Optimal sequence similarity thresholds for clustering of molecular operational taxonomic units in DNA metabarcoding studies

Aurélie Bonin[1,2]*, Alessia Guerrieri[1], G. Francesco Ficetola[1,3]

1) Department of Environmental Science and Policy, University of Milan. Via Celoria 10, 20126 Milano Italy
2) Argaly, Bâtiment CleanSpace, 354 Voie Magellan, 73800 Sainte-Hélène-du-Lac, France
3) Univ. Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, LECA, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France

*Corresponding author: aurelie.bonin@argaly.com

Please cite this paper as:

Bonin, A., Guerrieri, A., Ficetola, G.F., 2023. Optimal sequence similarity thresholds for clustering of molecular operational taxonomic units in DNA metabarcoding studies. Molecular Ecology Resources 23, 368-381.

Abstract

Clustering approaches are pivotal to handle the many sequence variants obtained in DNA metabarcoding datasets, therefore they have become a key step of metabarcoding analysis pipelines. Clustering often relies on a sequence similarity threshold to gather sequences in Molecular Operational Taxonomic Units (MOTUs), each of which ideally representing a homogeneous taxonomic entity, e.g. a species or a genus. However, the choice of the clustering threshold is rarely justified, and its impact on MOTU over-splitting or over-merging even less tested. Here, we evaluated clustering threshold values for several metabarcoding markers under different criteria: limitation of MOTU over-merging, limitation of MOTU over-splitting, and trade-off between over-merging and over-splitting. We extracted sequences from a public database for nine markers, ranging from generalist markers targeting Bacteria or Eukaryota, to more specific markers targeting a class or a subclass (e.g. Insecta, Oligochaeta). Based on the distributions of pairwise sequence similarities within species and within genera, and on the rates of over-splitting and over-merging across different clustering thresholds, we were able to propose threshold values minimizing the risk of over-splitting, that of over-merging, or offering a trade-off between the two risks. For generalist markers, high similarity thresholds (0.96-0.99) are generally appropriate, while more specific markers require lower values (0.85-0.96). These results do not support the use of a fixed clustering threshold. Instead, we advocate a careful examination of the most appropriate threshold based on the research objectives, the potential costs of over-splitting and over-merging, and the features of the studied markers.

Keywords

metabarcoding marker; sequence variant; COI; MOTU over-splitting, MOTU over-merging; alpha diversity

Introduction

DNA metabarcoding studies are typically based on a succession of experimental steps governed by important methodological choices (Zinger et al., 2019). These include a) the definition of sampling design and the selection of sampling sites (Dickie et al., 2018), b) the approach used for the preservation of the starting material (Tatangelo et al., 2014, Guerrieri et al., 2021), c) the protocol used for DNA extraction (Taberlet et al., 2012, Eichmiller et al., 2016, Zinger et al., 2016, Lear et al., 2018, Capo et al., 2021), d) the selection of appropriate primers to amplify a taxonomically-informative genomic region (Elbrecht et al., 2016, Fahner et al., 2016, Ficetola et al., 2021), e) the strategy adopted for DNA amplification and high-throughput sequencing of amplicons (Nichols et al., 2018, Taberlet et al., 2018, Bohmann et al., 2022), f) the pipeline selected for bioinformatics analyses (Boyer et al., 2016, Calderón-Sanou et al., 2020, Capo et al., 2021, Couton et al., 2021, Macher et al., 2021, Mächler et al., 2021), and g) the statistical approach used to translate metabarcoding data into ecological information (Paliy & Shankar 2016, Chen & Ficetola 2020). Each of these methodological choices can heavily influence the reliability and interpretation of results (Alberdi et al., 2018, Zinger et al., 2019), and there is thus a critical need for development, proper assessment and optimization of methods specially dedicated to DNA metabarcoding.

