BMC Bioinformatics

RESEARCH ARTICLE

Open Access

To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography

Adrià Antich[1], Creu Palacin[2], Owen S. Wangensteen[3*] and Xavier Turon[1*]

*Correspondence: owen.wangensteen@uit.no; xturon@ceab.csic.es
[1] Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB-CSIC), Blanes (Girona), Catalonia, Spain[3] Norwegian College of Fishery Science, UiT The Arctic University of Norway, Tromsö, Norway
Full list of author information is available at the end of the article

Abstract

Background: The recent blooming of metabarcoding applications to biodiversity studies comes with some relevant methodological debates. One such issue concerns the treatment of reads by denoising or by clustering methods, which have been wrongly presented as alternatives. It has also been suggested that denoised sequence variants should replace clusters as the basic unit of metabarcoding analyses, missing the fact that sequence clusters are a proxy for species-level entities, the basic unit in biodiversity studies. We argue here that methods developed and tested for ribosomal markers have been uncritically applied to highly variable markers such as cytochrome oxidase I (COI) without conceptual or operational (e.g., parameter setting) adjustment. COI has a naturally high intraspecies variability that should be assessed and reported, as it is a source of highly valuable information. We contend that denoising and clustering are not alternatives. Rather, they are complementary and both should be used together in COI metabarcoding pipelines.

Results: Using a COI dataset from benthic marine communities, we compared two denoising procedures (based on the UNOISE3 and the DADA2 algorithms), set suitable parameters for denoising and clustering, and applied these steps in different orders. Our results indicated that the UNOISE3 algorithm preserved a higher intra-cluster variability. We introduce the program DnoisE to implement the UNOISE3 algorithm taking into account the natural variability (measured as entropy) of each codon position in protein-coding genes. This correction increased the number of sequences retained by 88%. The order of the steps (denoising and clustering) had little influence on the final outcome.

Conclusions: We highlight the need for combining denoising and clustering, with adequate choice of stringency parameters, in COI metabarcoding. We present a program that uses the coding properties of this marker to improve the denoising step. We recommend researchers to report their results in terms of both denoised sequences (a proxy for haplotypes) and clusters formed (a proxy for species), and to avoid collapsing the sequences of the latter into a single representative. This will allow studies at the cluster (ideally equating species-level diversity) and at the intra-cluster level, and will ease additivity and comparability between studies.

Keywords: Metabarcoding, Metaphylogeography, COI, Denoising, Clustering, Operational taxonomic units

Background

The field of eukaryotic metabarcoding is witnessing an exponential growth, both in the number of communities and substrates studied and the applications reported (reviewed in [1–4]). In parallel, technical and conceptual issues are being discussed (e.g., [5, 6]) and new methods and pipelines generated. In some cases, however, new practices are established after a paper reporting a technique is published and followed uncritically, sometimes pushing its application outside the context in which it was first developed.

A recently debated matter concerns the treatment of reads by denoising procedures or by clustering techniques [7]. Both methods are often presented as alternative approaches to the same process (e.g., [7–11]). However, both are philosophically and analytically different [12]. While denoising strives to detect erroneous sequences and to merge them with the correct "mother" sequence, clustering tries to combine a set of sequences (without regard to whether they contain or not errors) into meaningful biological entities, ideally approaching the species level, called OTUs or MOTUs (for Molecular Operational Taxonomic Units). Usually only one representative sequence from each MOTU is kept (but note that this is only common practice, not a necessary characteristic of the method). Thus, while both procedures result in a reduced dataset and in error correction (by merging reads of erroneous sequences with the correct one or by combining them with the other reads in the MOTU), they are not equivalent. More importantly, they are not incompatible at all and can (and should) be used together.

A recent paper [13] proposes that denoised sequences should replace MOTUs as the unit of metabarcoding analyses. We contend that it may be so for ribosomal DNA datasets such as the one used in that paper, but this notion has gained momentum also in other fields of metabarcoding for which it is not adequate. In particular, when it comes to highly variable markers such as COI. This proposal misses the fact that sequence clusters are a proxy for species-level entities, the basic unit in eukaryotic biodiversity studies. The 3′ half (also called Leray fragment) of the standard barcode fragment of COI (Folmer fragment) is becoming a popular choice for metabarcoding studies addressed at metazoans or at eukaryotic communities at large [14], reaching now 28% of all metabarcoding studies [15]. Metabarcoding stems from studies of microbes where 16S rRNA is the gene of choice, and the concept was then applied to analyses of the 18S rRNA gene of eukaryotes. With the recent rise of COI applications in metabarcoding, programs and techniques developed for rDNA are sometimes applied to COI without reanalysis and with no parameter adjusting given the highly contrasting levels of variation of these markers.

The idea that denoising should be used instead of clustering has been followed by some (e.g., [16–20]), while other authors have combined the two approaches (e.g., [21–23]). Indeed, denoising has the advantages of reducing the dataset and to ease pooling or comparing studies, which is necessary in long term biomonitoring applications. However, with COI there is a wealth of intraspecific information that is missed if only denoising is applied [24]. COI has been a prime marker of phylogeographic studies to date [25, 26], and these studies can be extended to metabarcoding datasets by mining the distribution of haplotypes within MOTUs (metaphylogeography [12]). The latter authors suggested to perform clustering first, and that denoising should be done within MOTUs to provide the right context of sequence variation and abundance skew. They also advised to perform a final abundance filtering step. In other studies, denoising is performed first, followed by clustering and refining steps (e.g., [22, 23]).

There are several methods for denoising (reviewed in [27]) and for clustering (reviewed in [28]). We will use two of the most popular denoising techniques, based on the DADA2 algorithm (Divisive Amplicon Denoising Algorithm, [29]) and the UNOISE3 algorithm [30]. The results of the former are called Amplicon Sequence Variants (ASVs) and those of the latter ZOTUs (zero-radius OTUs). In practice, the terminology is mixed and ASV, ZOTU, ESV (Exact Sequence Variant), sOTU (sub-OTU) or ISU (Individual Sequence Variant), among others, are used more or less interchangeably. For simplicity, as all of them are equivalent, we will use henceforth the term ESV. Clustering, on the other hand, can be performed using similarity thresholds (e.g., [31, 32]), Bayesian Methods (CROP, [33]), or methods based on single-linkage-clustering (SWARM, [34]), among others. We will focus on de novo clustering methods (i.e., independent of a reference database), while denoising is always de novo by its very nature [13]. We will here use SWARM as our choice of clustering program due to its good performance compared to other methods [28]. It is noteworthy that all these programs were originally developed and tested on ribosomal DNA datasets. When applied to other markers, often no indication of parameter setting is given (i.e., omega_A for DADA2, α for UNOISE3, d for SWARM), suggesting that default parameter values are used uncritically.

In this article, we aim to use a COI metabarcoding dataset of benthic littoral communities to (1) set the optimal parameters of the denoising and clustering programs for COI markers, (2) compare results of the DADA2 algorithm with the UNOISE3 algorithm, (3) compare the results of performing only denoising, only clustering, or combining denoising with clustering in different orders, and (4), suggest and test improvements in the preferred denoising algorithm to take into account the fact that COI is a coding gene. We implement these modifications in the new program DnoisE. Our aims are to provide guidelines for using these key bioinformatic steps in COI metabarcoding and metaphylogeography. The conceptual framework of our approach is sketched in Fig. 1.

Methods

The dataset

We used as a case study an unpublished dataset of COI sequences obtained from benthic communities in 12 locations of the Iberian Mediterranean. Information on the sampling and sample processing is given in Additional File 1. Sequences were obtained in a full run of an Illumina MiSeq (2 * 250 bp paired-end reads).

Bioinformatic analyses

The initial steps of the bioinformatic pipeline followed [12] and were based on the OBItools package [35]. Reads were paired and quality filtered, demultiplexed, and dereplicated. A strict length filter of 313 bp was used. We also eliminated sequences with only one read. Chimera detection was performed on the whole dereplicated

