The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data

Elizabeth L. Clare[1], Frédéric J.J. Chain[2], Joanne E. Littlefair[1], Melania E. Cristescu[2]

1. School of Biological and Chemical Sciences, Queen Mary University of London. Mile End Rd. London, E1 4NS, UK

2. Department of Biology, McGill University, 1205 Docteur Penfield, Montréal, Québec, H3A 1B1, Canada

Corresponding Author: EL Clare. e.clare@qmul.ac.uk

Abstract

The combination of DNA barcoding and high-throughput (next-generation) sequencing (metabarcoding) provides many promises but also serious challenges. Generating a reliable comparable estimate of biodiversity remains a central challenge to the application of the technology. Many approaches have been used to turn millions of sequences into distinct taxonomic units. However, the extent to which these methods impact the outcome of simple ecological analyses is not well understood. Here we performed a simple analysis of dietary overlap by skinks and shrews on Ile Aux Aigrettes, Mauritius. We used a combination of filtering thresholds and clustering algorithms on a COI metabarcoding dataset and demonstrate that all bioinformatics parameters will have interacting effects on molecular operational taxonomic unit recovery rates. These effects generated estimates covering two orders of magnitude. However, the magnitude of the effect on a simple ecological analysis was not large and, despite the wide variation, the same ecological conclusion was drawn in most cases. We advise that a conservative clustering programme coupled with larger sequence divergences to define a cluster, the removal of singletons, rigorous length filtering and stringent match criteria for Molecular Identifier tags are preferable to avoid MOTU inflation and that the same parameters be used in all comparative analyses.

Key-words: metabarcoding, DNA barcoding, eDNA, ecological simulations, MOTU

Introduction

Molecular methods of species identification are becoming a pervasive technique from regulatory and legal applications to pure research objectives (e.g. Clare et al. 2014, Cristescu 2014). Metabarcoding is rapidly expanding in application to all areas of ecological research and biodiversity science (Pompanon et al. 2012; Bohmann et al. 2014; Clare 2014; Cristescu 2014, Adamozicz 2015). As these investigations probe new geographic and research areas encountering unknown taxa (Trontelj and Fišer 2009), one of the most difficult research aspects is how to provide a reliable comparable estimate of species counts and biodiversity assessments linking sequences to biological species. Turning millions of sequences into manageable and accurate datasets remains the central challenge to the application of high-throughput (next-generation) sequencing to ecological investigation.

The metabarcoding method refers to the combination of traditional DNA barcoding (Hebert et al. 2003) and high-throughput sequencing technologies. Defining molecular operational taxonomic units (MOTU) is the most common approach to analyse metabarcoding sequences (Floyd et al. 2002). MOTU can be assigned taxonomy using reference databases of known sequences, left as unknowns for statistical analysis, or treated using some combination of these approaches. The advantage of using MOTU is that both known and unknown taxa can be included in analyses. Identifications of MOTU tend to be biased towards larger, more charismatic species that are better known and appear in

reference collections, but unknowns are equally important in most ecological investigations and should dominate in relatively unknown fauna (Trontelj and Fišer 2009). A number of analytical programmes are used to define MOTU (e.g. Caporaso et al. 2010; Jones et al. 2011; Ratnasingham & Hebert, 2013), and most rely on some sort of clustering or threshold approach. As a standard, 3% sequence divergence is often applied and may function well in simple communities (Brown et al. 2015), and is particularly popular in bacterial research where metabarcoding techniques have been used for some time and a default 3% is generally accepted, though this represents a somewhat arbitrary choice (Yang et al. 2013). The problem is more difficult when dealing with more complex datasets of the more recent eukaryotic metabarcoding efforts, where the automatic adoption of workflows from the bacterial literature is ill advised. The problem with MOTU-based approaches is the same as all species concepts, that no rule or metric will apply universally to all genetic markers and all taxonomic groups (Brown et al. 2015). Thus, the MOTU approach is an attempt at a reasonable estimate of species richness and should either be tailored to each dataset uniquely or standardized across datasets for meta-analyses.