When analyzing metabarcoding data, bioinformatic pipelines generally produce a list of detected sequences that can be assigned to a given taxon with a more or less precise taxonomic resolution. However, the number of unique sequences obtained after bioinformatic treatment is generally much higher than the number of taxa actually present in the sample (Calderón-Sanou et al., 2020, Mächler et al., 2021). This stems from multiple reasons including genuine intraspecific diversity of the selected markers and errors occurring during the amplification or sequencing steps. Consequently, sequence clustering approaches are often used to collapse very similar sequences into one single Molecular Operational Taxonomic Unit (MOTU), which does not necessarily correspond to a species in the traditional sense (Kopylova et al., 2016, Froslev et al., 2017, Bhat et al., 2019, Antich et al., 2021). Sequence clustering can be performed using similarity thresholds, Bayesian approaches, or through single-linkage (Antich et al., 2021). Approaches based on similarity thresholds can have excellent performance and they display several advantages such as flexibility and easy implementation (Kopylova et al., 2016, Wei et al., 2021). However, when performing clustering based on sequence similarity, two key parameters have to be determined *a priori*. The first one is the sequence to be selected as representative of the cluster. In the case of metabarcoding studies, keeping the most abundant sequence of the cluster as the cluster representative is a convenient way of merging sequence variants generated during the PCR or sequencing steps with the original sequence they derive from (Mercier et al., 2013). The second parameter is the similarity threshold (clustering threshold) used to build MOTUs (Clare et al., 2016, Calderón-Sanou et al., 2020, Wei et al., 2021). The choice of this threshold is delicate without prior knowledge of the maker and its intrinsic level of diversity. A too low threshold can collapse different taxa into the same MOTU (over-merging), while a too high threshold can create too many MOTUs (over-splitting) compared to the actual diversity level (Clare et al., 2016, Roy et al., 2019, Schloss 2021).

Some works suggest that the ecological interpretation of metabarcoding data can be relatively robust to the threshold selected for sequence clustering. For instance, Botnen et al. (2018) used thresholds of sequence similarity ranging from 0.87 to 0.99 to analyze multiple microbial communities, and obtained community structures highly coherent across thresholds. Nevertheless, levels of alpha diversity can be heavily impacted by the threshold selection.

Ideally, the threshold used for clustering would depend on a trade-off between MOTU over-splitting and MOTU over-merging. A growing number of markers are currently being used in metabarcoding studies (Taberlet et al., 2018), with some allowing broad-scale biodiversity assessment but having limited taxonomic resolution (e.g. 18S rDNA primers amplifying all eukaryotes; Guardiola et al., 2015) and others being highly specific to one single class or even family (e.g. Baamrane et al., 2012, Ficetola et al., 2021). Biodiversity surveys generally aim to generate a set of MOTUs that are each associated with a unique taxon, all taxa being ideally situated at the same level in the taxonomic tree, in order to facilitate comparisons. In these conditions, optimal clustering thresholds probably differ strongly across markers. One can for example expect high values for highly conserved markers, and lower values for markers showing high variability (Kunin et al., 2010, Brown et al., 2015). However, there is limited quantitative assessment of how optimal clustering thresholds vary across markers (but see Alberdi et al., 2018).

In this study, we analyzed sequences from a public database (EMBL - European Molecular Biology Laboratory) to identify clustering thresholds for different markers and under different criteria. We considered nine metabarcoding markers (Table 1), ranging from generalist markers (i.e. targeting Bacteria or Eukaryota) to more specific markers (e.g. targeting Oligochaeta [earthworms], Insecta [insects] or Collembola [springtails]), and amplifying fragments situated either in protein coding (e.g. cytochrome c oxidase subunit 1 mitochondrial gene) or non-protein coding (e.g. rDNA genes) genomic regions. We evaluated how clustering thresholds can change for each marker and taxonomic group, depending on the criterion adopted to set the threshold. We used two alternative strategies to identify thresholds, each time with different objectives in mind. First, following a procedure similar to the one adopted in barcoding studies (Machida et al., 2009; Meyer & Paulay 2005), we compared the distribution probabilities of sequence similarities among different individuals of the same species and among different species of the same genus to identify values: *i*) minimizing the risk that different sequences of the same species are split in different MOTUs (i.e. risk of over-splitting); *ii*) minimizing the risk that distinct but related species are clustered in the same MOTU (i.e. risk of over-merging); *iii*) balancing the risk of over-splitting and over-merging (Figure 1A). Second, we calculated the over-splitting and over-merging rates of the studied markers for a range of clustering thresholds, to identify values that minimize the two error rates (Figure 1B). We expect that, if researchers want to minimize over-splitting, they should select lower clustering thresholds than if they want to minimize over-merging. Furthermore, we expect higher clustering thresholds for generalist markers compared to markers targeting one class or more restricted taxonomic groups, because of the lower taxonomic resolution and slower evolutionary rate of the former.