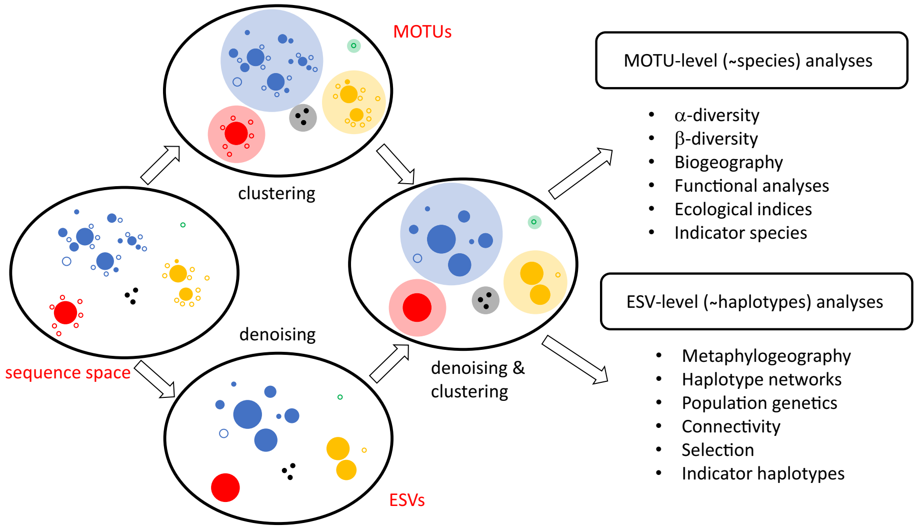

Fig. 1 Conceptual overview of the denoising and clustering processes. The oval on the left sketches a fragment of the sequence space with four biological species plus an artefact divergent sequence (denoted by colours). Correct sequences are indicated by filled circles and artefacts by empty circles, with indication of abundance (circle size). Denoising results in the detection of putatively correct sequences to which the reads of putatively incorrect sequences are merged (leading to a reduced dataset). The outcome of denoising should ideally approach the true haplotype composition of the samples. Clustering generates MOTUs without regard as to whether the grouped sequences are erroneous or not. This is usually accompanied by read pooling and keeping only one representative sequence per MOTU (leading to a reduced dataset). The outcome of clustering should ideally approach the species composition of the samples. Combining both processes results in a dataset that is reduced in size, comparable across studies, and amenable to analyses at the MOTU (species) and ESV (haplotype) levels. Note that errors likely persist in the final dataset both as artefact MOTUs and artefact ESVs within MOTUs, and carefully designed filters should be used to minimize them (abundance filtering, chimera filtering, numts removal)

dataset with uchime3_denovo as embedded in unoise3 (USEARCH 32-bit free version, [36]). We used minsize = 2 to include all sequences. Those identified as chimeras were recovered from the –tabbedout file and eliminated from the dataset. Sequences with small offsets (misaligned), identified as shifted in the output, were likewise deleted. The working dataset thus comprised well-aligned, chimera-free, unique sequences which had appeared with at least two reads in the samples.

Note that for this technical study we didn't consider the sample distribution of the reads. A complete biogeographic study of the samples is ongoing and will be published elsewhere. For the present analysis, for each unique sequence only the actual DNA sequence and the total number of reads were retained.

The denoisers: UNOISE3 and DADA2

Comparing denoising algorithms is challenging because each method comes with a different software suite with embedded features and recommendations [27]. For instance, uchime3_de novo is embedded in the unoise3 command as implemented in USEARCH, while a chimera removal procedure (removeBimeraDenovo) is an optional feature in the DADA2 pipeline. Furthermore, while UNOISE3 uses paired reads, DADA2 recommends denoising forward and reverse reads separately, and then performing a merging step. We have tried to isolate the algorithms from their pipelines for comparability. This was done by generating a Python script [37] that implements the algorithm described in [30] and by using DADA2 from its R package v. 1.14.1 and not as embedded into the qiime2 pipeline [38].

For UNOISE3, our program (henceforth DnoisE) was compared on the working dataset described above with command unoise3 in USEARCH with minsize = 2, alpha = 5 and without the otutab step. That is, we recovered the ESV composition and abundance with an R script directly from the output of unoise3 (using the output files –tabbedout and –ampout), without a posterior re-assignment of sequences to ESVs via otutab. This step was not necessary as all sequences were included in the ESV calculations. The results of DnoisE and unoise3 were > 99.99% identical in ESVs recovered and reads assigned to them, so we continued to use our script for performing the comparisons and for further improvements of the algorithm (see below).

The recommended approach for DADA2 is to denoise separately the forward and reverse reads of each sequence. This complicates the technical comparison, as all initial filtering steps cannot be equally performed (e.g., we won't know if there is just one read of a particular sequence, or if the merged pair will be discarded for low quality of the assembly or for unsuitable final length) and thus we cannot have two identical starting datasets. More importantly, we cannot use this procedure when we test the effects of denoising at later steps (i.e., after clustering), so we would be unable to compare the denoisers at this level. Thus, for our comparative analysis we need to use DADA2 on paired reads. According to Callahan et al. [29], this can result in a loss of accuracy, but this point has never been tested to our knowledge. We addressed this issue by comparing denoising before and after pairing on half of the reads in the final dataset. After this analysis, we decided to continue our comparison of DADA2 and UNOISE3 on paired reads.

Additionally, denoising before pairing is not optimal if a PCR-free library preparation protocol is used, as in our case, because half of the reads are in one direction and the other half are in the opposite direction (hence the use of half of the reads in the above comparison). Forward and reverse reads can of course be recombined to generate new files with all reads in the same direction, but the quality of the reads with original forward and reverse orentation is different. Alternatively, two rounds of DADA2 (one per orientation) must be performed and combined at later steps.

To run DADA2 on paired reads, we entered them in the program as if they were the forward reads and did not use a merging step after denoising. In all DADA2 runs we did not perform the recommended chimera removal procedure as the input sequences were already chimera-free according to uchime3_de novo. Note that, when denoising was done after clustering, we used error rates calculated for the whole dataset, and not for each MOTU separately (most of them do not have enough number of sequences for a reliable estimation of error rates).

UNOISE3 relies heavily on the stringency parameter α, which weights the distance between sequences as a function of the number of differences between them [30]. In short, lower values of α tend to merge sequences more strongly, while higher values recovered higher numbers of ESVs. The default, and the value used in most studies with ribosomal DNA, is 2. However, for COI three independent approaches, based on mock communities [39], entropy changes [12], and co-sequenced control DNA [40] suggested that for this marker $\alpha = 5$ is the optimal value. For DADA2 the key parameter is omega_A, which indicates the probability threshold at which a sequence *i* is considered an error derived from another sequence *j* given their abundance values and the inferred error rates. If the observed value is higher than omega_A, then sequence *i* is considered an error of sequence *j*. Omega_A is by default set to a very low value (10^{-40}), but no study has analysed the impact of changing this parameter for COI datasets. To our knowledge, only [41], based on a comparison of 3 values, concluded that the default value of omega_A was adequate for a marker based on the control region of the mitochondrial DNA.

The clustering algorithm

Our preferred clustering method is SWARM v3 [42], as it is not based on a fixed distance threshold and is independent of input order. It is a very fast procedure that relies on a single-linkage method with a clustering distance (*d*), followed by a topological refining of the clusters using abundance structures to divide MOTUs. As we were interested in keeping all sequences within MOTUs, and not just a representative sequence, we mined the SWARM output with an R script to generate MOTU files, each with its sequence composition and abundance.

The crucial parameter in this approach is *d*, the clustering distance threshold for the initial phase. The default value is 1 (that is, amplicons separated by more than one difference will not be clustered together), and this value has been tested in ribosomal DNA. However, Mahé et al. [42] pointed out that higher *d* values can be necessary for fast evolving markers (such as COI) and advised to analyse a range of *d* to identify the best fitting parameter (i.e., avoiding over- or under-clustering) for a particular dataset or scientific question. A *d* value of 13 (thus, allowing 13 differences over ca. 313 bp to make a connection) has been recently used for the Leray fragment of COI (e.g., [43–47]), but a formal study of its adequacy has not been published yet.

Setting the right parameters

With our dataset, we assessed the best-fitting parameters for UNOISE3, DADA2 and SWARM as applied to COI data. For the first two, we used changes in diversity values per codon position (measured as entropy, [48]), as calculated with the R package *entropy* [49]. Coding sequences have properties that can be used in denoising procedures [12, 41]. They have naturally a high amount of variation concentrated in the third position of the codons, while errors at any step of the metabarcoding pipeline would be randomly distributed across codon positions. Thus, examining the change in entropy values according to codon position can guide the choice of the best cleaning parameters. Turon et al. [12] suggested to use the entropy ratio (Er) between position 2 of the codons (least variable) and position 3 (most variable). In a simulation study these authors showed that Er decreased as more stringent denoising was applied until reaching a plateau, which was taken as the indication that the right parameter value had been reached.

Using the Er to set cut-points, we re-assessed the adequate value of α in UNOISE3 testing the interval of $\alpha = 1$ to 10. With the same procedure, we tested DADA2 for values of omega_A between $10^{-0.05}$ (ca. 0.9) and 10^{-90}.

For SWARM, we compared the output of SWARM with a range of values of d from 1 to 30 applied to our dataset (prior to denoising). We monitored the number of MOTUs generated and the mean intra- and inter-MOTU distances to find the best-performing value of d for our fragment.

The impact of the steps and their order

With the selected optimal parameters for each method, we combined the two denoising procedures and the clustering step in different orders. We therefore combined denoising (Du for UNOISE3 algorithm implemented in DnoisE, Da for DADA2) and clustering with SWARM (S) and generated and compared datasets of ESVs and MOTUs as follows (for instance, Da_S means that the dataset was first denoised with DADA2, then clustered with SWARM):

ESVs: Du, Da
MOTUs: Du_S, Da_S, S_Du, S_Da

For comparison of datasets, we used Venn diagrams and an average match index of the form

$$\text{Match Index (A,B)} = \left(\mathrm{N}_{\mathrm{match\_A}}/\mathrm{N}_{\mathrm{A}} + \mathrm{N}_{\mathrm{match\_B}}/\mathrm{N}_{\mathrm{B}}\right)/2$$

where $\mathrm{N}_{\mathrm{match\_A}}$ is the number of a particular attribute in dataset A that is shared with dataset B, and N_{A} is the total number of that attribute in dataset A. The same for $\mathrm{N}_{\mathrm{match\_B}}$ and N_{B}. The matches can be the number of ESVs shared, the number of MOTUs shared, the number of ESVs in the shared MOTUs, or the number of reads in the shared ESVs or MOTUs, depending on the comparison.