Conservative approaches to MOTU definition attempt to reduce the number of MOTU in datasets by increasing the divergence threshold by which MOTU will be defined, eliminating rare and/or artifactual sequences, or removing rare MOTU themselves. (NOTE: We use “increased divergence threshold” to refer to a larger absolute value, e.g. 5% of sequence divergence, to define a MOTU cluster

compared to “decreased threshold”, e.g. 2%, which generates more MOTU and is thus less conservative.) Rare MOTU are MOTU found infrequently within the dataset, usually only once (see a discussion in Salinas-Ramos et al. 2015). A conservative approach may be advisable, as many MOTU programmes appear to overestimate species diversity. For example, using a mock community of 61 zooplankton species, Flynn et al. (2015) tested the effect on MOTU recovery rate of using a variety of programmes and metrics. They found that estimates ranged over orders of magnitude (22-22191) and were particularly influenced by the retention of rare sequences (singletons). This is particularly true when metabarcoding targets are length and copy-number variable regions like ribosomal genes that have different evolutionary properties affecting clustering behaviours and cannot undergo much length filtering. While insertions and deletions are thought to be rarer than substitutions in most high-throughput sequencing profiles (though it is platform dependent), their distribution appears to be non-random with reports of insertions more likely than deletions and concentrations of errors around specific sequence locations (Schirmer et al. 2015). This length variation through sequencing error may artificially increase MOTU estimates depending on how gaps are treated in alignments and clustering methods, and indeed some clustering approaches have opted to ignore any position with a gap or indeterminate base (Jones et al. 2011).

While this may be less of a problem for coding regions like fragments of COI used in DNA barcoding where rigorous length filtering can be applied,

sequencing errors of any kind will have greater impact when divergence thresholds are less conservative. For example, Razgour et al. (2011) estimated a 12% overestimate of lepidopteran diversity in a dietary analysis. A few base pair errors may have marginal effects when the threshold is set at 4% divergence but will increase in their impact if the threshold leads to smaller sequence divergences being meaningful for generating new MOTU (e.g. 2%). Alternatively, a threshold of 6% or 4% will not be influenced strongly by random error but may lump different taxa together in the same MOTU and generate more conservative MOTU estimates. There is thus a trade off between estimates from larger clustering thresholds that risk lumping taxa, and from smaller thresholds that risk artificially increasing MOTU numbers from errors. Many additional informatics steps will similarly alter MOTU detection. The allowance of gaps or substitutions in the recognition of Molecular Identifier (MID) tags and the retention of rare sequences (e.g. singletons) will both increase the MOTU number as they may preferentially include more sequences of lower quality. The consequence of overestimation of MOTU number has impacts on both data interpretation (Clare 2014) and downstream applications such as conservation management (Cristescu 2014). While some of these are obvious, such as the ranking of areas by biodiversity for managements practice, others are less predictable.

One simple ecological analysis is the measurement of species overlap between any two samples. This may be the diet of two predators or the diversity of species occupying two geographic areas. It provides a measure of shared similarity

between samples and can be modeled in simple ecological packages. One method commonly used is Pianka's (1973) measure of niche overlap, which can be modeled in many ways including with the program EcoSim (version 7; http://grayentsminger.com/ecosim.htm). In this program null models are used to test whether the extent of overlap is greater than expected by chance by comparing observed and simulated matrices of randomized MOTU composition using the equation:

$$O_{jk} = \frac{\sum_{i}^{n} p_{ij}\, p_{ik}}{\sqrt{\sum_{i}^{n} p_{ij}^2 \sum_{i}^{n} p_{ik}^2}},$$

where *Pij* is the proportion that resource *i* is of the total resources used by species *j*; *Pik* is the proportion that resource *i* is of the total resources used by species *k*; and *n* is the total number of resource states (total number of MOTU).

Here we are interested in whether trade-offs between MOTU conservatism and artificially increased MOTU estimates have an impact on the outcome of simple ecological analyses and whether choices in data processing alone can lead to alternative interpretations of a simple ecological model. In this investigation we take a real dataset, an analysis of dietary niche overlap by predatory skinks and shrews (Brown et al. 2014), and measure the effect of increasing and decreasing MOTU estimates by using different sequence clustering programmes and

parameters (Figure 1). The resulting data are compared using the Pianka's measure of niche overlap as reported in the original paper. We test the hypothesis that changing MOTU definition parameters has a predictable impact on the outcomes of ecological analyses, and we measure the magnitude of the effect of changing MOTU definition parameters on the outcome. We further make recommendations on what biases such decisions may impose on the outcomes of these analytical methods.