Methods

Markers examined and construction of sequence datasets

We focused on a set of nine DNA metabarcoding markers (Bact02, Euka02, Fung02, Sper01, Arth02, COI-BF1/BR2, Coll01, Inse01, Olig01) targeting different taxonomic groups and different genomic regions (Table 1). Four of these markers can be considered as generalist, i.e. targeting entire superkingdoms or kingdoms: Bact02 targeting Bacteria, Euka02 targeting Eukaryota, Fung02 targeting Fungi, and Sper01 targeting Spermatophyta (vascular plants). Two markers were intermediate (Arth02 and COI-BF1/BR2, both targeting arthropods, i.e. the most species-rich phylum on Earth). Finally, three markers were more specific, i.e. targeting groups from classes to subclasses: Coll01 targeting Collembola (springtails), Inse01 targeting Insecta, and Olig01 targeting Oligochaeta (earthworms). Eight of these markers are situated in non-protein coding genes (Bact02, Arth02, Coll01, Inse01 and Olig01: 16 rDNA gene; Euka02: 18S rDNA gene; Fung02: ITS1 nuclear rDNA gene; Sper01: P6 loop of the intron of the chloroplastic *trnL* gene). The last marker, COI-BF1/BR2, is situated in the cytochrome c oxidase subunit 1 (COI) mitochondrial gene (Table 1).

For each of these markers, a sequence database was built from EMBL (European Molecular Biology Laboratory) release 140 (final sequence databases available at https://doi.org/10.5061/dryad.crjdfn353) as follows. An *in silico* PCR was first carried out by running the program *ecoPCR* (Ficetola et al., 2010) using the corresponding primers (Table S1). Three mismatches per primer were allowed (-e option), and amplicon length (without primers) was restricted (-l and -L options) to the expected length interval (Table S1). The amplified sequences were further filtered by keeping only those belonging to the target taxonomic group, showing a taxonomic assignment (i.e. taxid) at the species and genus levels and having no ambiguous nucleotides. This allowed assembling a working dataset, from which we extracted two sub-datasets. The “within-species” dataset was built by keeping only species for which at least two sequences (identical or not) were available; if >2 sequences were available for a given species, we randomly selected two sequences for that species using the *obiselect* command of the OBITools. The “within-genus” dataset was built by keeping only genera for which at least two sequences were available; if >2 sequences were available for a given genus, we randomly selected two sequences for that genus using the *obiselect* command. For some markers (Bact02, Euka02, Fung02, Inse01, Sper01), the within-species dataset and sometimes the within-genus dataset still contained a very large number of sequences (>10,000). To limit computation time for these markers, we randomly selected a subset of 5000 different taxa, to reach a final number of sequences equal to 10,000. An example of dataset preparation is provided in Script1_Arth02_DatasetsPreparation.sh (Supplementary Material), and Table S2 summarizes the number of sequences in the different datasets.

Calculation of sequence similarities and probability distributions

As a measure of sequence similarity, we computed the pairwise LCS (Longest Common Subsequence) scores between pairs of sequences in the within-species and within-genus datasets using the *sumatra* program (Mercier et al., 2013; see Script2A_Arth02_PairwiseSimilarities_Sumatra.sh from the Supplementary Material). Methodological comparisons showed that this algorithm provides an excellent balance between performance and computation efficiency (Jackson et al., 2016, Kopylova et al., 2016, Bhat et al., 2019). As *sumatra* provides pairwise scores for all possible pairs of sequences, the similarity scores resulting from the within-species dataset were filtered in R (R Core Team 2020) to keep only those representing similarities between sequences of the same species. Similarly, the scores resulting from the within-genus dataset were filtered to keep only those representing similarities between different species of the same genus (see first part of Script2B_Arth02_DensityPlots.Rmd from the Supplementary Material).