Improving the denoising algorithm

The preferred denoising algorithm (UNOISE3, see Results) has been further modified in two ways. Let i be a potential error sequence derived from sequence j. The UNOISE3 procedure is based on two parameters: the number of sequence differences between i and j (d, as measured by the Levenshtein distance) and the abundance skew (β, abundance i/abundance j) between them. These parameters are related by the simple formula [30]:

$$\beta(\mathrm{d}) = 1/2^{\alpha d+1}$$

where β(d) is the threshold abundance skew allowed between two sequences separated by distance d so that below it the less abundant would be merged with the more abundant, and α is the stringency parameter. Thus, presumably incorrect “daughter” sequences are merged with the correct “mother” sequences if the number of sequence differences (d) is small and the abundance of the incorrect sequence with respect to the correct one (abundance skew) is low. The higher the number of differences, the lower the skew should be for the sequences to be merged.

For COI, however, the fact that it is a coding gene is a fundamental difference with respect to ribosomal genes. In a coding fragment, the amount of variability is substantially different among codon positions. This is not considered in the UNOISE3 formulation (nor in DADA2 or other denoising programs that we knew of, for that matter). We suggest to incorporate this information in DnoisE by differentially weighting the d values according to whether the change occurs in the first, second, or third codon position. Note that our sequences are all aligned and without indels, which makes this weighting scheme straightforward. The differences in variability can be quantified as differences in entropy values [48]; position 3 of the codons has the highest entropy, followed by position 1 and position 2. In other words, two sequences separated by n differences in third positions are more likely to be naturally-occurring sequences than if the n differences happen to occur in second positions, because position 3 is naturally more variable. To weight the value of d, we first record the number of differences in each of the three codon positions (d(1) to d(3)), we then correct the d value using the formula

$$\mathrm{d_{corr}} = \sum_{\mathrm{i}=1}^{3} \mathrm{d(i)}^{*}\mathrm{entropy(i)}^{*}3/\big(\mathrm{entropy(1)} + \mathrm{entropy(2)} + \mathrm{entropy(3)}\big)$$

where i is the position in the codon, and $\mathrm{d_{corr}}$ is the corrected distance that will be used in the UNOISE3 formula instead of d.

With this formula, two sequences separated by just one difference in each codon position will continue to have a d of 3, but a change in a high entropy position (3) will translate in a higher d than the same change in a low entropy position (2), thus the program will tend to keep the former and to merge the later. The entropy of the three positions of the codons for the weighting was obtained from the original dataset prior to any denoising, thus entropy(1) = 0.473, entropy(2) = 0.227, and entropy(3) = 1.021. Note that d(i) is based on the number of differences occurring at each codon position. The Levenshtein distance used in the non-corrected d measures is not adequate for this purpose, as it cannot keep track of codon positions. However, for sequences of equal length, aligned, and without indels, as in our case, the number of differences is in practice equivalent to the Levenshtein distance.

The present algorithm of UNOISE3 gives precedence to the abundance skew over the number of differences (d) because sequences are considered in order of decreasing abundance. Thus, a very abundant sequence will form a centroid that can "capture" a rare one even if d is relatively high. Other, somewhat less abundant, sequences can be more similar (less d) to the rare sequence and can fulfil the conditions to capture it, but this will never happen as the rare sequence will be incorporated to the first centroid and will become unavailable for further comparisons. In our modification, DnoisE does not automatically join sequences to the first centroid that fits the condition. Rather, for each sequence the potential "mothers" are stored (with their abundance skew and d) and the sequences are left in the dataset. After the round of comparisons is completed, for each daughter sequence we can choose, among the potential mothers, the one whose abundance skew is lower (precedence to abundance skew, corresponding to the usual UNOISE3 procedure), the mother with the lowest distance (precedence to d), or the one for which the ratio (abundance skew/max abundance skew for the observed d, β(d)) is lower, thus combining the two criteria.

We compared in our dataset the results of the different formulations of DnoisE: precedence to abundance skew, precedence to distance, combined precedence, and correcting distances according to codon position of the differences. A beta version of DnoisE is available from [37].

Benchmarking

Ground truthing is a difficult task in metabarcoding studies. Constructing mock communities is the most common method. However, mock communities, even the largest ones, are orders of magnitude simpler than complex biological communities. Thus, some technical aspects cannot be tested accurately. For instance, metabarcoding results of mock communities in general lack true sequences at very rare abundances (the most problematic ones). For complex communities, we need to rely on metrics that can evaluate the fit of denoising and filtering procedures. The coding properties of COI can help design useful parameters, such as the entropy ratio mentioned above. Another possible metric stems from the evaluation of the prevalence of incorrect ESVs (defined by having indels or stop codons) across denoising and filtering procedures [50].

In this work, we have performed two benchmarking procedures that rely on taxonomic assignment of the MOTUs. This assignment was done using the ecotag procedure in OBItools against the db-COI_MBPK database [51], containing 188,929 eukaryote COI reference sequences (available at [52]). Ecotag assigns a sequence to the common ancestor of the candidate sequences selected in the database, using the NCBI taxonomy tree. This results in differing taxonomic rank of the assignments depending on the density of the reference database for a given taxonomic group.

First, we checked the performance of the entropy correction of DnoisE by examining the percent of incorrect to total ESVs. To this end, we retained only the MOTUs assigned to metazoans and, following [12], examined the presence of stop codons and changes in the 5 aminoacids present in the fragment amplified that are conserved among metazoans [53]. To be on the conservative side, for a given MOTUs we evaluated the different genetic codes and selected the ones that produced the smaller number of stop codons. The five aminoacids were then checked using these codes and the minimal number of "wrong" aminoacids was recorded. The R package Biostrings [54] was used for the translations. The ESVs featuring stop codons and/or aminoacid changes in the five conserved positions were labelled as erroneous. The rationale is that a suitable denoising procedure would reduce the ratio of error vs total ESVs.

Second, we performed a taxonomic benchmarking. As MOTUs should ideally reflect species-level entities, we selected those sequences assigned at the species level as a benchmark for the MOTU datasets. We also enforced a 97% minimal best identity with the reference sequence. We traced these sequences in the output files of our procedures and classified the MOTUs containing them into three categories (following the terminology in [9]): closed MOTUs, when they contain all sequences assigned to a species and only those; open MOTUs, when they contain some, but not all, sequences assigned to one species and none from other species, and hybrid MOTUs. The latter included MOTUs with sequences assigned to more than one species, or MOTUs with a combination of sequences assigned to one species and sequences not assigned (i.e., they don't have species-level assignment, or they do with less than 97% similarity).

This analysis was intended as a tool for comparative purposes, to benchmark the ability of the different MOTU sets generated to recover species-level entities. In other words, which procedure retains more ESVs with species-level assignment and places them in closed (as opposed to open or hybrid) MOTUs.

Results

The dataset

After pairing, quality filters, and retaining only 313 bp-long reads, we had a dataset of 16,325,751 reads that were dereplicated into 3,507,560 unique sequences. After deleting singletons (sequences with one read), we kept 423,164 sequences (totalling 10,305,911 reads). Of these sequences, 92,630 were identified as chimeras and 152 as misaligned sequences and eliminated. Our final dataset for the study, therefore, comprised 330,382 sequences and 9,718,827 reads (the original and the refined datasets were deposited in Mendeley Data, [55]).

For testing the performance of DADA2 on unpaired and paired reads on a coherent dataset, we selected the reads that were in the forward direction, that is, the forward primer was in the forward read (R1). As expected, they comprised ca. half of the reads (4,892,084). For these reads we compared the output of applying DADA2 before and after pairing, as detailed in Additional File 2. The results were similar, with most reads placed in the same ESVs in both datasets, albeit 21% more low-abundance ESVs were retained using the paired reads. Henceforth we will use DADA2 on paired sequences, as this was necessary to perform our comparisons.

Setting the right parameters

We used the change in entropy ratio (Er) of the retained sequences of the global dataset (330,382 sequences and 9,718,827 reads) for selecting the best performing α–value in UNOISE3 and the best omega_A in DADA2 across a range of values. We also assessed the number of ESVs resulting from the procedures.

For UNOISE3 as implemented in our DnoisE script, the Er diminished sharply for α–values of 10 to 7, and more smoothly afterwards (Fig. 2a). The number of ESVs detected likewise decreased sharply with lower α–values, but tended to level off at $\alpha = 5$ (Fig. 2a). The value of 5 seems a good compromise between minimizing the Er and keeping the maximum number of putatively correct sequences.

For the DADA2 algorithm we tested a wide range of omega_A from $10^{-0.05}$ to 10^{-90} (we set parameter omega_C to 0 in all tests, so all erroneous sequences were corrected). The results showed that, even at the highest value ($10^{-0.05}$, or ca. 0.9 p-value, thus accepting as new partitions a high number of sequences), there was a substantial drop in number of sequences (ca. 75% reduction) and in Er with respect to the original dataset (Fig. 2b). Both variables remained relatively flat with a slight decrease between omega_A 10^{-2} and 10^{-15}, becoming stable again afterwards (Fig. 2b).