Methods

Sample data

The data used were taken from an analysis of dietary overlap by skinks and shrews on Ile Aux Aigrettes, Mauritius (Brown et al. 2014). In this system endemic Telfair's skinks (*Leiolopisma telfairii*) are thought to be under threat from invasive Asian Musk Shrews (*Suncus murinus*). At some periods of the year they are thought to be mutually predatory on each other's young, but most of the time they are thought to compete for the same insect resources. The primary goal of their analysis was to determine which prey they might share and to what extent their diet overlapped. The sequences were produced by targeting a small fragment from the cytochrome *c* oxidase subunit 1 "DNA barcode" using mini-barcode primers (LCO-1490/Uni-MiniBar-R). All sequencing was performed on a Roche 454 GS-FLX (Roche Applied Sciences) using the emPCR Lib-L method at the Genepool Edinburgh. Their main conclusion was that the two predators

overlap strongly in their use of common prey MOTU but only marginally when all prey MOTU were considered.

Bioinformatics analysis

Raw sequencing data was processed with custom scripts (appendix) that use a combination of software and various thresholds for comparison. Pooled pyrosequencing data were first de-multiplexed using *fastx_barcode_splitter.pl* from the FASTX-toolkit (hannonlab.cshl.edu/fastx_toolkit) based on the forward and reverse MID indices of each sample. Four different thresholds (0, 1, 2, and 3) were used for both the number of mismatches allowed in the MIDs (--mismatches) and the number of non-overlapping bases (--partial), which is similar to allowing indels/gaps. For each resulting sample file, primer adapter and MID removal was performed using *fastx_clipper* in which sequences without primers were discarded (-c). Reads were then dereplicated using -*derep_fulllength* in USEARCH v8.0.1517 (Edgar 2010) and concatenated into a master file for clustering. Additional files were created with the absence of singletons (uniquely occurring sequences within a sample). For each master file, sequences were treated using two different length-filtering criteria. First, reads were only kept if they were between 122bp and 132bp in length. Second, reads were only kept if they were between 126bp and 128bp in length. Clustering of reads into MOTU was performed using two different approaches, UPARSE (Edgar 2013) and SWARM (Mahé et al. 2014). For UPARSE, clustering was performed using five different sequence divergence thresholds (1%, 2%, 3%, 4%,

5%), following the UPARSE manual (see appendix for specific commands). For SWARM the number of differences between the sequences was explored across six thresholds (from 1 to 6). A total of 176 MOTU files were produced from these combinations (Table 1). The SWARM threshold was interpreted as an approximate sequence divergence threshold so that it could be analysed alongside UPARSE data (e.g. 5/122 – 5/132bp = 4.10 - 3.79% divergence). For comparative purposes, the MOTU results from the category with 5bp differences was merged with the 4bp group during ecological simulations as the majority of the sequences are 127bp long, and therefore the majority will be ≈4% divergence threshold and this merging generates a better matched between methods.

Ecological Simulations

We performed ecological simulations using Pianka's niche overlap with 1000 bootstrap replications in EcoSim (as described above) on all datasets where the MOTU count was less than 800 (large values become computationally impractical for simulations). We analysed two cases, one where all data are retained (All-MOTU analysis) and one where rare MOTU found only once in the entire dataset are removed (Common-MOTU analysis). This approach was used in the original paper (Brown et al. 2014) to remove MOTU that may be the result of sequencing error or may represent rare prey not contributing substantially to the diet of either predator. See Figure 1 for an outline of study design.

Statistical Analysis

All analyses were performed in R version 3.2.1 (R Development Core Team 2015). Common-MOTU and all-MOTU datasets were analysed separately. Initially, data were visualised and linear regressions performed to determine the relationship between mean niche overlap and MOTU number. In both the all-MOTU analysis and common-MOTU analysis, MOTU number was log-transformed, so that the data conformed to the assumptions of the model and improved the fit of the model residuals.

A linear model was fitted to examine the effects of different bioinformatics parameters on mean niche overlap in the all-MOTU dataset. Clustering programme, sequence divergence threshold, bp length filtration, and the presence or absence of singleton sequences (and their two-way interactions) were all added to the initial model as explanatory factors. The model was simplified using deletion tests based on partial F tests until a minimal adequate model was achieved (Crawley 2007). Explanatory variables with a p value of <0.05 were retained in the minimal adequate model. Model validation plots were examined for deviations from the assumptions of a linear model.

