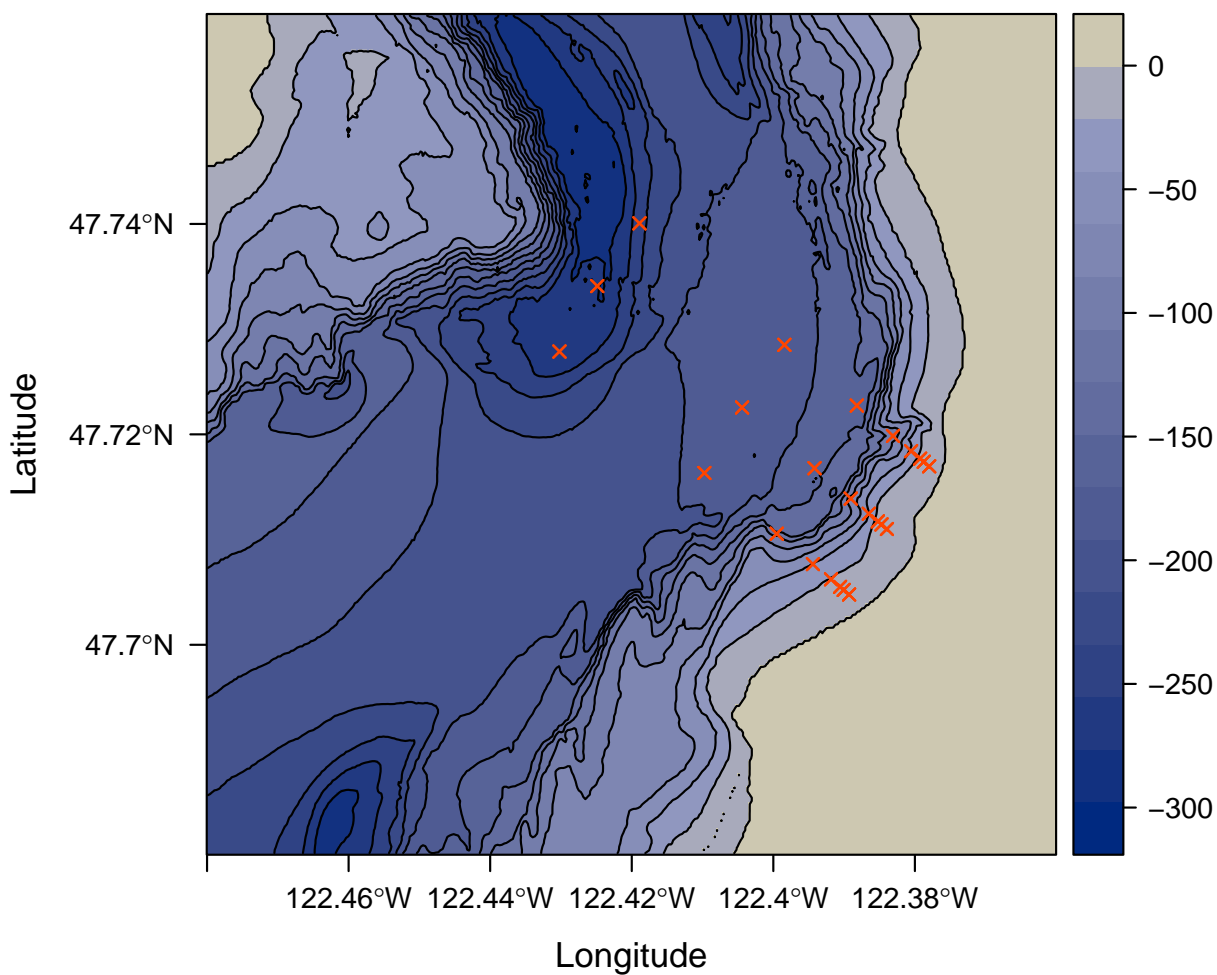


Figure 1: Map of study area. Depth in meters below sea level is indicated by shading and 25 meter contours. Sampled locations are indicated by red points.