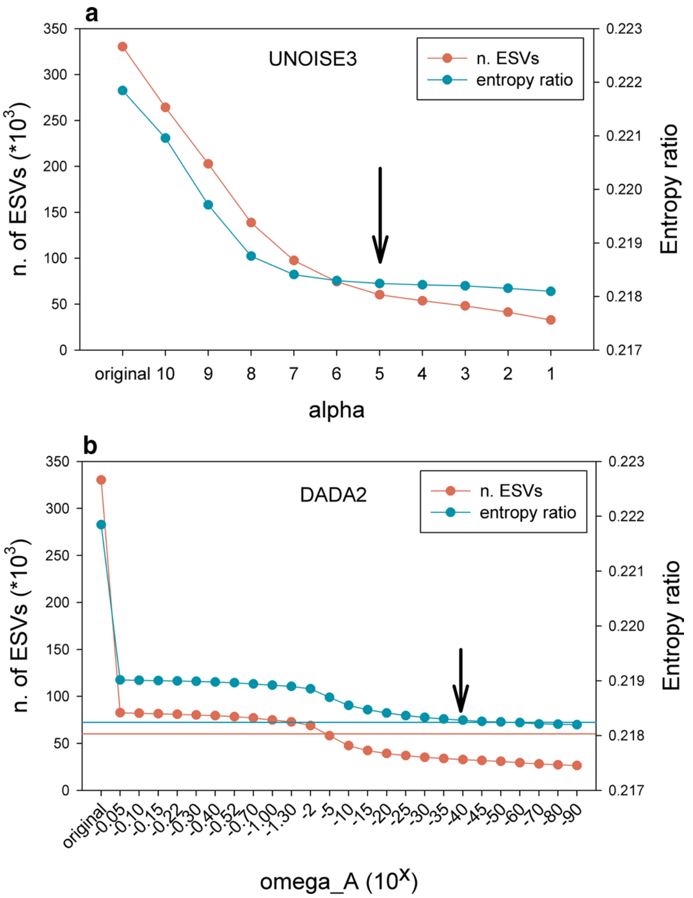

Fig. 2 Values of the Entropy ratio (Er) of the set of ESVs obtained with the UNOISE3 algorithm at decreasing values of α (**a**), and of those obtained with the DADA2 algorithm at decreasing values of omega_A (**b**). Arrows point at the selected value for each parameter. Horizontal blue line in (**b**) represents the Er value reached in (**a**) at α = 5, horizontal red line marks the number of ESVs detected in (**a**) at α = 5

The number of ESVs retained was considerably lower than for UNOISE3. In fact, the number obtained at α = 5 by the latter (60,198 ESVs) was approximately reached at omega_A = 10^{-5} (58,191 ESVs). On the other hand, the entropy value obtained at α = 5 in UNOISE3 (0.2182) was not reached until omega_A = 10^{-60}. As a compromise, we will use in this study the default value of the dada function (10^{-40}), while acknowledging that the behaviour of DADA2 with changes in omega_A for the parameters analysed was unexpected and deserves further research.

For the clustering algorithm SWARM v.2, we monitored the outcome of changing the *d* parameter between 1 and 30. For each value, we tracked the number of clusters formed (separately for all MOTUs and for those with 2 or more sequences), as well as the mean intra-MOTU and the mean inter-MOTU genetic distances (considering only the most abundant sequence per MOTU for the latter). The goal was to find the value that maximizes the intra-MOTU variability while keeping a sharp difference between both values (equivalent to the barcode gap).

The total number of MOTUs decreased sharply from 38,560 (*d* = 1) to around 19,000 with a plateau from *d* = 9 to *d* = 13, and then decreased again (Fig. 3a). If we only

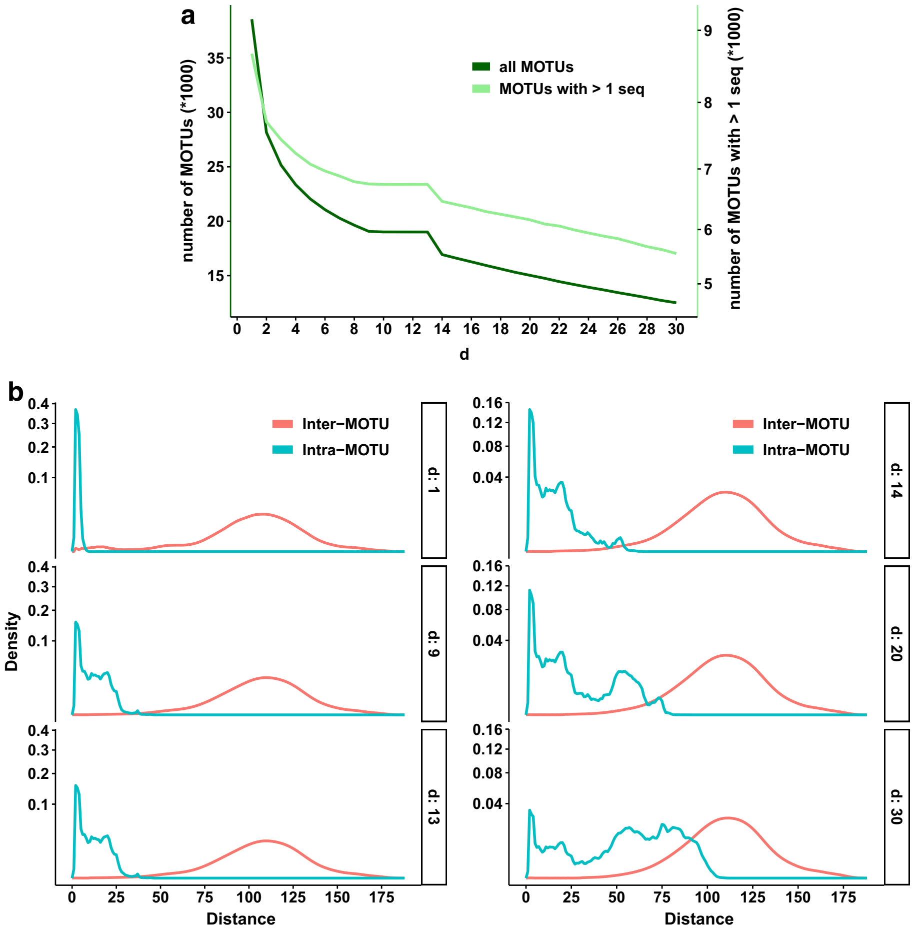

Fig. 3 **a** Number of MOTUs obtained at different values of *d* using SWARM. Total number of MOTUs (dark green) and of MOTUs with two or more sequences (light green) are represented (note different Y-axes). **b** Density plots (note quadratic scale) showing the distribution of number of differences between different clusters (inter-MOTU, red) and sequences within clusters (intra-MOTU, blue) obtained by SWARM for selected values of the parameter *d* (1, 9, 13, 14, 20 and 30)

consider the MOTUs with 2 or more sequences, the overall pattern is similar, albeit the curve is much less steep. The numbers decreased from 8684 for $d = 1$ to 6755 at $d = 12$ and 13, and decreased again at higher values (Fig. 3a).

Inter-MOTU distances had a similar distribution with all values of the parameter *d*, albeit with a small shoulder at distances of 10–20 differences with $d = 1$ (selected examples in Fig. 3b). Intra-MOTU distances, on the other hand, became more spread with higher values of *d* as expected. Values from 9 to 13 showed a similar distribution of number of differences, but for *d* values higher than 14, intra-MOTU distances started to overlap with the inter-MOTU distribution (Fig. 3b). The value of $d = 13$ seems, therefore, to be the best choice to avoid losing too much MOTU variability (both in terms of number of MOTUs and intra-MOTU variation), and at the same time keeping intra- and inter-MOTU distances well separated. The mean intra-MOTU distance in our dataset at $d = 13$ was 9.10 (equivalent to 97.09% identity), and the mean inter-MOTU distance was 108.78 (65.25% identity).