Results

Number of MOTU and niche overlap

The original paper analysing the diet overlap of skinks and shrews (Brown et al. 2014) used jMOTU (Jones et al. 2011) to describe the MOTU richness. Although

jMOTU is too computationally intensive for many of the larger datasets now produced, it offers a conservative approach by disregarding gaps that can be legitimate taxonomic characters particularly in ribosomal or intron data (Jones et al. 2011). The original authors (Brown et al. 2014) determined the presence of an apparent “barcode gap” (Meyer and Paulay 2005, Ratnasingham and Hebert 2013) using 4bp differences to define separate MOTU. They reported significant diet overlap using Pianka’s simulation (O_{jk} = 0.55, p = 0.0012) when all MOTU were considered, and when rare MOTU were excluded this value increased dramatically to O_{jk} = 0.80 (p = 0.002).

In our analysis, 176 MOTU estimates were generated by manipulating combinations of sequence filtering and clustering parameters including (1) MID match criteria, (2) length filtration, (3) retention or removal of rare (singleton) sequences, (4) clustering programme, and (5) clustering threshold (Figure1, Table 1). MOTU content varied from 54 to 6238. Eight of the 10 highest estimates (all ≥ 2855) were generated from UPARSE, while 9 of the smallest 10 values (all ≤ 75) were generated from SWARM. A total of 136 combinations of clustering parameters generated files with <800 MOTU and were thus considered in our analysis for ecological simulations. Values substantially above this are inefficient to analyse given the computational requirements for randomized matrices in the available programme and thus we report neither overlap nor statistical testing for these. In many cases these are likely not biologically reasonable in any case e.g. ≈16 of these represent clustering with divergences

equivalent to <1% (Table1). Of the analysed outcomes, mean diet overlap was estimated as 0.526-0.623, with 58% of simulations suggesting overlap was statistically significant. No cases suggested overlap was significantly less than expected by chance (resource partitioning).

All-MOTU and Common-MOTU analysis

When all MOTU were considered (rare MOTU retained) we found a significant negative relationship between mean niche overlap and log MOTU number (Figure 2, $F_{1,134} = 26.7$, $p < 0.0001$). Three interactions and one first-order main effect determined mean niche overlap in the linear model. The interaction between clustering programme and the presence or absence of singletons in the dataset was a highly significant determinant of mean niche overlap (Figure 3, $F_{1,128} = 19.2$, $p < 0.0001$), with the SWARM programme interacting with the presence of singletons to produce a lower mean niche overlap than the other treatments. Clustering threshold interacted with the presence or absence of singletons (Figure 4, $F_{1,128} = 44.4$, $p < 0.0001$). Clustering programme and clustering threshold level had a significant interaction (Figure 5, $F_{1,128} = 53.4$, $p < 0.0001$). Finally, 126-128bp read length filtration produced a significantly higher mean niche overlap than 122-132bp (Figure 6, $F_{1,128} = 61.1$, $p < 0.0001$). When only common MOTU were retained (rare MOTU removed), there was a significant positive relationship between logged common MOTU number and mean niche overlap (Figure 2, $F_{1,170} = 37.9$, $p < 0.0001$).

Discussion

Defining molecular operational taxonomic units (MOTU) from the millions of sequences generated in each next-generation sequencing run remains one of the central challenges of metabarcoding. Here we tested the hypothesis that altering the parameters of MOTU clustering impacts the number of MOTU recovered and has predictable impacts on ecological analyses. Our analysis of two predator diets from a single dataset generated 176 variations of MOTU definition spanning two orders of magnitude. This demonstrates that wide variation in taxonomic richness estimates can be created. We found that these estimates had a small but unpredictable impact on the measurement of ecological niche overlap. When all MOTU were included in the analysis, niche overlap dropped as MOTU counts increased. When we considered only common MOTU, there was a positive relationship between MOTU number and niche overlap. All parameters tested altered MOTU counts and thus had measureable effects on estimates of niche overlap, but also interacted with each other to complicate the analysis. Despite these measurable effects, the actual values of niche overlap did not vary greatly, and in the majority of cases the same ecological conclusion, that the two prey species overlap in their use of resources, would have been made regardless of the parameters used. We suggest that while MOTU parameters that are less conservative lead to lower estimates of niche overlap, general ecological conclusions are robust to most parameter choices.

