

Response to Reviews

January 13, 2017

Manuscript Title: Spatial distribution of environmental DNA in a nearshore marine habitat
Journal: PeerJ
Manuscript ID: 14649
Authors: O'Donnell JL, Kelly RP, Shelton AO, Samhouri JF, Lowell NC, Williams GD

We are grateful to the editor and reviewers for their insightful feedback. Reviewer feedback is below in regular typeface, with our responses indented and colored. **Line numbers refer to the track changes version of the manuscript.**

Editor

Both reviewers commented favourably on this manuscript and have made some useful and constructive comments which you should consider and respond to. I would also like you to respond to the comment by Bernd Haenfling on your pre-print as he is in agreement with reviewer 1 in relation to the degree of novelty this work represents (line 100 and your concluding sentence). Personally, I found your discussion of the results in the light of the advantages and limitations of this methodology interesting and highly informative. I wondered what the community analyses results would look like if you compartmentalised by e.g. life history or type (e.g. fish)?

Thank you for the positive feedback. We thank the commenters for pointing out recent literature that we'd missed when developing the original manuscript. We have now altered the language in our paper to avoid suggesting that our study is the first to do any kind of explicit spatial analysis, but (as noted below in the response to the commenter), because we have reservations about the way the referenced study was conducted, we have avoided citing it as an example. However, the suggestions of reviewer 1 (Wangensteen) on this matter are well taken and we have added the appropriate citations and corrected our wording.

Like you, we were curious about the effect of the data subset you mention, and conducted that analysis using the life history data. Unfortunately, we found no meaningful patterns, and have made an attempt to clarify this in the text (L356).

Reviewer 1 (Owen Wangensteen)

Basic reporting

The article is well written. The structure is correct. The introduction and the material and methods sections are a little bit long, but can be easily improved by deleting some sentences (detailed below in this review). The references used are relevant and updated. The raw data and the analysis scripts are all publicly available.

Thank you for the positive feedback. We feel the introduction is appropriately detailed to put our study in the context of both the field of metabarcoding and the field of spatial ecology. We strove for excruciating detail in the methods section as a counter to the chronic under-reporting of details in the metabarcoding field, which hinders reproducible research. We took your later suggestions into

account, but please note that the other reviewer requested more analytical detail in the methods. We would happily move extraneous details to the supplement if the editor feels that manuscript length is an issue as written (e.g. molecular laboratory details such as PCR parameters).

Experimental design

The objectives are clearly established and the experimental design is perfect for addressing the proposed questions. The laboratory methods are rigorous and the analysis pipeline includes steps for assuring excellent quality of the resulting data set.

Thank you; we agree.

Validity of the findings

The findings are very robust. The high level of replication in the PCR analyses and the high-standards bioinformatic filtering steps used for removing any sequence attributable to possible contamination assure an excellent resulting data set to work with. The statistical analysis of the data are very clearly explained and the results are highly trustworthy.

Thank you; we agree.

Comments for the Author

The manuscript is, in general, very well written and the robust results obtained from the enhanced laboratory procedures and bioinformatic pipelines used are presented in a clear manner. I think that the introduction and material and methods sections are a little too long, but they can be enhanced by deleting some sentences (detailed below).

Thank you; please see our previous comment on manuscript length, as well as our responses to your specific comments on extraneous details.

I think that it would be informative to add a paragraph in the discussion section commenting the caveats of using 16S as metabarcoding marker for taxonomic assignment of the MOTUs (i.e. the issues with the assignment of many sequences which can be assigned only to the family or lower levels), compared to other more variable markers with denser reference databases, such as COI, which, in most cases, allow assignments to the genus or species level.

We agree with your point that a locus with only coarse-scale resolution (16S) may obfuscate patterns of diversity compared to a locus with fine-scale resolution. A subset of us has recently published an in-depth comparison of multiple loci for metabarcoding, and we have added a more substantial comment on this and referenced to this now-published work (L452).

Specific Corrections

Some specific corrections are detailed below:

L 36 to 40. This whole paragraph may be deleted to shorten the introduction, without any significant information loss.

We can understand the reviewer's perspective, but feel this point is a stylistic one and would prefer to leave the manuscript as it is currently worded to increase the appeal of our work to a broader audience.

L90. Correct "Actiniaria" and "Bryozoans".

Thanks for catching these errors. We made these changes (L99, L100).

L91. Delete “(Decapoda)”. These habitats hold many other Crustacea other than Decapoda as well. Actually, I would write “barnacles (Sessilia) and other crustaceans” in line 89.

Excellent suggestion; we made this change (L98).

L93: “which provide”

We made this change (L102).