Table 1 Main characteristics of the original and the generated datasets

	n. ESVs (*)	n. MOTUs	Single-ESV MOTUs	ESVs/MOTU (*)	Reads/MOTU
Original	330,382	–	–	–	–
Du (**)	60,198	–	–	–	–
Da	32,798	–	–	–	–
Du_e (***)	113,133	–	–	–	–
S	330,382	19,012	12,257	17.378	511.194
Du_S	60,198	19,058	12,471	3.159	509.961
S_Du	75,069	19,012	12,433	3.949	511.194
Da_S	32,798	19,167	15,565	1.711	507.060
S_Da	35,376	19,012	15,198	1.861	511.194
Du_d_S	60,198	19,058	12,471	3.159	509.960
Du_c_S	60,198	19,058	12,471	3.159	509.960
Du_e_S	113,133	19,016	12,365	5.949	511.087
Du_e_d_S	113,133	19,016	12,365	5.949	511.087
Du_e_c_S	113,133	19,016	12,365	5.949	511.087

All datasets had 9,718,827 reads. 1-ESV MOTUs refer to the number of MOTUs with just one ESV. Codes of the datasets: Du, denoised with UNOISE3 algorithm (unless otherwise stated, it refers to the original formulation giving precedence to abundance ratio); Da, denoised with DADA2 algorithm; S, clustered with SWARM algorithm; Du_S, denoised (UNOISE3) and clustered; S_Du, clustered and denoised (UNOISE3); Da_S, denoised (DADA2) and clustered; S_Da, clustered and denoised (DADA2); Du_d_S, denoised (UNOISE3) with precedence to distance and clustered; Du_c_S, denoised (UNOISE3) with combined precedence and clustered; Du_e _S, denoised (UNOISE3) with correction taking into account the entropy of the codon positions and clustered; Du_e_d_S, denoised (UNOISE3) with correction plus precedence to distance and clustered; Du_e_c_S, denoised (UNOISE3) with correction plus combined precedence and clustered

*For the original and S datasets the number of sequences instead of ESVs is used

**The same values apply to Du_d (distance precedence) and Du_c (combined precedence)

***The same values apply to Du_e_d (distance precedence) and Du_e_c (combined precedence)

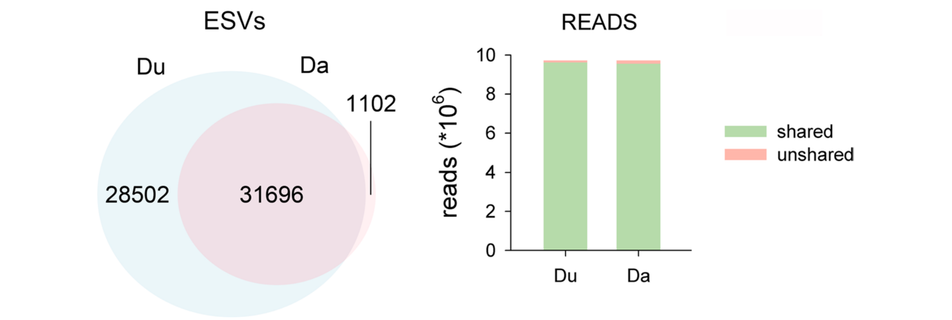

Fig. 4. Venn Diagram showing the number of ESVs shared between the two denoising procedures (Du vs Da). Bar chart shows the number of reads in the shared and unshared ESVs

The impact of the steps and their order

Table 1 shows the main characteristics of the original and the generated datasets, as well as the datasets obtained by modifying the UNOISE3 algorithm (see below). All datasets are available from Mendeley Data [55]).

We first compared the outcomes of denoising the original reads with UNOISE3 and DADA2 (Du vs Da), with the stringency parameters set as above. The error rates of the different substitution types as a function of quality scores were highly correlated in the DADA2 learnErrors procedure. The lowest Pearson correlation was obtained between the substitutions T to C and A to G ($r=0.810$), and all correlations (66 pairs of substitution types) were significant after a False Discovery Rate correction [56].

The main difference found is that the Du dataset retained almost double number of ESVs than the Da dataset: 60,198 vs 32,798. Of these, 31,696 were identical in the two datasets (Fig. 4), representing a match index of 0.746. Of the shared ESVs, 20,691 (65.28%) had exactly the same number of reads, suggesting that the same reads have been merged in these ESVs.

On the other hand, the shared ESVs concentrated most of the reads (Fig. 4): the match index for the reads was 0.986. This is coherent with the fact that most of the non-shared ESVs of the Du dataset had a low number of reads (mean = 3.66). Thus, the two denoising algorithms with the chosen parameter values provided similar results as for the abundant ESVs, but UNOISE3 retained a high number of low abundance ESVs as true sequences.

We then evaluated the output of combining denoising and clustering, using either of them as a first step. Thus, we compared the datasets Du_S, S_Du, Da_S, and S_Da. The results showed that the final number of MOTUs obtained was similar (ca. 19,000) irrespective of the denoising method and the order used (Table 1). Moreover, the shared MOTUs (flagged as MOTUs that have the same representative sequence) were the overwhelming majority (Fig. 5), with MOTU match indices over 0.96 in all comparisons.

As for the number of ESVs, clustering first results in a higher number of retained sequence variants than clustering last, ca. 25% more for Du and ca. 8% for Da. In all comparisons, the majority of ESVs were to be found in the shared MOTUs, and the same applies to the number of reads (Fig. 6, match indices for the ESVs, all > 0.95, match indices for the reads, all > 0.99). Ca. 2/3 of the MOTUs comprised a single ESV when using Du, and this number increased notably with Da (ca. 80% of MOTUs, Table 1). In both cases, clustering first resulted in a slight decrease of the number of single-ESV MOTUs.

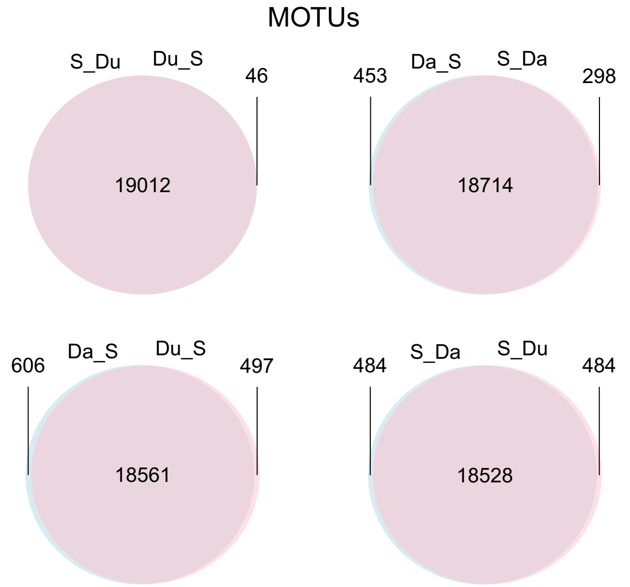

Fig. 5. Venn diagrams showing the number of MOTUs shared between the two denoising procedures and a clustering step performed in different orders

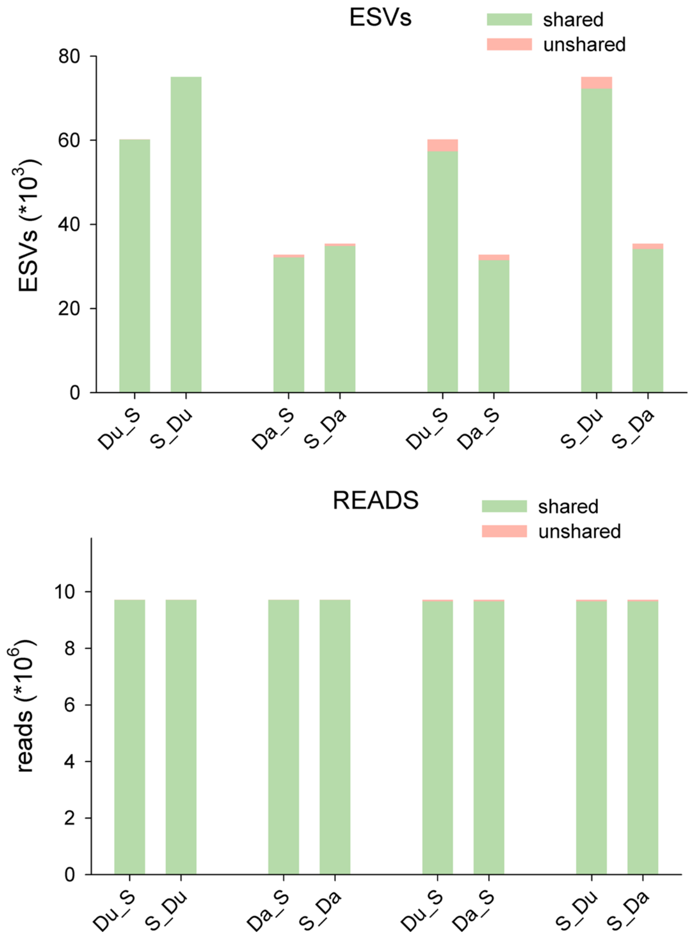

Fig. 6. Bar charts of the number of ESVs and the number of reads found in the shared and unshared MOTUs in the same comparisons as in Fig. 5

Improving the denoising algorithm

We tried different options of our DnoisE algorithm. The use of the Levenshtein distance without any correction and with priority to abundance skew corresponds to the original UNOISE3 algorithm (i.e., the Du dataset used previously). We also tried priority to distance and a combination of skew and abundance to choose among the potential "mother" ESVs to which a given "daughter" sequence will be joined. The same three options were applied when correcting distances according to the entropy of each codon position. In this case we used a pairwise distance accounting for the codon position where a substitution was found. We further applied a clustering step (SWARM) to the DnoisE results to generate MOTU sets (Du_S, Du_d_S, Du_c_S, Du_e_S, Du_e_d_S, Du_e_c_S, see Table 1 for explanation of codes) for comparison with those obtained previously.

The three ways to join sequences have necessarily the same ESVs, only the sequences that are joined under each centroid can vary and, thus, the abundance of each ESV and how these are clustered in MOTUs. However, this had a very small effect in our case. For the three datasets generated without distance correction, most MOTUs were shared, and the shared MOTUs comprised most ESVs. In turn most ESVs have the same number of reads, suggesting that the same sequences have been grouped in each ESV. All match indices were ca. 0.99. The same was found for the three entropy-corrected datasets.