The influence of MOTU number on ecological analyses

The combination of different parameters generated MOTU estimates that varied across two orders of magnitude. While this is considerable, it is not as dramatic as the variation reported by Flynn et al. (2015) on ribosomal genes. In our case, the use of a coding gene region without length variation may buffer the effect somewhat as length filtering can be quite stringent. When all MOTU were used, more conservative MOTU estimates increased the estimate of niche overlap. This is not unexpected since these phenomena are normally sensitive to rare events (Clare 2014), and conservative clustering approaches tend to homogenize samples by lumping taxa. While this negative relationship does exist, the actual values do not vary greatly. Mean overlap estimates were almost all below 0.6 (most between 0.56 and 0.6). In the common-MOTU analysis all measures were above 0.6, and there was a weak positive relationship between MOTU count and niche overlap. This relationship is unexpected but still would not have altered the conclusions of the original paper that the two predators overlapped strongly in their use of common prey items but not as strongly when all MOTU were considered (Brown et al. 2014).

Specific parameter choices

Factors that lead to the retention of more data contribute to larger MOTU counts and should reduce measures of mean niche overlap. In our case this was largely true though the filtering and clustering parameters interacted in some unexpected ways. The retention of singleton sequences, broader sequence length filtering, and more permissive MID match criteria all contributed to increased MOTU

counts. However, interactions between factors made it difficult to tease apart which parameters have a larger effect on MOTU counts. For any clustering threshold (Figure 4) the inclusion of singletons led to a drop in mean niche overlap. The exclusion of rare sequences is a common analytical step, the assumption being that many of these will constitute sequencing error (Kunin et al. 2010). It has been shown using mock community analyses that most are sequencing artefacts, and the inclusion of these requires that the divergence threshold for MOTU clustering be increased (Brown et al. 2015) to maintain good correspondence between MOTU and taxonomic designations (e.g. from 2% to 4%). In some cases authors have found that rare taxa are often only represented within the singletons (Zhan et al. 2013), but they cause massive MOTU inflation at the same time (Flynn et al. 2015). Therefore, the trade-off between keeping bad data and excluding good data is likely balanced towards the latter but will dependent on how important the possibility of rare taxa is to the analysis (Flynn et al. 2015).

Interestingly we did not observe a consistent effect of clustering thresholds in both the SWARM and UPARSE programmes (Figure 5). This is counterintuitive since thresholds closer to 1% should generate larger MOTU counts, reducing mean overlap. A reasonable biological explanation is that the most common prey items shared between predators were disproportionately split into numerous MOTU, artificially amplifying niche overlap. Alternatively, some of the largest files could not be analysed because of computational demands. Large matrices

become computationally difficult in our simulation software thus we excluded any files that generated >800 MOTU for practical reasons. Because of this, the effect here may be due to the exclusion of the largest files, somewhat correcting MOTU counts. The high niche overlap among the UPARSE MOTU at 4% and 5% divergence may partly be technical; clustering with a threshold bigger than 3% divergence with UPARSE is not recommended and involves a slightly different computational procedure compared to lower thresholds (e.g. 1-2%) (appendix, http://www.drive5.com/usearch/manual/uparse_otu_radius.html). Clustering thresholds have been shown to impact measures of community composition. Yang et al. (2013) demonstrated that altering the clustering threshold from 4% to 1% divergence shifted the relative proportion of MOTU assigned to different taxonomic levels in a complex community. It may be prudent to use less conservative clustering thresholds only when rare sequences are excluded. This conservative tactic of reducing MOTU numbers by using thresholds that are based around increased sequence divergence has been used in some dietary studies where DNA degradation and over interpretation of biological effects are both likely and some data reduction is required (e.g. Salinas-Ramos et al. 2015) but should be balanced against the risk of lumping taxa. We also found that more conservative length filtering increased mean overlap (Figure 6). Because this region of COI is not length variable, relaxed filtering will always mean the inclusion of more sequences with errors resulting in the inclusion of spurious MOTU generated from sequences of more divergent length. When including data with known length variation it is important to consider whether gaps lead to the

site being ignored (Jones et al. 2011) or may be treated as single mutations or multiple mutations based on gap length.