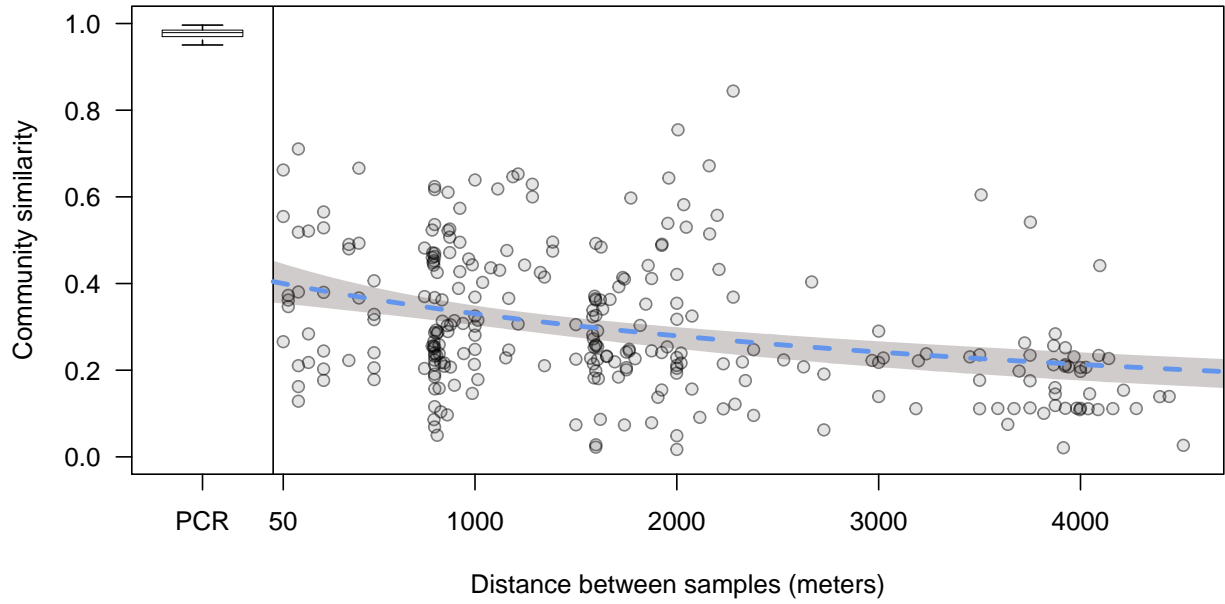


Figure 2: Distance decay relationship of environmental DNA communities. Each point represents the Bray-Curtis similarity of a site sampled along three parallel transects comprising a 3000 by 4000 meter grid. Blue dashed line represents fit of a nonlinear least squares regression (see Methods), and shading denotes the 95% confidence interval. Boxplot is comparisons within-sample across PCR replicates, separated by a vertical line at zero, where the central line is the median, the box encompasses the interquartile range, and the lines extend to 1.5 times the interquartile range. Boxplot outliers are omitted for clarity.



Figure 3: Cluster membership of sampled sites. Distance from onshore starting point is log scaled. Sites are colored and labeled by their assignment to a cluster by PAM analysis for number of clusters (K) chosen based on a priori expectations (2) and mean silhouette width (8).