On the other hand, if we consider the entropy of codon positions the results change notably in terms of ESV recovered. The corrected datasets have 113,133 ESVs (against 60,198 of the uncorrected datasets). So, when considering the entropy in distance calculations the number of retained ESVs increased by 88%. This is the result of accepting sequences that have variation in third codon positions as legitimate. When comparing the entropy-corrected and uncorrected datasets 57,318 ESVs were found in common (ESV match index of 0.729). These ESVs comprise a majority of reads, though (read match indices of ca. 0.97 in all possible comparisons). Figure 7 illustrates one of these comparisons (Du vs Du_e_c).

When clustering the ESVs obtained with the different methods, the final number of MOTUs obtained was similar to those generated in the previous sections (ca. 19,000 in all cases, Table 1). This indicates that the entropy corrected datasets provided more intra-MOTU variability, but no appreciable increase in the number of MOTUs. As an example, the mean number of ESVs per MOTU was 3.159 for the Du_S dataset, and 5.949 for the Du_e_c_S dataset. The number of single-ESV MOTUs decreased slightly (12,471 for Du_S, 12,365 for Du_e_c_S). Taking this comparison as an example, most MOTUs (as indicated by identity in the representative sequence) were shared between datasets. In addition, most of the ESVs and most of the reads were found in the shared MOTUs (match indices for MOTUs, ESVs and reads > 0.99).

Benchmarking

We computed the percent of erroneous ESVs (either because they have stop codons or changes in the five conserved aminoacids) in the MOTUs assigned to metazoans for the datasets obtained with and without entropy correction. The original dataset clustered without any denoising (dataset S) had 9,702 erroneous ESVs (or 4.65% of the total number of ESVs). The denoised dataset Du_S had 559 erroneous ESVs (1.58%), while the dataset denoised considering the variability of the codon positions (Du_e_c_S) had 500 erroneous ESVs (0.70%). Thus, albeit the uncorrected UNOISE3 procedure reduced the proportion of errors to one third, when a correction for codon position is applied the absolute number of errors is reduced, out of almost double total number of ESVs, thus the relative number is cut by more than one half.

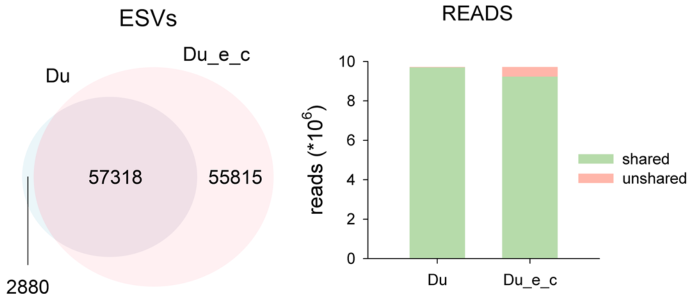

Fig. 7 Venn Diagram showing the number of ESVs shared between two denoised datasets (Du vs Du_e_c). Bar chart shows the number of reads in the shared and unshared ESVs

The results of the taxonomic benchmarking are given in detail in Additional File 3, while the obtained species-level dataset is available as Additional File 4. In short, all datasets recovered a majority of closed MOTUs, meaning that ESVs assigned to a given species were placed in the same MOTU. The proportion of hybrid MOTUs was lower for the more stringent DADA2 datasets. On the contrary, the proportion of species recovered and the proportion of ESVs with species-level assignment was lowest for the DADA2 datasets and highest for the entropy-corrected UNOISE3 datasets.

Discussion

After adjusting the different parameters of the algorithms based on ad hoc criteria for COI amplicons, between ca. 33,000 and ca. 113,000 ESVs were obtained depending on the denoising procedure used. Irrespective of the method, however, they clustered into ca. 19,000 MOTUs. This implies that there was a noticeable intra-MOTU variability even for the most stringent denoising method. The application of SWARM directly to the original dataset (without any denoising) generated likewise ca. 19,000 MOTUs. This suggests that the SWARM algorithm is robust in recovering alpha-diversity even in the presence of noisy sequences. Thus, denoising and clustering clearly accomplish different functions and, in our view, both are complementary and should be used in combination. The fact that some studies detect more MOTUs than ESVs when analysing datasets using clustering and denoising algorithms separately (e.g., [8, 57]) reflects a logical flaw: MOTUs seek to recover meaningful species-level entities, ESVs seek to recover correct sequences. There should be more sequences than species, otherwise something is wrong with the respective procedures. It has even been suggested that ESVs or MOTUs represent a first level of sequence grouping and that a second round using network analysis is convenient [9]. We contend that, with the right parameter settings, this is unnecessary for eukaryotic COI datasets.

We do not endorse the view of Callahan et al. [13] that ESVs should replace MOTUs as the standard unit analysis of amplicon-sequencing datasets. Using information at the strain level may be useful in the case of prokaryotes, and in low-variability eukaryote markers such as ribosomal 18S rDNA there may be correspondence between species and unique sequences (indeed, in many cases different species share sequences). But even in more variable nuclear markers such as ITS, a clustering step is necessary [58]. In eukaryotes the unit of diversity analyses is the species. MOTUs and not ESVs target species-level diversity and, in our view, should be used as the standard unit of analyses for most ecological and monitoring applications. Most importantly, that ESVs are organized into MOTUs is highly relevant information added at no cost. We do not agree that clustering ESVs into MOTUs eliminates biological information [29]. This only happens if only one representative sequence per MOTU is kept. We strongly advocate here for keeping track of the different sequences clustered in every MOTU and reporting them in metabarcoding studies. In this way analyses can be performed at the MOTU level or at the ESV level, depending on the question addressed.

Denoising has been suggested as a way to overcome problems of MOTU construction and to provide consistent biological entities (the correct sequences) that can be compared across studies [13]. We fully agree with the last idea: ESVs are interchangeable units that allow comparisons between datasets and can avoid generating too big datasets when combining reads of, say, temporally repeated biomonitoring studies. But clustering ESVs into MOTUs comes as a bonus, provided the grouped sequences are kept and not collapsed under a representative sequence, thus being available for future reanalyses.

The denoising and clustering methods here tested have been developed for ribosomal markers and uncritically applied to COI data in the past, with default parameter values often taken at face value (in fact, parameters are rarely mentioned in methods sections). We confirm that the UNOISE3 parameter $\alpha = 5$ is adequate for COI data, in agreement with previous research using three independent approaches [12, 39, 40]. We also tested and confirmed the suitability of a d value of 13 for SWARM that has been used in previous works with COI datasets (e.g., [43–47]). As Mahé et al. [42] noted, higher d values can be necessary for fast evolving markers. They advised to track MOTU coalescing events as d increases to find the value best-fitting the sequence marker chosen. We have followed this approach, together with the course of the intra- and inter-MOTU distances, to select the d-value for the COI marker. In our view, fixed-threshold clustering procedures should be avoided, as even for a given marker the intra- and interspecies distances can vary according to the group of organisms considered. With SWARM, even if the initial clusters were made at $d = 13$ (for a fragment of 313 this means an initial threshold of 4.15% for connecting sequences), after the refining procedure the mean intra-MOTU distances obtained was 2.91%, which is in line with values suggested using the whole barcoding region of COI [59]. Furthermore, in our taxonomic benchmarking, we found a high proportion of closed MOTUs, irrespective of the denoising method used, indicating that the SWARM procedure adequately and robustly grouped the sequences with known species-level assignments.

Our preferred algorithm for denoising is UNOISE3. It is a one-pass algorithm based on a simple formula with few parameters, it is computationally fast and can be applied at different steps of the pipelines. It keeps almost double ESVs than DADA2 and, combined with a clustering step, results in less single-sequence MOTUs and a higher number of ESVs per MOTU, thus capturing a higher intra-MOTU diversity. It also produced 60% more closed group MOTUs than DADA2 in our taxonomic benchmarking. Edgar et al. [30], by comparing both algorithms in mock and in vivo datasets, also found that UNOISE had comparable or better accuracy than DADA2. Similarly, Tsuji et al. [41] found that UNOISE3 retained less false haplotypes than DADA2 in samples from tank water containing fish DNA. We also found that the entropy values of the sequences changed as expected when denoising becomes more stringent with UNOISE3, indicating that the algorithm performs well with coding sequences. We also suggest ways of improving this algorithm (see below).

DADA2, on the other hand, is being increasingly used in metabarcoding studies but its suitability for a coding gene such as COI remains to be demonstrated. We had to use paired reads (against recommendation) to be able to make meaningful comparisons, but our results indicate that with unpaired sequences the number of ESVs retained would have been even lower. The DADA2 algorithm, when tested with increasingly stringent parameters, did not progressively reduce the entropy ratio values that should reflect an adequate denoising of coding sequences. Further, the high correlation of error rates between all possible substitution types suggests that the algorithm may be over-parameterized, at least for COI, which comes at a computational cost. Comparisons based on known communities (as in [41]) and using COI are needed to definitely settle the appropriateness of the two algorithms for metabarcoding with this marker.