UPARSE vs. SWARM

The two clustering programmes UPARSE and SWARM behaved somewhat differently. Most of the largest files were generated from UPARSE and the smallest from SWARM. This may create an anomalous interaction (Figure 3), because some files with singletons included were too large to be practically analysed (see above). Therefore, despite the apparent outcome that UPARSE generated files with larger mean overlap (Figure 3), this could be an effect of having removed all the larger files. SWARM appears to be more conservative, generating smaller MOTU counts. However, it is important to note that a direct comparison between the two programmes is complicated by the different approaches employed, in which UPARSE uses a greedy clustering algorithm and SWARM uses an agglomerative single-linkage-clustering algorithm. Whereas UPARSE assigns reads to MOTU “centroid sequences” based on a global percent divergence threshold, SWARM uses a combined clustering approach of first delineating MOTU based on a sequence difference threshold, followed by MOTU refinement based on the read abundance and structure of the clusters. The initial iterative clustering by SWARM is meant to group similar sequences together in a progressive manner without using a global threshold applied to a centroid sequence, and in this way appears to achieve a more conservative number of MOTU with fewer singletons.

Interpreting ecological analyses

The remaining consideration is how ecological analyses should be interpreted. In this case, increasing conservatism in our estimates of MOTU will have a predictable impact, biasing our analyses towards the detection of resource sharing. This may be beneficial in some cases. It has been argued that metabarcoding approaches are actually too sensitive in predator-prey systems (Clare 2014). Many predators will not be able to discriminate between prey at the species level, and thus they cannot make adaptive decisions and will consume many food items based on encounter frequency (see a discussion in Clare (2014)). In this way, we have a tendency to over interpret the data towards resource partitioning and specific predator choices that may not be a biological reality. If more conservative analyses can counter this by decreasing that likelihood, it may help improve our interpretation of data. Critically here, the interpretation of the system does not change regardless of the parameters selected in almost every case. This raises the question of whether the niche overlap simulation tests are sensitive to resource partitioning for metabarcoding data. Comparisons with the same parameters will be more reliable, but the actual value of the overlap reached should be treated with caution. Further analyses should consider the conditions under which significant partitioning would be recovered using these data types.

Ideally we need to construct and test parameter choices on mock communities based around the taxa of interest or, failing that, general communities that contain similar taxonomic diversity. Currently mock communities are not a common control in these analyses, and we know of few robust analysis of such a community. Failing this we must rely on established parameters and a degree of common sense about the research priorities. In the example of the skinks and shrews, a 4bp threshold was originally used which approximates a 3% divergence (Brown et al. 2014). With highly mixed arthropod communities such as those anticipated for the diet in this case, no threshold is “correct” and thus “consistent” is probably the best choice. The authors’ attempt to find the barcode gap is a reasonable metric but in highly diverse communities this is difficult. There is some argument about what an appropriate threshold for consistency would be. Some have advocated 2-3% for insects, which approximates some of the observations of real communities reported in the literature (e.g. Hajibabaei et al. 2006), though this may actually end up overinflating taxonomic estimates in some next-generation sequencing approaches. Sequencing error is common in these datasets, and while removing rare haplotypes will substantially reduce error, many authors have chosen to use a threshold in the range of 4-6% so that any retained error is caught in the “fuzz” in the outside of MOTU clusters. This risks lumping, but reduces the risk of MOTU inflation. In the area of bacterial metagenomics and metabarcode approaches a standard 3% has been used by default. The biological reality of this threshold has been minimally tested, especially considering the diverse uncultured bacterial world, but this approach

does achieve consistency across studies, making them comparable. Ultimately the choice of threshold will need to be determined by a combination of the taxa, the question being asked, and the level of conservatism desired by the researcher.

We suggest that in the wider context of data analysis the statistical effect of parameter choice is not likely to have a strong impact on the actual ecological conclusions if the error is equal among samples and treatments, particularly if the parameters used are the best estimates that can be made given the circumstances. More crucially we argue that comparisons between datasets generated using different methods are likely meaningless whether it is the analysis of ecological models or the ranking of areas by biodiversity for management decisions. This extends to comparisons of data generated across sequencing platforms with different sequencing depths and error rates and to the parameters used to define MOTU. While the latter may be under the control of a researcher in meta-analyses, the former are set and may constantly be shifting as platforms and chemistry change. Caution is warranted. Any factor which influences the rate at which new MOTU are recovered will cause biased conclusions, and it is thus necessary to always use the same analytical pipeline for comparative analyses.