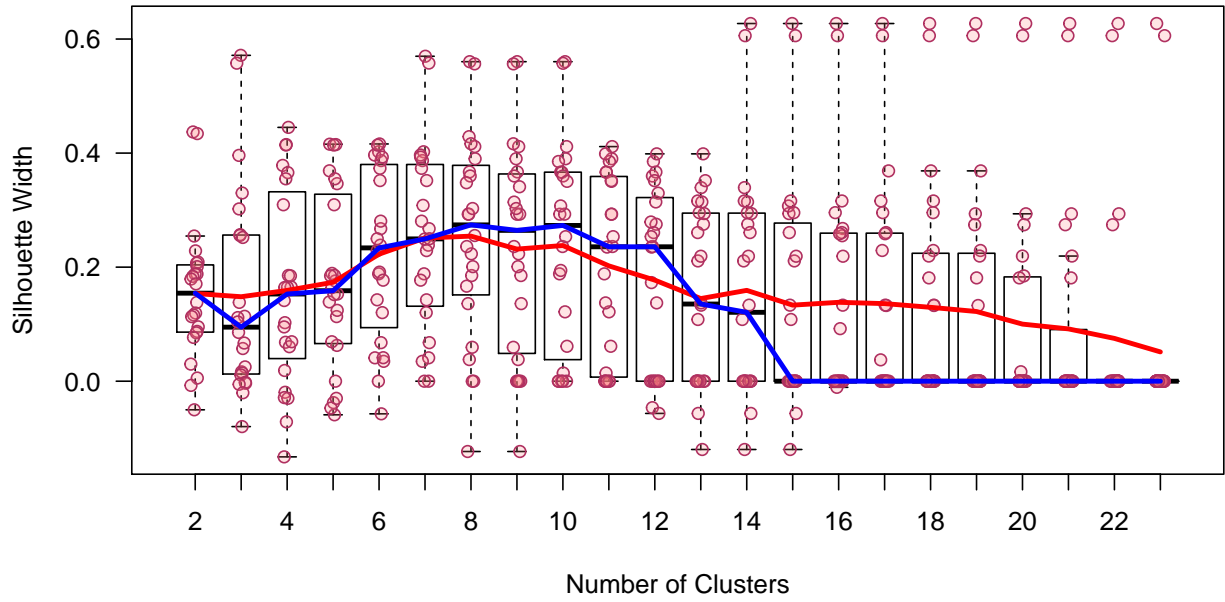


Figure 4: Silhouette widths from PAM analysis. Points are the width of the PAM silhouette of each sample at each number of clusters ( $K$ ). Red line is the mean, blue line is the median. Boxes encompass the interquartile range with a line at the median, and the whiskers extend to 1.5 times the interquartile range. Boxplot outliers are omitted for clarity.