L94: The microscopic plankton is not composed just by “diatoms and larvae”. Many important holoplanktonic organisms (e.g. copepods) could be mentioned here.

We did not mean to imply that the microscopic plankton is composed solely of diatoms and larvae, we made a statement and qualified it with two examples to give meaning and context: ‘a diverse assemblage of microscopic plankton including diatoms and larvae’. We mention diatoms because they are a widely-known, numerically abundant, and ecologically important group that falls outside the scope of this study on metazoans. We mention larvae because it subtly hints at our major finding that our samples captured many organisms best known for their sessile, benthic stages (e.g. barnacles). We added copepods to this list, though we recognize that it is necessarily incomplete (L104).

L100: I am not sure that this is “the first explicitly spatial analysis of eDNA-derived community similarity”, since some other papers have been published which reported patterns of spatial variability in marine systems (although using less quantitative approaches); e.g. De Vargas et al. 2015. I think that the authors should change this to “an explicitly spatial analysis of eDNA-derived community similarity”.

Indeed, our wording did not clearly convey our meaning. Our intent was to indicate the novelty of the experimental design which was intended to capture relatively fine-scale spatial distribution of eDNA in marine waters. We have made the suggested change (L110).

L117 to 122. This whole paragraph can be deleted. This methodology and the steps involved are now commonplace among aquatic biodiversity researchers and need not be enumerated.

We appreciate this perspective, but disagree that the method is commonplace to the extent that it can be omitted. Our perspective is that few ecologists understand eDNA methodology, and will likely skip the details we outline in the methods. We include this very brief overview for those readers in order to give perspective to the complicated methodology.

L166 to 170. Why not use “ μ l” instead of “microliter”?

Our experience is that such characters are often mangled at some point the publication process. The journal is welcome to make this change if they so desire.

L169. I guess the authors mean “1.25 microliter genomic or eDNA template”.

We intended ‘genomic eDNA’ to mean the full DNA extracted from samples, but your suggestion is a good one and we have implemented it (L182).

L171 to 172. Use “95 °C” instead of “95C”.

We made this change (L154, L159, L162, L184, L185).

L195. “(machine NS500615, run 115, flowcell H3LFLAFX)”. I don’t think we need to know these details!

While this does seem like overkill, it is not unheard of for machines or sequencing facilities to develop problems that can bias data, and we included this with the intent of being as transparent as possible. Nevertheless, we removed this information (L209).

L196. “to enhance sequencing depth”. PhiX is added “to increase sequence diversity” and not to enhance sequencing depth. Actually, the more PhiX is spiked-in, the less amplicon reads will be obtained, thus decreasing (not enhancing) sequencing depth.

We are grateful that you have read the manuscript with such a careful eye! Both statements are correct for subtle reasons, and we have incorporated your phrasing to our explanation. Because

PhiX increases sequence diversity, its effect is to increase the number of clusters which can be correctly differentiated from one another on the flowcell, thus increasing sequencing depth. Similar to your statement, the **less** PhiX is spiked-in, the fewer clusters will be correctly differentiated from one another, and the fewer reads passing filter will be obtained, thus also decreasing sequencing depth (L210).

L223. Delete “a” in “This allows for a samples”.

We made this change (L237).

L270-272. I think that these sentences may be confusing due to the use of some terms related to “dispersal range”, “type of fertilization” and “type of larval development”. Some “internally fertilized species” may have larvae with planktonic development and long dispersal capacity (e.g. some barnacles), whereas some broadcast spawners (e.g. some gorgonians) may be surface brooders with low dispersal ability. I think that the main traits affecting “dispersal range” is the type of larval development (planktonic or benthic) and the total planktonic larval duration (PLD), and not the type of fertilization or the strategy for gamete releasing, per se. I think that “internally fertilized species” in L271 could be changed to “benthic brooding species” and “broadcast spawners” could be changed to “species with long planktonic larval development”. Anyway, this semi-quantitative life-history information recorded by the authors in their data tables does not seem to have been used in their further analyses (since only qualitative considerations regarding two brooding species are discussed below).

You make an excellent point that we were not clear in our wording. We did not mean to suggest that we made only one general estimate of dispersal per taxon; rather, we made separate estimates for each stage. That is, even if a species (e.g. *Balanus glandula*) is internally fertilized and has sessile adults in the intertidal, if it has a larval stage with high potential for dispersal, we cannot rule out an expectation that it will be most abundant in the sequence data from samples far from shore due to larval abundance. We have attempted to clarify this in the text (L284). That said, you are also correct that this data is almost unused except to single out *Cymatogaster aggregata* and *Cupolaconcha mero-clista* as taxa for which an a priori expectation based on adult distribution should be met. We added a concluding sentence to the paragraph to explain the reason for these data (L292). We chose to keep the description here to show that we have been diligent in investigating potential factors obscuring taxon-specific patterns.