Conclusions

We demonstrate that all parameters in the bioinformatics analysis of COI metabarcoding data will have interacting effects on MOTU recovery rates and that modifying only a few of these can generate estimates that cover two orders of magnitude from the same input data. However, the magnitude of the effect on a simple ecological analysis is not as large and, despite the wide variation in MOTU estimates, the same ecological conclusions would be drawn in most cases. While the accuracy of MOTU counts may be inadequate, the repeatability of analyses is high. To make more conservative MOTU estimates we suggest the use of a more conservative clustering programme coupled with larger sequence divergence, the removal of singletons, rigorous length filtering, and more stringent MID match criteria. However when the detection of rare variants or taxa is important less conservative choices are desirable.

References

Adamowicz, S.J. 2015. International Barcode of Life: Evolution of a global research community. Genome. **58**(5):151-162.

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., Yu, D.W., and de Bruyn, M. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. Trends Ecol. Evol. **29**(6): 358–367. doi:http://dx.doi.org/10.1016/j.tree.2014.04.003.

Brown, D.S., Symondson, W.O.C., Burger, R., Cole, N., Vencatasamy, D., Clare, E.L., and Montaxam, A. 2014. Dietary competition between the alien Asian Musk Shrew (*Suncus murinus*) and a reintroduced population of Telfair's Skink (*Leiolopisma telfairii*). Mol. Ecol. **23**: 3695–3705.

Brown, E.A., Chain, F.J.J., Crease, T.J., MacIsaac, H.J., and Cristescu, M.E. 2015. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? Ecol. Evol. **5**(11): 2234–2251. doi:10.1002/ece3.1485.

Caporaso, G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Costello, E., Fierer, N., Pena, A., Goodrich, J., Gordon, J., Huttley, G., Kelley, S., Knights, D., Koenig, J., Ley, R., Lozupone, C., McDonald, D., Muegge, B., Pirrung, M., Reeder, J., Sevinsky, J., Turnbaugh, P., Walters, W., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Meth **7**(5): 335–336. Nature Publishing Group. doi:doi: 10.1038/nmeth.f.303.

Clare, E.L. 2014. Molecular detection of trophic interactions: emerging trends, distinct advantages, significant considerations and conservation applications. Evol. Appl. **7**: 1144–1157.

Cristescu, M.E. 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends Ecol. Evol. **29**(10): 566–571. Elsevier Ltd. doi:10.1016/j.tree.2014.08.001.

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinforma. **26** (19): 2460–2461. doi:10.1093/bioinformatics/btq461.

Edgar, R.C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Meth **10**(10): 996–998. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. Available from http://dx.doi.org/10.1038/nmeth.2604.

Floyd, R., Abebe, E., Papert, A., and Blaxter, M. 2002. Molecular barcodes for soil nematode identification. Mol. Ecol. **11**(4): 839–850. England. Available from http://www.ncbi.nlm.nih.gov/pubmed/11972769.

Flynn, J.M., Brown, E.A., Chain, F.J.J., MacIsaac, H.J., and Cristescu, M.E. 2015. Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. Ecol. Evol. **5**(11): 2252–2266. doi:10.1002/ece3.1497.

Hebert, P.D.N., Cywinska, A., Ball, S.L., and DeWaard, J.R. 2003. Biological identifications through DNA barcodes. Proc. R. Soc. B-Biological Sci. **270**(1512): 313–321. England. doi:10.1098/rspb.2002.2218.

Jones, M., Ghoorah, A., and Blaxter, M. 2011. jMOTU and Taxonerator: turning DNA barcode sequences into annotated operational taxonomic units. PLoS One **6**(4): e19259. doi:10.1371/journal.pone.0019259.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. **12**(1): 118–123. Blackwell Publishing Ltd. doi:10.1111/j.1462-2920.2009.02051.x.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. 2014. Swarm: robust and fast clustering method for amplicon-based studies. PeerJPrePrints: 2:e593.

Meyer, C.P., Paulay, G. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. **3**: 2229–2238.

Pompanon, F., Deagle, B.E., Symondson, W.O.C., Brown, D.S., Jarman, S.N., and Taberlet, P. 2012. Who is eating what: diet assessment using next generation sequencing. Mol. Ecol. **21**: 1931–1950. Blackwell Publishing Ltd. doi:10.1111/j.1365-294X.2011.05403.x.

R Development Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ratnasingham, S., and Hebert, P.D.N. 2013. A DNA-based registry for all animal species: the barcode index number (BIN) system. PLoS One **8**(7): e66213. doi:10.1371/journal.pone.0066213.

Razgour, O., Clare, E.L., Zeale, M.R.K., Hanmer, J., Schnell, I.B., Rasmussen, M., Gilbert, T.P., and Jones, G. 2011. High-throughput sequencing offers insight into mechanisms of resource partitioning in cryptic bat species. Ecol.