In addition, PCR-free methods now popular in library preparation procedures complicate the use of DADA2 as there is no consistent direction (forward or reverse) of the reads. We acknowledge that our paired sequences still included a mixture of reads that were originally in one or another direction and, thus, with different error rates. However, the non-overlapped part is only the initial ca. 100 bp, and these are in general good quality positions in both the forward and reverse reads.

Another choice to make is to decide what should come first, denoising or clustering. Both options have been adopted in previous studies (note that clustering first is not possible with DADA2 unless paired sequences are used). Turon et al. [12] advocated that denoising should be made within MOTUs, as they provide the natural "sequence environment" where errors occur and where they should be targeted by the cleaning procedure. We found that clustering first retained more ESVs, because sequences that would otherwise be merged with another from outside its MOTU were preserved. It also resulted in less single-ESV MOTUs, retaining more intra-MOTU variability. It can also be mentioned that denoising the original sequences took approximately 10 times more computing time than denoising within clusters, which can be an issue depending on the dataset and the available computer facilities. We acknowledge, however, that most MOTUs are shared and most ESVs and reads are in the shared MOTUs when comparing the two possible orderings, irrespective of denoising algorithm. The final decision may come more from the nature and goals of each study. For instance, a punctual research may go for clustering first and denoising within clusters to maximize the intra-MOTU variability obtained. A long-term research that implies multiple samplings over time that need to be combined together may use denoising first and then perform the clustering procedure at each reporting period with the ESVs obtained in the datasets collected so far pooled.

There are other important steps at which errors can be reduced and that require key choices, but they are outside the scope of this work as we addressed only clustering and denoising steps. In particular, nuclear insertions (numts) may be difficult to distinguish from true mitochondrial sequences [50, 60]. Singletons (sequences with only one read) are also a problem for all denoising algorithms (as it is difficult to discern rare sequences from errors). Singletons are often eliminated right at the initial steps, as we did in this work. Likewise, a filtering step, in which ESVs with less than a certain amount of reads are eliminated, is deemed necessary to obtain biologically reliable datasets. A 5% relative abundance cut-off value was suggested by [39], while [12] proposed an absolute threshold of 20 reads. However, the procedure and the adequate threshold are best adjusted according to the marker and the study system, so, albeit we acknowledge that a filtering step is necessary, this has not been addressed in this paper.

We recommend that the different denoising algorithms be programmed as standalone steps (not combined, for instance, with chimera filtering) so anyone interested could combine the denoising step with the preferred choices for other steps. We also favour open source programs that could be customized if needed. For UNOISE3 algorithm we suggest that a combination between distance and skew ratio be considered to assign a read to the most likely centroid. This had little effect in our case, but can be significant in other datasets. For DADA2 algorithm, we advise to weight the gain of considering the two reads separately vs using paired sequences. The advantages of the latter involve a higher flexibility of the algorithm as it does not need to be performed right at the beginning of the pipeline. For both algorithms, we think it is important to consider the natural variation of the three positions of the codons of a coding sequence such as COI, which can allow a more meaningful computation of distances between sequences and error rates. This of course applies to other denoising algorithms not tested in the present study (e.g., AmpliCI [27], deblur [61]). Our DnoiSE program, based on the UNOISE3 algorithm, includes the option of incorporating codon information in the denoising procedure. With this option, we found ca 50,000 more ESVs than with the standard approach. Importantly, this fact did not increase the proportion of erroneous sequences, as determined using aminoacid substitution patterns in metazoan MOTUs. Rather, this proportion was cut by one-half, and erroneous sequences were less even in absolute numbers. In our taxonomic benchmarking, a higher proportion of ESVs with species-level matches in the reference database were detected with the codon position-corrected method. We used a dataset of fixed sequence length and eliminated misaligned sequences. The correction for codon position would be more complicated in the presence of indels and dubious alignments. We also acknowledge the lack of a mock community to ground truth our method, but we contend that mock communities are hardly representative of highly complex communities such as those here analysed. We hope our approach will be explored further and adequately benchmarked in future studies on different communities.

Conclusions

COI has a naturally high intraspecies variability that should be assessed and reported in metabarcoding studies, as it is a source of highly valuable information. Denoising and clustering of sequences are not alternatives. Rather, they are complementary and both should be used together to make the most of the inter- and intraspecies information contained in COI metabarcoding datasets. We emphasize the need to carefully choose the stringency parameters of the different steps according to the variability of this marker.

Our results indicated that the UNOISE3 algorithm preserved a higher intra-cluster variability than DADA2. We introduce the program DnoisE to implement the UNOISE3 algorithm considering the natural variability (measured as entropy) of each codon position in protein-coding genes. This correction increased the number of sequences retained by 88%. The order of the steps (denoising and clustering) had little influence on the final outcome.

We provide recommendations for the preferred algorithms of denoising and clustering, as well as step order, but these may be tuned according to the goals of each study, feasibility of preliminary tests, and ground-truthing options, if any. Other important steps of metabarcoding pipelines, such as abundance filtering, have not been addressed in this study and should be adjusted according to the marker and the study system.

We advise to report the results in terms of both MOTUs and ESVs included in each MOTU, rather than reporting only MOTU tables with collapsed information and just a representative sequence. We also advise that the coding properties of COI should be used both to set the right parameters of the programs and to guide error estimation in denoising procedures. We wanted to spark further studies on the topic, and our procedures should be tested and validated or refined in different types of community.

There is a huge amount of intra- and inter-MOTU information in metabarcoding datasets that can be exploited for basic (e.g., biodiversity assessment, connectivity estimates, metaphylogeography) and applied (e.g., management) issues in biomonitoring programs, provided the results are reported adequately.

Abbreviations

ASV: amplicon sequence variant; COI: cytochrome *c* oxidase subunit 1; ESV: exact sequence variant; MOTU: molecular operational taxonomic unit; OTU: operational taxonomic unit; ZOTU: zero-radius operational taxonomic unit.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04115-6.

Additional file 1 Format: .pdf. Details of the dataset used in the analysis of the article, including information of the sampling localities in the Iberian Peninsula and the sample processing steps prior to sequencing.

Additional file 2 Format: .pdf. Comparison of DADA2 on paired and unpaired reads.

Additional file 3 Format: .pdf. Details of the taxonomic benchmarking.

Additional file 4 Format: .csv. Species-level dataset. Table with the ESVs identified at the species level with >97% similarity. The taxonomy assigned is indicated, as well as the best-match in the reference database, the taxid, and the sequence.

Acknowledgements

We thank Daniel Sanromán for his help with the field sampling.

Authors' contributions

XT and OSW conceived the study; XT, AA and CP performed the sampling; AA performed the laboratory work; AA and OSW wrote code; XT, CP, AA and OSW performed analysis; XT and AA drafted the ms; all authors contributed intellectual content and revised the ms.

Funding

This study has been funded by Project PopCOmics (CTM2017-88080, MCIU/AEI/FEDER/UE) and project BigPark (Autonomous Organism of National Parks, OAPN 2462/2017) from the Spanish Government.

Availability of data and materials

All the datasets supporting the conclusions of this article are available in the Mendeley Data repository in https://data.mendeley.com/datasets/84zypvmn2b/ and in Additional File 4. The Phyton program DnoisE is available in the Github repository in https://github.com/adriantich/DnoisE.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB-CSIC), Blanes (Girona), Catalonia, Spain. [2] Department of Evolutionary Biology, Ecology and Environmental Sciences, University of Barcelona and Research Institute of Biodiversity (IRBIO), Barcelona, Catalonia, Spain. [3] Norwegian College of Fishery Science, UiT The Arctic University of Norway, Tromsö, Norway.