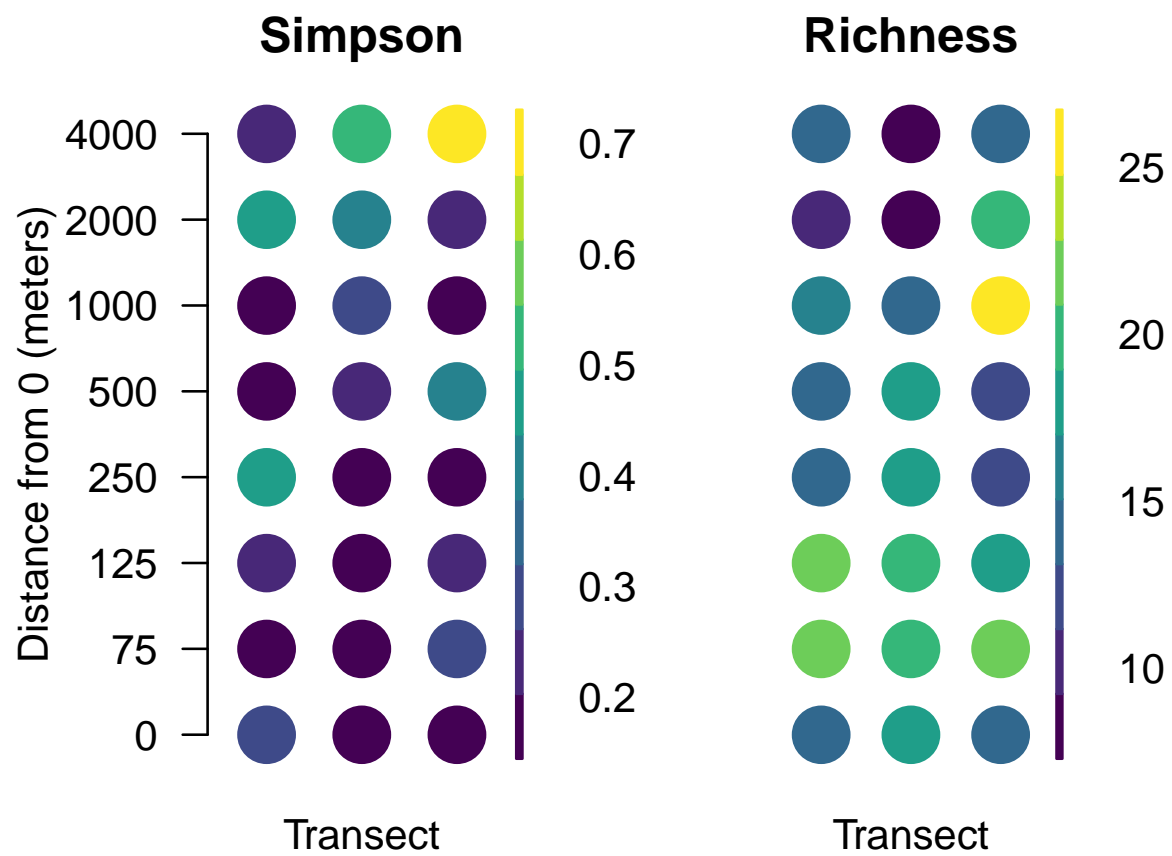


Figure 5: Aggregate measures of diversity at each sample site.

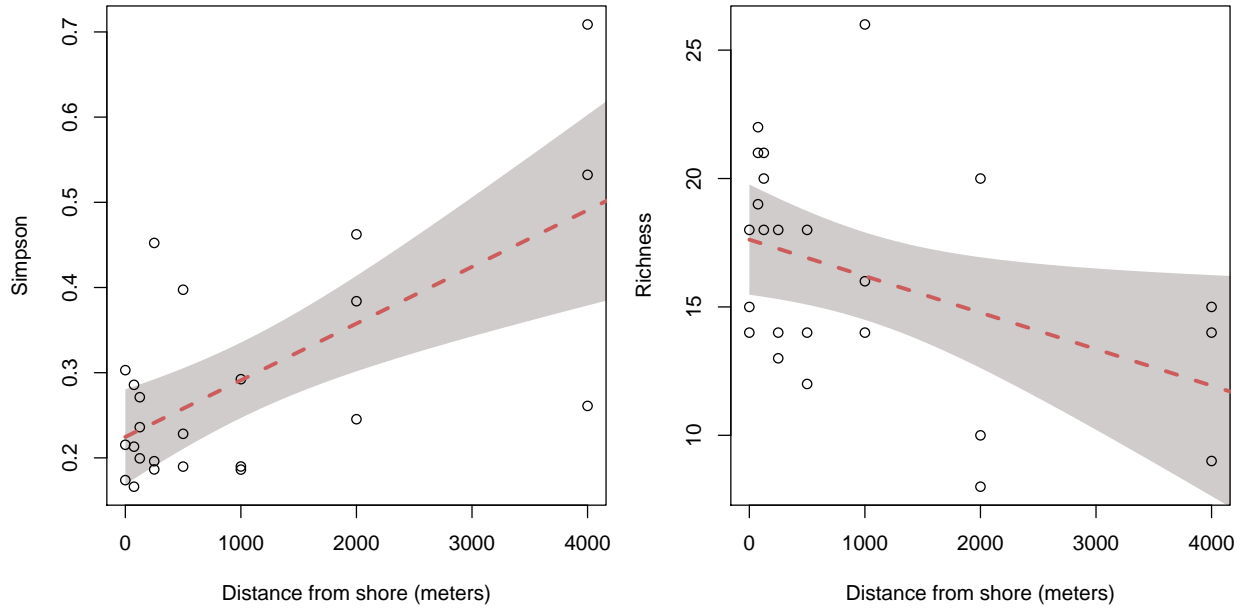


Figure 6: Aggregate diversity metrics of each site plotted against distance from shore. Both Simpson's Index (left) and richness (right) are shown, and have been computed from the mean abundance of unique DNA sequences found across 4 PCR replicates at each of 24 sites. Lines and bands illustrate the fit and 95