Evol. **1**(4): 556–570. doi:10.1002/ece3.49.

Salinas-Ramos, V.B., Herrera Montalvo, L.G., León-Regagnon, V., Arrizabalaga-Escudero, A., and Clare, E.L. 2015. Dietary overlap and seasonality in three species of mormoopid bats from a tropical dry forest. Mol. Ecol. **24**(20): 5296–5307. doi:10.1111/mec.13386.

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., and Quince, C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. **5**: 2234–2251. doi:10.1093/nar/gku1341.

Trontelj, P., and Fišer, C. 2009. Cryptic species diversity should not be trivialised. Syst. Biodivers. **7**(01): 1–3.

Yang, C.X., Ji, Y.Q., Wang, X.Y., Yang, C.Y., and Yu, D.W. 2013. Testing three pipelines for 18S rDNA-based metabarcoding of soil faunal diversity. Sci. China Life Sci. **56**(1): 73–81. doi:10.1007/s11427-012-4423-7.

Zhan, A., Hulák, M., Sylvester, F., Huang, X., Adebayo, A.A., Abbott, C.L., Adamowicz, S.J., Heath, D.D., Cristescu, M.E., and MacIsaac, H.J. 2013. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. Methods Ecol. Evol. **4**(6): 558–565. doi:10.1111/2041-210X.12037.

Table 1. Number of MOTU using a combination of different programmes (UPARSE and SWARM) with various clustering divergence thresholds, MID mismatches, length filtering criteria, and singleton removal. Darker backgrounds represent higher MOTU numbers.

length			122-132bp				126-128bp			
singletons			absence		presence		absence		presence	
clustering			UPARSE	SWARM	UPARSE	SWARM	UPARSE	SWARM	UPARSE	SWARM
Clustering divergence threshold & number of MID mismatches allowed (mm)	1	0 mm	185	189	4826	2349	121	148	2723	1133
		1 mm	189	196	4988	2429	127	158	2820	1182
		2 mm	216	227	6145	2962	145	184	3429	1399
		3 mm	216	227	6238	3018	141	178	3456	1404
	2	0 mm	185	125	2224	984	121	94	1036	465
		1 mm	189	133	2276	1023	127	101	1082	486
		2 mm	216	146	2855	1213	145	115	1298	546
		3 mm	216	146	2872	1267	141	112	1325	570
	3	0 mm	185	111	1280	510	121	84	611	271
		1 mm	189	115	1340	541	127	88	637	293
		2 mm	216	125	1626	634	145	99	765	321
		3 mm	216	125	1656	663	141	96	768	331
	4	0 mm	139	100	963	320	84	74	442	184
		1 mm	138	100	970	349	88	75	449	197
		2 mm	162	110	1184	391	103	85	518	216
		3 mm	166	111	1204	409	101	83	523	226
	5	0 mm	124	89	748	236	78	64	335	139
		1 mm	124	90	738	253	75	66	345	150
		2 mm	134	100	896	283	93	76	391	162
		3 mm	143	99	921	288	93	73	410	164
	6	0 mm		79		189		54		117
		1 mm		80		202		57		124
		2 mm		90		224		68		135
		3 mm		90		234		67		141

Figure 1: A diagram demonstrating the variables and analytical steps used in the study design.

Figure 2: Regression between log MOTU and mean niche overlap for both all-MOTU analysis (rare MOTU are retained in the dataset, represented by filled circles, $y = -0.0131 x + 0.641$, $p < 0.0001$) and common-MOTU analysis (rare MOTU are removed from the dataset, open circles, $y = 0.0129 x + 0.651$, $p < 0.0001$).

Figure 3: Interaction between clustering programme (SWARM and UPARSE) and the presence or absence of singletons in the dataset on mean niche overlap; bars are 95% confidence intervals.

Figure 4: Significant interaction between clustering divergence threshold (1-5%) and the presence or absence of singletons on mean niche overlap; bars are 95% confidence intervals. Note: singletons at clustering threshold = 1 generated files too large for analysis and are thus excluded.

Figure 5: Significant interaction between clustering divergence threshold (1-5%) and clustering programme (SWARM or UPARSE) on mean niche overlap; bars are 95% confidence intervals.

Figure 6: Significant effect of read length filtering on mean niche overlap; bars are 95% confidence intervals.