Received: 9 January 2021 Accepted: 30 March 2021
Published online: 05 April 2021

References

1. Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Mol Ecol. 2017;26:5872–95.
2. Aylagas E, Borja A, Muxika I, Rodríguez-Ezpeleta N. Adapting metabarcoding-based benthic biomonitoring into routine ecological status assessment networks. Ecol Ind. 2018;95:194–202.
3. Bani A, De Brauwer M, Creer S, Dumbrell AJ, Limmon G, Jompa J, von der Heyden S, Beger M. Informing marine spatial planning decisions with environmental DNA. Adv Ecol Res. 2020;62:375–407.
4. Compson ZG, McClenaghan B, Singer GAC, Fahner N, Hajibabaei M. Metabarcoding from microbes to mammals: comprehensive bioassessmenton a global scale. Front Ecol Evol. 2020;8:581835.
5. Mathieu C, Hermans SM, Lear G, Buckley TR, Lee KC, Buckley HL. A systematic review of sources of variability and uncertainty in eDNA data for environmental monitoring. Front Ecol Evol. 2020;8:135.
6. Rodríguez-Ezpeleta N, Morisette O, Bean CW, Manu S, Banerjee P, Lacoursière-Roussel A, Beng KC, Alter SE, Roger F, Holman LE, Stewart KA, Monaghan MT, Mauvisseau Q, Mirimin L, Wangensteen OS, Antognazza CM, Helyar SJ, de Boer H, Monchamp ME, Nijland R, Abbott CL, Doi H, Barnes MA, Leray M, Hablützel PI, Deiner K. Trade-offs between reducing complex terminology and producing accurate interpretations from environmental DNA: comment on 'Environmental DNA: What's behind the term?' by Pawlowski et al. (2020). EcoEvoRxiv. 2020. https://doi.org/10.32942/OSF.IO/KGNYD.
7. Porter TM, Hajibabaei M. Putting COI metabarcoding in context: the utility of exact sequence variants (ESV) in biodiversity analysis. Front Ecol Evol. 2020;8:248.
8. Macheriotou L, Guilini K, Bezerra TN, Tytgat B, Nguyen DT, Nguyen TXP, Noppe F, Armenteros M, Boufahja F, Rigaux A, Vanreusel A, Derycke S. Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. Ecol Evol. 2019;9:1211–26.
9. Forster D, Lentendu G, Filker S, Dubois E, Wilding TA, Stoeck T. Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. Environ Microbiol. 2019;21(11):4109–24.
10. O'Rourke DR, Bokulich NA, Jusino MA, MacManes MD, Foster JT. A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. Ecol Evol. 2020;10:9721–9.
11. Giebner H, Langen K, Bourlat SJ, Kukowka S, Mayer C, Astrin JJ, Misof B, Fonseca VG. Comparing diversity levels in environmental samples: DNA sequence capture and metabarcoding approaches using 18S and COI genes. Mol Ecol Resour. 2020;20:1333–45.
12. Turon X, Antich A, Palacín C, Praebel K, Wangensteen OS. From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecol Appl. 2020;30:e02036.
13. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11:2639–43.
14. Andujar C, Arribas P, Yu DW, Vogler AP, Emerson BC. Why the COI barcode should be the community DNA metabarcode for the Metazoa. Mol Ecol. 2018;27:3968–75.
15. van der Loos LM, Nijland R. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Mol Ecol. 2020. https://doi.org/10.1111/MEC.15592.
16. Tapolczai K, Keck F, Bouchez A, Rimet F, Kahlert M, Vasselon V. Diatom DNA metabarcoding for biomonitoring: strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. Front Ecol Evol. 2019;7:409.
17. Holman LE, de Bruyn M, Creer S, Carvalho G, Robidart J, Rius M. Consistent marine biogeographic boundaries across the tree of life despite centuries of human impacts. bioRxiv. 2020. https://doi.org/10.1101/2020.06.24.169110.
18. Steyaert M, Priestley V, Osborne O, Herraiz A, Arnold R, Savolainen O. Advances in metabarcoding techniques bring us closer to reliable monitoring of the marine benthos. J Appl Ecol. 2020;57:2234–45.
19. Zamora-Terol S, Novotny A, Winder M. Reconstructing marine plankton food web interactions using DNA metabarcoding. Mol Ecol. 2020;29:3380–95.
20. Pearman JK, Chust G, Aylagas E, Villarino E, Watson JR, Chenuil A, Borja A, Cahill AE, Carugati L, Danovaro R, David R, Irigoien X, Mendibil I, Moncheva S, Rodríguez-Ezpeleta N, Uyarra MC, Carvalho S. Pan-regional marine benthic cryptobiome biodiversity patterns revealed by metabarcoding autonomous reef monitoring structures. Mol Ecol. 2020;29:4882–97.
21. Brandt MI, Trouche B, Quintric L, Wincker P, Poulain J, Arnaud-Haond S. A flexible pipeline combining bioinformatic correction tools for prokaryotic and eukaryotic metabarcoding. bioRxiv. 2020. https://doi.org/10.1101/717355.
22. Nguyen BN, Shen EW, Seemann J, Correa AMS, O'Donnell JL, Altieri AH, Knowlton N, Crandall KA, Egan SP, McMillan WO, Leray M. Environmental DNA survey captures patterns of fish and invertebrate diversity across a tropical seascape. Sci Rep. 2020;10:6729.
23. Laroche O, Kersten O, Smith CR, Goetze E. Environmental DNA surveys detect distinct metazoan communities across abyssal plains and seamounts in the western Clarion Clipperton Zone. Mol Ecol. 2020;29:4588–604.
24. Zizka VMA, Weiss M, Leese F. Can metabarcoding resolve intraspecific genetic diversity changes to environmental stressors? A test case using river macrozoobenthos. Metabarcoding Metagenom. 2020;4:23–34.
25. Avise JC. Phylogeography: retrospect and prospect. J Biogeogr. 2009;36:3–15.
26. Emerson BC, Cicconardi F, Fanciulli PP, Shaw PJA. Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. Philos Trans R Soc B. 2011;366:2391–402.
27. Peng X, Dorman K. AmpliCI: A high-resolution model-based approach for denoising Illumina Amplicon data. Bioinformatics. 2020. https://doi.org/10.1093/bioinformatics/btaa648.
28. Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahé F, He Y, Zhou HW, Rognes T, Caporaso JG, Knight R. Open-source sequence clustering methods improve the state of the art. mSystems. 2020;1(1):e00003–15.
29. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.
30. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS aplicon sequencing. bioRxiv. 2016. https://doi.org/10.1101/081257.

31. Edgar RC. UPARSE: hihgly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013;10(10):996–1000.
32. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.
33. Hao X, Jiang R, Chen T. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics. 2011;27(5):611–8.
34. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ. 2015;3:e1420.
35. Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E. OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. Mol Ecol Resour. 2016;16:176–82.
36. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;2010(26):2460–1.
37. Antich A. DnoisE, Distance denoise by Entropy. GitHub repository. https://github.com/adriantich/DnoisE. Accessed 20 November 2020.
38. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–7.
39. Elbrecht V, Vamos EE, Steinke D, Leese F. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ. 2018;6:e4644.
40. Shum P, Palumbi SR. Testing small-scale ecological gradients and intraspecific differentiation from hundreds of kelp forest species using haplotypes from metabarcoding. Mol Ecol. 2021. https://doi.org/10.1111/mec.15851.
41. Tsuji S, Miya M, Ushio M, Sato H, Minamoto T, Yamanaka H. Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: a case study using tank water. Environ DNA. 2020;2:42–52.
42. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ. 2014;2:e593.
43. Siegenthaler A, Wangensteen OS, Soto AZ, Benvenuto C, Corrigan L, Mariani S. Metabarcoding of shrimp stomach content: Harnessing a natural sampler for fish biodiversity monitoring. Mol Ecol Resour. 2019;19:206–20.
44. Garcés-Pastor S, Wangensteen OS, Pérez-Haase A, Pèlachs A, Pérez-Obiol R, Cañellas-Boltà N, Mariani S, Vegas-Vilarrúbia T. DNA metabarcoding reveals modern and past eukaryotic communities in a high-mountain peat bog system. J Paleolimnol. 2019;62:425–41.
45. Bakker J, Wangensteen OS, Baillie C, Buddo D, Chapman DD, Gallagher AJ, Guttridge TL, Hertler H, Mariani S. Biodiversity assessment of tropical shelf eukaryotic communities via pelagic eDNA metabarcoding. Ecol Evol. 2019;9:14341–55.
46. Atienza S, Guardiola M, Praebel K, Antich A, Turon X, Wangensteen OS. DNA metabarcoding of deep-sea sediment communities using COI: community assessment, spatio-temporal patterns and comparison with 18S rDNA. Diversity. 2020;12:123.
47. Antich A, Palacin C, Cebrian E, Golo R, Wangensteen OS, Turon X. Marine biomonitoring with eDNA: Can metabarcoding of water samples cut it as a tool for surveying benthic communities? Mol Ecol. 2021. https://doi.org/10.1111/mec.15641.
48. Schmidt AO, Herzel H. Estimating the entropy of DNA sequences. J Theor Biol. 1997;3:369–77.
49. Hausser J, Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. J Mach Learn Res. 2009;10:1469–84.
50. Andújar C, Creedy TJ, Arribas P, López H, Salces-Castellano A, Pérez-Delgado AJ, Vogler AP, Emerson BC. Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. Mol Ecol Resour. 2021. https://doi.org/10.1111/1755-0998.13337.
51. Wangensteen OS, Palacin C, Guardiola M, Turon X. DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. Peer J. 2013;6:e4705.
52. Wangensteen OS. Reference-databases Metabarpark. GitHub repository. http://github.com/metabarpark/Reference-databases. Accessed 23 December 2020.
53. Pentinsaari M, Salmela H, Mutanen M, Roslin T. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. Sci Rep. 2016;6:35275.
54. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.58.0. https://bioconductor.org/packages/Biostrings. Accessed 10 March 2021.
55. Antich A, Palacin C, Wangensteen OS, Turon X. Dataset for "To denoise or to cluster? That is not the question. Optimizing pipelines for COI metabarcoding and metaphylogeography". Mendeley Data. 2021. https://data.mendeley.com/datasets/84zypvmn2b/.
56. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;55(1):289–300.
57. Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ. 2018;6:e5364.
58. Estensmo EL, Maurice S, Morgado L, Martin-Sanchez P, Skrede I, Kauserud H. The influence of intraspecific sequence variation during DNA metabarcoding: a case study of eleven fungal species. Authorea. 2020. https://doi.org/10.22541/au.160071155.58915559.

59. Ratnasingham S, Hebert PDN. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PLoS ONE. 2013;8(8):6.
60. Porter TM, Hajibabaei M. Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. bioRxiv. 2021. https://doi.org/10.1101/2021.01.24.427982.
61. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017;2(2):e00191-16.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.