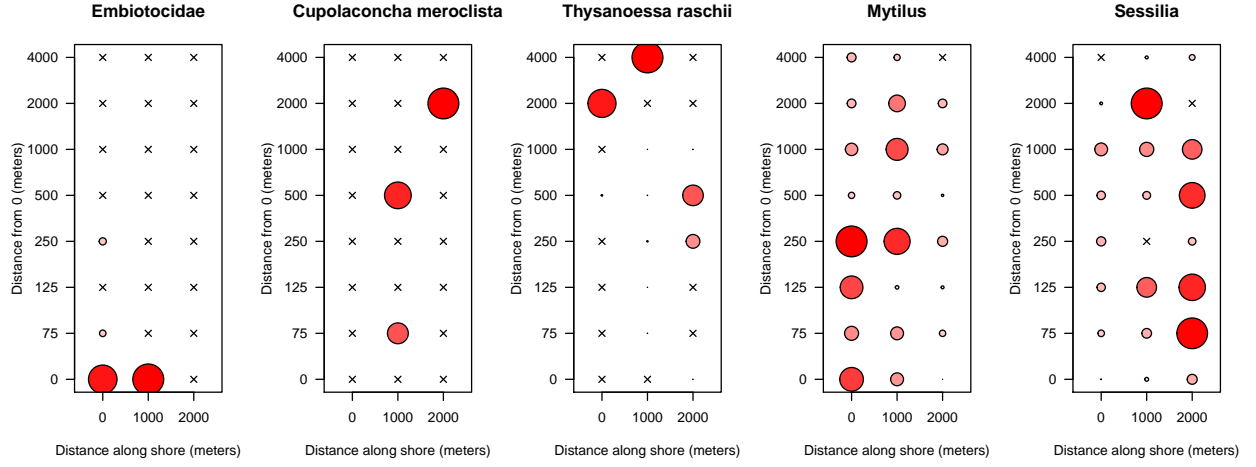


Figure 7: Distribution of eDNA from select taxa. Circles are colored and scaled by the proportion of that taxon's maximum proportional abundance. That is, the largest circle is the same size in each of the panels, and occurs where that taxon contributed the greatest proportional abundance of reads to that sample.



## 630 Supplemental Material

### 631 Methods

#### 632 Bioinformatics

633 Reads passing the preliminary Illumina quality filter were demultiplexed on the basis of the adapter  
634 index sequence by the sequencing facility. We used fastqc to assess the fastq files output from the  
635 sequencer for low-quality indications of a problematic run. Forward and reverse reads were merged  
636 using PEAR v0.9.6 Zhang et al. (2014) and discarded if more than 0.01 of the bases were uncalled.  
637 If a read contained two consecutive base calls with quality scores less than 15 (i.e. probability of  
638 incorrect base call = 0.0316), these bases and all subsequent bases were removed from the read.  
639 Paired reads for which the probability of matching by chance alone exceeded 0.01 were not assembled  
640 and omitted from the analysis. Assembled reads were discarded if assembled sequences were not  
641 between 50 and 168 bp long, or if reads did not overlap by at least 100 bp.

642 We used vsearch v2.1.1 (Rognes et al., 2016) to discard any merged reads for which the sum of the  
643 per-base error probabilities was greater than 0.5 (“expected errors”) Edgar (2010). Sequences were  
644 demultiplexed on the basis of the primer index sequence at base positions 4-9 at both ends using the  
645 programming language AWK. Primer sequences were removed using cutadapt v1.7.1 Martin (2011),  
646 allowing for 2 mismatches in the primer sequence. Identical duplicate sequences were identified,  
647 counted, and removed in python to speed up subsequent steps by eliminating redundancy, and  
648 sequences occurring only once were removed. We checked for and removed any sequence likely to be  
649 a PCR artifact due to incomplete extension and subsequent mis-priming using a method described  
650 by Edgar (2010) and implemented in vsearch v2.0.2. Sequences were clustered into operational  
651 taxonomic units (OTUs) using the single-linkage clustering method implemented by swarm version  
652 2.1.1 with a local clustering threshold (d) of 1 and fastidious processing (Mahé et al., 2014).