L285. “333,537.9 ± 112,200.5”. It looks somehow awkward to use decimal values for numbers of reads here. You could round the values to “333,538 ± 112,200”

Because these are summary statistics (mean and standard deviation), we believe the use of decimals is appropriate.

L314-315. You can delete “the probability that two reads chosen at random from a sample belong to different species”, since it has been already explained in the Materials and Methods section.

We made this change (L332).

L344. “For the first time in a marine environment, we document four key patterns”. The authors should definitely use a more modest sentence here. At least the first two patterns they claim to report “for the first time” have been already reported previously in several works in the marine environment. E.g.: Fonseca et al. 2014, Guardiola et al. 2015, Lallias et al. 2015, De Vargas et al. 2015, etc., although possibly never with a sampling design specifically designed to get a quantitative assessment of distance decay.

Our intent was precisely to convey the novelty of the sampling design; we failed to do that clearly. We removed ‘For the first time in a marine environment’ (L264), and we have clarified the novelty of our study in the context of recently published work in the introduction (L76).

Reviewer 2 (Stephen Moss)

Basic reporting

There were a couple of minor grammatical errors in the manuscript. 1) The final sentence (L292-294) of the “Sequence Data Processing (Bioinformatics)” paragraph on page 12 in the Results section.

We corrected this (L309).

2) The first sentence (L359) of the “Communities far from one another tend to be less similar than those that are nearby” paragraph on page 15 of the Discussion section.

It is unclear what the specific error is, but we acknowledge this sentence is poorly worded. We changed this to read ‘We demonstrate that distant locations have less-similar eDNA communities than proximate locations in Puget Sound, a dynamic marine environment’ (L371).

I was a bit confused initially over discussion of massively parallel sequencing of 16S amplicons in relation to metazoan diversity in the abstract, and it wasn’t until into the introduction (L98) that I realised it was referring to 16S sequences from the mtDNA. Perhaps some clarity on this earlier on would be helpful?

This is a good suggestion; we have modified the abstract to clarify this as early as possible (L9).

Perhaps also discussing the reasoning behind the decision to use 16S rDNA from the mtDNA rather than 18S rDNA would be helpful (e.g. Tang et. al., 2012; <http://www.pnas.org/content/109/40/16208>)?

We added a brief explanation of our marker choice to the section Methods:PCR Amplification (L166), and a more general discussion of marker choice to the discussion (L452).

I couldn’t find any reference to the details of the assigned clusters in Figure 3. I thought it might be useful to include a reference to figures or tables, perhaps as supplementary information, outlining the taxonomic composition of the clusters?

We had a similar first reaction to the results of the classification analysis: *Do these correspond to known ecological communities?* The utility of clustering methods is that they simplify the complexity of highly dimensional data; in our case, hundreds to thousands of OTUs, depending on the data subset. While additional analyses, such as PCA on a reduced-dimensionality dataset, may help identify which OTUs driving community differentiation, this result will not necessarily correspond to the clusters reported using PAM, especially considering the uncertainty regarding the appropriate number of clusters in our data. In short, we are unaware of an analytical or visual method for further simplification of the classification results. If you feel this is imperative information, we could potentially modify our stacked bar plot in order to indicate the cluster to which each sample was assigned, though making sense of this result in a spatial context is not straightforward.

I felt the label for Figure 5 needed expanding to provide more detailed information on what it represents, making it more meaningful/understandable as a stand-alone figure.

Thank you for noticing this; we have corrected the legend to be meaningful as a standalone figure (Figure 5 legend).

The raw data is listed as available under Data Availability, though I assume this is still under embargo? The link on page 18 (L463) will need updating upon publication.

Yes; these links were a placeholder until the manuscript is accepted and have now been updated (L514).

A GitHub DOI can be provided by Zenodo (<https://zenodo.org>) and will need updating upon publication.

We are aware of this and will create a GitHub version and DOI of the repository once the final version is obtained.

Experimental design

When detailing how samples were collected (L126-128) under the Environmental Sampling header in the Methods section, I felt that information on how sampling artefacts were mediated (e.g. from L409-414) could have been provided sooner.

Perhaps we misunderstand your comment, but in the relevant part of the methods section, we state: 'Each environmental sample was collected in a clean 1 liter high-density polyethylene bottle, the opening of which was covered with 500 micrometer nylon mesh to prevent entry of larger particles' (L140).