Approach to identify clustering thresholds on the basis of within-species and within-genus sequence similarities

We first examined within-species and within-genus sequence similarities to evaluate four different strategies (Figure 1A) and determine the similarity value that: *i*) avoids over-splitting; *ii*) avoids over-merging; *iii*) finds a balance between over-splitting and over-merging, with two distinct procedures based on the intersection (*iii*-a) or on modes (*iii*-b) of the density probability distributions (see Script2B_Arth02_DensityPlots.Rmd from the Supplementary Material). These strategies are analogous to those adopted in traditional barcoding studies to set the limit between intra-specific and inter-specific diversity (Meyer & Paulay 2005).

i) **Avoid over-splitting**

In this case, the aim is to avoid distributing different sequences belonging to the same species in different clusters, i.e. to limit the probability of generating additional spurious MOTUs. For this purpose, we selected as clustering threshold the 10% quantile of the distribution of similarities between sequences from the same species (within-species dataset). With this approach, the sequences belonging to the same species according to EMBL are gathered in the same cluster in 90% of the cases.

ii) **Avoid over-merging**

In this case, the aim is to avoid gathering sequences attributed to different species of the same genus in the same cluster, i.e. to limit the probability of merging related species in the same MOTU. For this purpose, we selected as clustering threshold the 90% quantile of the distribution of similarities between different species belonging to the same genus. With this approach, the sequences attributed to different species belonging to the same genus are assigned to different clusters in 90% of the cases.

iii) **Find a balance between over-splitting and over-merging**

In this case, the aim was to minimize both over-splitting and over-merging. We considered two distinct approaches. First, we obtained the probability distribution of within-species and within-genus sequence pairwise similarities using the *density* function from R, with biased cross-validation (bw=“bcv”) as smoothing bandwidth selector and a Gaussian smoothing kernel (kernel=“gaussian”; Venables & Ripley 2002). We tested other possible smoothing bandwidth selectors, but biased cross-validation was the approach best fitting the score histograms for all markers and all datasets (Figures S1 to S9). The balance threshold *iii*-a was then identified as the intersection between the probability distributions of the within-species and within-genus similarities. As an alternative approach to balance over-merging and over-splitting (*iii*-b), we calculated the midpoint between the modes of the within-species and within-genus probability distributions.

Rates of over-merging and over-splitting

For each marker, over-merging and over-splitting rates were evaluated at different clustering thresholds using the within-species dataset described in the paragraph “Markers examined and construction of sequences datasets”. This dataset contains two sequences at random, identical or not, for a number of species belonging to the taxonomic group of interest.

For each within-species dataset, clustering was performed using the *sumaclust* program (Mercier et al., 2013, see Script3A_Arth02_Clustering.sh from the Supplementary material) with the *-n* option (normalization by alignment length) based on the sequence similarities first calculated using the *sumatra* program (see above; Mercier et al., 2013). Threshold values (*-t* option) ranging from 0.90 to 1 at 0.01 steps were tested for all markers except Coll01 and Olig01 for which wider ranges ([0.70 – 1] and [0.80 – 1], respectively) were selected based on the within-genus and within-species sequence similarity probability distributions determined previously (see Figure 2). Clustered datasets were then explored to calculate five different variables at each clustering threshold (see Script3B_Arth02_Oversplitting_Overmerging.Rmd from the Supplementary Material): 1) the number of clusters; 2) the percentage of MOTUs containing one single species; 3) the percentage of MOTUs containing one single genus; 4) the percentage of species gathered in one single MOTU; 5) the percentage of genera gathered in one single MOTU among genera represented by several sequences. Variables 2 and 3 are indicative of appropriate MOTU merging of sequences at the species and genus levels, respectively, while variables 4 and 5 are indicative of appropriate MOTU splitting at the species and genus levels, respectively.

These values were also used to calculate three measures of error. We defined the over-merging rate as *(100 - the percentage of MOTUs containing one single species)/100*; and the over-splitting rate as *(100 - the percentage of species gathered in one single MOTU)/100*. These two values belong to a [0,1] interval. The summed error rate was then calculated as the sum of the over-merging and over-splitting rates. For this estimate, we assigned the same weight to over-splitting and over-merging.

Results

Our *in-silico* PCRs amplified between 101,955 (Arth02) and 3,202,507 (Bact02) sequences per marker (Table S2). After data filtering, we retained between 510 (Coll01) and 707,874 (Bact02) sequences per marker. The within-species dataset comprised between 118 (Coll01) and 10,000 (Bact02, Euka02, Fung02, Sper01, COI-BF1/BR2, Inse01) sequences, while the within-genus dataset comprised between 74 (Coll01) and 10,000 (Euka02 and Sper01) sequences per marker.

Clustering thresholds determined from probability distributions of within-species and within-genus sequence similarities

The probability distributions of within-species and within-genus sequence similarities showed very contrasting patterns between the generalist and the specific markers (Figure 2).