653 Cross-contamination of environmental, DNA, or PCR samples can result in erroneous inference  
654 about the presence of a given DNA sequence in a sample. However, other processes can contribute  
655 to the same signature of contamination. For example, errors during oligonucleotide synthesis or  
656 sequencing of the indexes could cause reads to be erroneously assigned to samples. The frequency  
657 of such errors can be estimated by counting the occurrence of sequences known to be absent from

658 a given sample, and of reads that do not contain primer index sequences in the expected position  
659 or combinations. These occurrences indicate an error in the preparation or sequencing procedures.  
660 We estimated a rate of incorrect sample assignment by calculating the maximum rate of occur-  
661 rence of index sequences combinations we did not actually use, as well as the rates of cross-library  
662 contamination by counting occurrences of primer sequences from 12S amplicons prepared in a lab  
663 more than 1000 kilometers away, but pooled and sequenced alongside our samples. This represents  
664 a general minimum rate at which we can expect that sequences from one environmental sample  
665 could be erroneously assigned to another, and so we considered for further analysis only those reads  
666 occurring with greater frequency than this across the entire dataset.

667 We checked for experimental error by evaluating the Bray-Curtis similarity (1 - Bray-Curtis  
668 dissimilarity) among replicate PCRs from the same DNA sample. We calculated the mean and  
669 standard deviation across the dataset, and excluded any PCR replicates for which the similarity  
670 between itself and the other replicates was less than 1.5 standard deviations from the mean.

671 To account for variation in the number of sequencing reads (sequencing depth) recovered per  
672 sample, we rarefied the within-sample abundance of each OTU by the minimum sequencing depth  
673 (Oksanen et al., 2016).

674 Because each step in this workflow is sensitive to contamination, it is possible that some se-  
675 quences are not truly derived from the environmental sample, and instead represent contamination  
676 during field sampling, filtration, DNA extraction, PCR, fragment size selection, quantitation, se-  
677 quencing adapter ligation, or the sequencing process itself. We take the view that contaminants  
678 are unlikely to manifest as sequences in the final dataset in consistent abundance across replicates;  
679 indeed, our data show that the process from PCR onward is remarkably consistent. Thus, after  
680 scaling to correct for sequencing depth variation, we calculated from our data the maximum number  
681 of sequence counts for which there is turnover in presence-absence among PCR replicates within an  
682 environmental sample. We use this number to determine a conservative minimum threshold above  
683 which we can be confident that counts are consistent among replicates and not of spurious origin,  
684 and exclude from further analysis observations where the mean abundance across PCR replicates  
685 within samples does not reach this threshold. For further analyses we use the mean abundance  
686 across PCR replicates for each of the 24 environmental samples.

687 In order to determine the most likely taxon from which each sequence originated, the representa-

688 tive sequence from each OTU was then queried against the NCBI nucleotide collection (GenBank;  
689 version October 7, 2015; 32,827,936 sequences) using the blastn command line utility (Camacho  
690 et al., 2009). In order to maximize the accuracy of this computationally intensive step, we imple-  
691 mented a nested approach whereby each sequence was first queried using strict parameters (e-value  
692 = 5e-52), and if no match was found, the query was repeated with decreasingly strict e-values (5e-48  
693 5e-44 5e-40 5e-36 5e-33 5e-29 5e-25 5e-21 5e-17 5e-13). Other parameters were unchanged among  
694 repetitions (word size: 7; maximum matches: 1000; culling limit: 100; minimum percent identity:  
695 0). Each query sequence can be an equally good match to multiple taxa either because of invariabil-  
696 ity among taxa or errors in the database (e.g. human sequences are commonly attributed to other  
697 organisms when they in fact represent lab contamination). In order to guard against these spurious  
698 results, we used an algorithm to find the lowest common taxon for at least 80% of the matched  
699 taxa, implemented in the R package taxize 0.7.8 (Chamberlain and Szöcs, 2013; Chamberlain et al.,  
700 2016). Similarly, we repeated analyses using the dataset consolidated at the same taxonomic rank  
701 across all queries, for the rank of both family and order.