I felt a little more effort could have been made to facilitate the ability to replicate the analyses in this study. The methods section provided good information, but the analyses as detailed on L473 needs more clarity. An outline of the workflow steps and the input/output at each stage would be very helpful here.

In our initial draft, we provided an outline of the bioinformatics procedures in the section Methods:Sequence Data Processing (Bioinformatics) (L213), and document the procedures in painstaking detail in the Supplemental Material section Methods:Bioinformatics (L725). However, we overlooked the inclusion of a direct link to the scripts used to process the raw sequence data, and have now added those (L515). We certainly appreciate any encouragement to move towards more reproducible research, though we are not yet where we would like to be. Any specific recommendations are welcome.

Validity of the findings

In response to the "(3) Richness declines and evenness increases with distance from shore" heading in the Discussion section, I felt that decline in richness and increase in evenness might be expected due to dilution and osmotic factors in the ocean. Perhaps some reference to tidal flushing and dilution might be pertinent here?

This is an excellent point. We have modified the relevant section to incorporate this potential mechanism (L411).

I felt that under the remit of the Ecological Analyses headings throughout the study, there could have been some focus on reporting the alpha diversity and rarefaction analysis in relation to whether effective sequencing depth had been achieved for the samples in question.

We believe we adequately addressed alpha diversity by reporting richness and evenness both in a 2-dimensional and linear spatial context, with accompanying figures. However, we have now also added an OTU accumulation figure to the supplement (Figure 10) and commented on it in the results (L310). We note that our goal is not to recover an absolute measure of alpha diversity, but to make comparisons among sites for standardized sampling effort (i.e. sequencing depth).

Comments for the Author

I thought the paper was very interesting. The study was well designed, implemented, and written. It was useful to clarify the expectations and hypotheses outlined within, however, I felt more effort could have been made to provide recommendations on how some of the limitations in eDNA sampling could be overcome.

Thank you for the kind words. We acknowledge there are many limitations to eDNA sampling; however, we believe that making more general recommendations than those we provide in the discussion is beyond the scope of this study. The more general limitations and recommendations can and do receive regular, detailed attention in the many reviews published on this subject. Of course, if there are specific limitations you feel we have erroneously overlooked and which are within the purview of our results, we are happy to revisit this issue.

The bioinformatics/statistical methods seemed to be very well thought out, and consistent with current best practices in the analyses of this type of sequencing data.

Thank you; we agree.

Comment on the preprint

Bernd Hänfling, Lori Lawson Handley, Dan Read and Ian Winfield

This is well designed study and we would like to congratulate the authors to this data set and its excellent analysis. However, we feel that the authors have made an important omission in the discussion.

Given that the main focus of this paper is to explore the spatial distribution of community eDNA we feel that the authors should consider the results of our study on lake fish communities (Hänfling et al. 2016). In our paper we have used extensive spatial sampling of three transects across our main study lake with a total of 66 individual samples and to our knowledge there is no other data set on macrobial eDNA metabarcoding with a comparable extent of explicit sampling. Our paper confirms that the spatial distribution of individual species reflects their abundance but also shows that the pattern of distribution is not random and ecologically meaningful. Although there might be important differences between freshwater and marine habitats we feel that the results from our study are extremely relevant here and should not be omitted especially when claiming novelty of discussion point 2.

Hänfling B, Lawson Handley L, Read DS, Hahn C, Li J, Nichols P, Blackman RC, Oliver A, Winfield IJ (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology* 25, 3101-3119.

Thank you for your comment and for drawing our attention to your paper. As you note therein, the rate of publication in this field can make it difficult to keep up. We commend your extensive environmental sampling; however, upon reading your paper, we have reservations about your interpretations of the results given the molecular protocols.

We note that you conducted replicate PCRs of each environmental sample, and pooled them prior to sequencing. While this practice is not uncommon in the field, it precludes the ability to test for any effects of PCR on the communities. We believe this is particularly problematic in the case of your study because, as far as we understand your methods, you used indexed ('tagged') primers in a single PCR reaction. We have shown previously that this can bias datasets in such a way that completely obscures any underlying pattern (see: <http://dx.doi.org/10.1371/journal.pone.0148698>). As we report in that study, indexed primers can bias relative read abundance by as much as 70%. That is, in one environmental sample, a species may comprise 70% of the reads in PCR replicates amplified with one indexed primer set, while being completely absent from replicates amplified by another indexed primer set. This would very likely result in the detection of a significant difference by various statistical tests, such as the Chi-squared test you conducted to compare lake basins.

Regardless, we have neutralized our statements on the novelty of our work on the basis of the points raised by your comment and those of the reviewers.