For Arth02 and most of the markers targeting broad taxonomic groups (Bact02, Euka02, and Sper01), the distributions of within-species and within-genus similarities were rather similar, both showing a mode at very high similarity values (Figure 2). Fung02 showed a slightly different pattern, as the within-genus similarities had a very broad distribution. Conversely, for COI-BF1/BR2 and the more specific markers (Coll01, Inse01, and Olig01), the distributions of sequence similarities were very different, with two clearly distinct peaks. Within-species similarities remained very high (mostly above 0.95), while within-genus similarities generally showed lower values (mode around 0.88-0.90 for COI-BF1/BR2 and Inse01, and below 0.80 for Olig01 and Coll01).

For all markers, criterion *i* (avoid over-splitting) yielded the lowest thresholds (Table 2), with very low values for Coll01 and Olig01. Conversely, criterion *ii* (avoid over-merging) yielded extremely high values, except for Coll01. For all generalist markers and Arth02, limiting over-merging would require setting clustering thresholds at 0.99 or higher. The same objective would entail a slightly lower threshold for COI-BF1/BR2 and Inse01 (0.98) and down to 0.94 for Olig01. For Coll01, criterion *ii* resulted in a very low threshold (0.77), because many within-genus comparisons showed very low similarity values.

Criteria *iii*-a and *iii*-b searching a balance between over-merging and over-splitting yielded somehow contrasting results across markers. For COI-BF1/BR2 and the three specific markers (Coll01, Inse01, and Olig01), the within-genus and within-species similarities showed clearly distinct peaks (Figure 2). As a consequence, the intersection between the two curves could effectively represent the point minimizing both over-merging and over-splitting (see discussion), and the midpoint between the modes also identified rather similar threshold values. On the contrary, for the generalist markers and Arth02, the within-species and within-genus similarities showed very high overlap and similar modes, and the density distributions actually intersected at values lower than both modes. The midpoint between the modes continued to identify threshold values intermediate between the peaks of within-species and within-genus similarities.

Rates of over-splitting and over-merging

For all markers, irrespective of the clustering threshold examined (values ≥ 0.70 for Coll01, ≥ 0.80 for Olig01 and ≥ 0.90 for the other markers), the percentage of MOTUs containing one single species was higher than 50%, and that of MOTUs containing one single genus was higher or close to 70% (Figure 3). Overall, for the generalist and intermediate markers, these two percentages showed a regular increase with the clustering threshold. For the specific markers as well as Fung02 and COI-BF1/BR2, they reached values close to 100% for high thresholds. Unsurprisingly, the two percentages tended to be lower for the generalist markers than for the specific markers at a given threshold, indicating that the former are more sensitive to over-merging. Fung02 was a notable exception, since about 87% and 97% of MOTUs contained one single species and one single genus, respectively, at the 0.97 threshold, which is frequently adopted as clustering threshold for fungal ITS sequences. These values were comparable to those observed for COI-BF1/BR2 and the specific markers, for which > 85% and > 98% of MOTUs contained one single species or one single genus, respectively, for thresholds ≥ 0.95.

The percentages of species and genera gathered in one single MOTU decrease at a similar rate with the clustering threshold, with generally a sharp drop at high thresholds (≥ 0.98; Figure 3). However, the pattern of MOTU splitting was less characteristic of generalist vs. specific markers. For some markers (Euka02, Sper01, Arth02, Inse01), the percentage of species or genera gathered in a single MOTU remained higher or close to 50% up to high thresholds (0.98). On the contrary, for Bact02, Fung02, COI-BF1/BR2, Coll01 and Olig01, these percentages dropped quickly when the clustering threshold increased, indicating that these markers are susceptible to over-splitting.

For all markers, the number of clusters generally increased regularly with the clustering threshold up to 0.97-0.98 (Figure 3), followed by a sharp rise up to 1 (which was however less obvious for Euka02 and Olig01). For example, for Bact02, the number of clusters more than doubled between 0.97 (2862 clusters) and 1 (6461 clusters).