## 702 **Alternative distance decay model formulations**

703 **Linear:** We fit a straight line through the points after log-transforming the spatial distances  
704 to estimate the intercept and slope. This model ignores the bounds of our response variable of  
705 community similarity.

706 **Michaelis-Menten:** We fit a Michaelis-Menten-like curve to our data. Our formulation can be  
707 thought of as a modification of the Michaelis-Menten equation, but with the addition of a parameter  
708 in the numerator which modifies the intercept.

$$y = \frac{AB + Cx}{B + x} \quad (2)$$

709 Where  $C$  is the asymptote of minimum similarity. This formulation allows us to estimate the  
710 maximum similarity in the system, and the rate at which it is achieved. If the value of the parameter  
711 ( $AB$ ) is 0 (e.g. if the intercept is 0), the form is identical to the Michaelis-Menten equation:

$$y = \frac{Cx}{B+x} \quad (3)$$

712 This is conceptually satisfying in that a fit through  $[0,1]$  reflects the theoretical expectation that  
713 samples at zero distance from one another are necessarily identical. Given an efficient sampling  
714 technique, replicate samples taken at the same position in space should be identical, and thus the  
715 intercept of the regression of similarity against distance should be 1, and deviation from 1 is an  
716 indicator of the efficiency of the sampling method.

717 Finally, we considered a model which estimates an asymptote as the total change in similarity  
718 ( $D$ ):

$$y = \frac{A + Dx}{B + x} \quad (4)$$

719 However, this model failed to converge and produced uninformative estimates of all parameters.

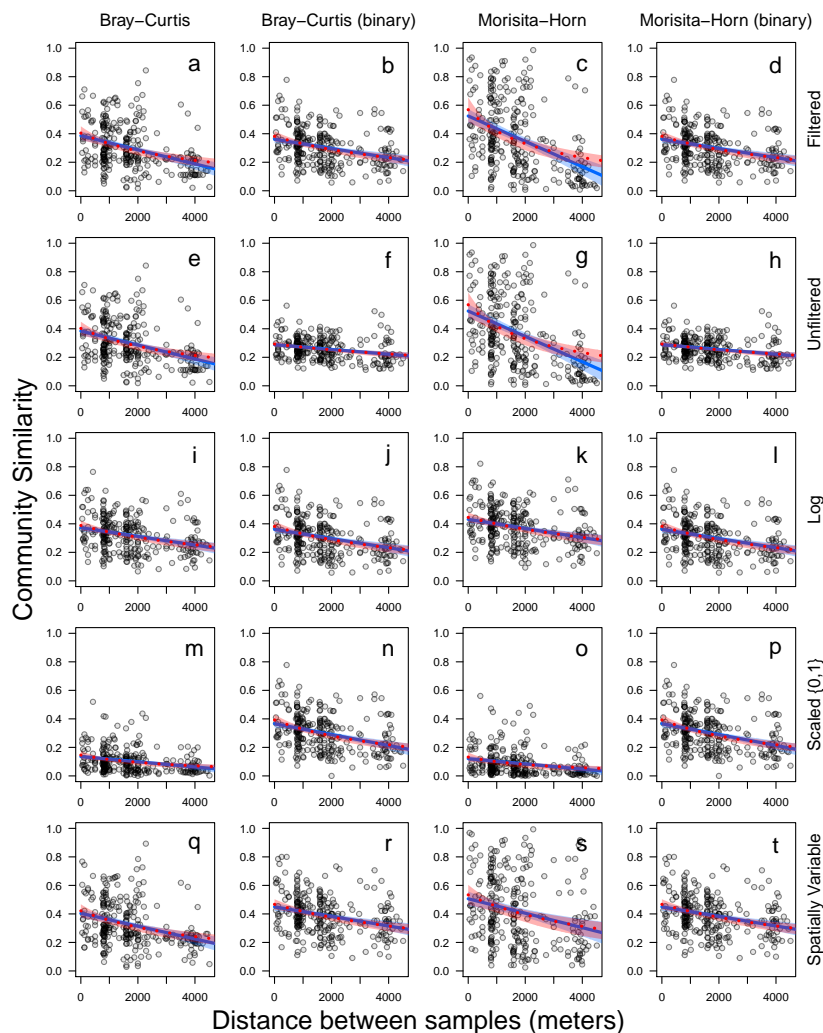


Figure 8: Distance decay relationship of environmental DNA communities using a variety of models, metrics, and data subsets. Each point represents the similarity of a site sampled along three parallel transects comprising a 3000 by 4000 meter grid. Each row of plots represents a different data subset indicated in the right margin, including the final filtered data reported in the main text (a-d), the unfiltered data including all rare OTUs (e-h), log-transformed ( $\log(x+1)$ ) data (i-l), OTU abundance scaled relative to within-taxon maximum (m-p), and exclusion of OTUs found at only one site (q-t). Columns indicate the similarity index used (Bray Curtis or Morisita-Horn) and whether the input was full abundance data or binary (0,1) transformed data. Lines and bands illustrate the fit and 95% confidence interval of both the main nonlinear model (red, dashed line) and a simple linear model (blue, solid line). Results using the Jaccard distance are omitted because of its similarity to Bray-Curtis.

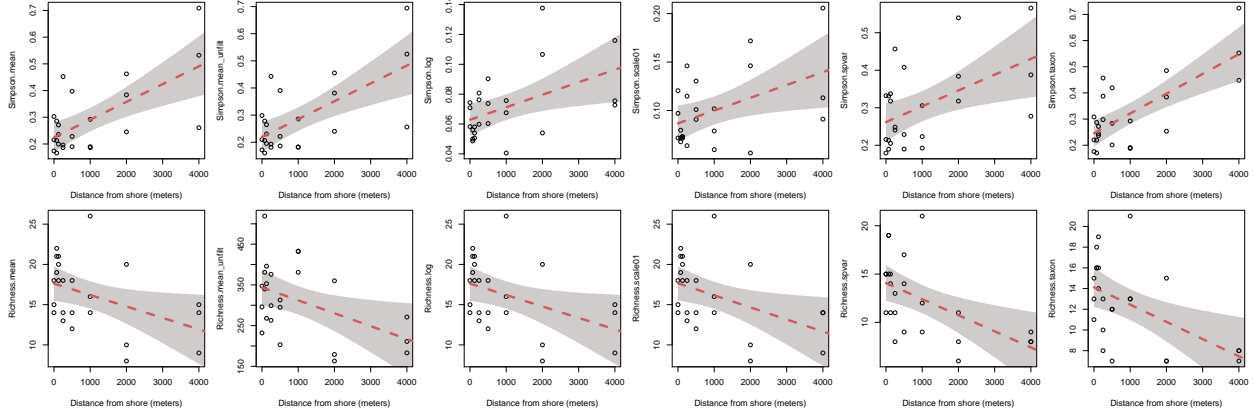


Figure 9: Aggregate diversity metrics of each site plotted against distance from shore. Both Simpson's Index (top) and richness (bottom) are shown for a variety of data subsets and transformations (left to right: mean, unfiltered mean,  $\log(x + 1)$ , transformed, scaled, spatially variable, and taxon clustered). Lines and bands illustrate the fit and 95% confidence interval of a linear model. See methods text for detailed data descriptions.