Our results showed clear patterns for over-merging and over-splitting rates, with over-splitting quickly increasing and over-merging quickly decreasing at high clustering thresholds (Figure 4). For several markers, the summed error showed a relatively clear minimum at specific clustering thresholds (Figure 4): 0.96-0.99 for Bact02, 0.97-0.99 for Euka02 and Arth02, 0.96-0.98 for Sper01, 0.93-0.96 for COI-BF1/BR2, and 0.94-0.97 for Inse01. The minimum was much less evident for Fung02, Coll01 and Oligo01, these markers showing relatively similar summed error rates over a broad range of clustering thresholds (Fung02: 0.91-0.98; Coll01: 0.89-0.97, with multiple minima; Oligo01: 0.84-0.96, with multiple minima).

DISCUSSION

Sequence clustering approaches are routinely used for the identification of MOTUs in metabarcoding studies, and they often resort to methods based on similarity values. Still, selecting a clustering threshold for a given marker more than often relies on common practices and rules of thumb rather than on proper scientific argument. By analyzing extensive sequence data deposited in public databases for a range of generalist and specialist markers, we showed that different thresholds can be selected depending on the marker and on the criterion favored by researchers. All studied markers but one (COI-BF1/BR2) are situated in non-protein coding genes (Table 1), and this has an influence on levels of sequence diversity. More variability might be expected in protein-coding genes due to the redundancy of the genetic code. Yet, for all markers including COI-BF1-BR2, the 10% quantile of the within-species similarity probability distribution was almost always lower than the 0.97 clustering threshold traditionally used in barcoding for markers targeting protein-coding genes like COI (Hebert et al., 2003), or for microbial MOTU delimitation (Bálint et al., 2016). This indicates indicating that some level of over-splitting can occur when using this threshold.

COI-BF1/BR2 is the only marker amplifying a fragment of a protein-coding gene, and it would have been logical to observe singular patterns for this marker. However, this was not the case, and COI-BF1/BR2, although designed to target arthropods (Elbrecht & Leese 2017) like Arth02, actually showed a behavior very similar to the more specific Inse01 targeting insects. The similarity between COI-BF1/BR2 and the more specific markers might be related to their high resolution, which allows the successful distinction of closely related species even on the basis of relatively short sequences (Elbrecht & Leese, 2017; Ficetola et al., 2021). Furthermore, at 0.94, which is a suitable clustering threshold for COI-BF1/BR2, about 88% of the MOTUs contain a single species, and about 88% of the species are gathered in a single MOTU (Figure 3), indicating that MOTU richness at this threshold is a reasonably good proxy for the number of species detected with this marker. This is corroborated by the number of clusters observed at this threshold (5659), which is comparable to the expected number of species (5000, Table S2) in the within-species dataset used to obtain Figure 3. Several COI markers are routinely used in metabarcoding, and COI-BF1/BR2 shows a large overlap with many of them (Elbrecht & Leese, 2017). We can thus expect that optimal clustering thresholds for COI-BF1/BR2 can also be rightfully applied to markers targeting a slightly different COI region.

Although the within-genus similarity values were generally lower than the within-species similarities for all the markers, the overlap between the two distributions was dependent on the generalist vs. specific nature of the marker. For some specific markers (e.g. Coll01 and Olig01), distinct peaks were visible for the two similarity metrics (Figure 2). Within-species similarities generally were >0.90, while within-genus values were <0.80. Such a pattern is expected for markers with an excellent taxonomic resolution and designed to identify taxa at the species level. Conversely for the generalist markers, within-species and within-genus similarity probability distributions largely overlapped and the differences between the peaks were minimal. Nevertheless, even for these markers, the density of within-species similarity distribution was consistently higher than that of within-genus similarity distribution at high similarity values. This suggests that the probability of observing the corresponding sequence similarity is higher within species than within genera. In other words, at high sequence similarities, a MOTU is more likely to represent a species than a genus. This result is confirmed by the fact that the percentage of MOTUs containing a single species is always higher than 50%, whatever the clustering threshold or the marker considered (Figure 3).

The sequences used as a primary source of information in this study were downloaded from the EMBL public database, therefore our results are probably highly dependent on the quality of the data deposited. Even though broad-scale analyses suggest that sequence data from public database are generally reliable (Leray et al., 2019), errors in the sequence itself (e.g. wrong nucleotide, or more complex errors like insertions, deletions, inversions, duplications or pseudogene sequences) and taxonomic mislabeling can occur. Organisms that are difficult to identify based on morphology are particular susceptible to wrong taxonomic information (Bridge et al., 2003, Bidartondo 2008, Valkiūnas et al., 2008, Mioduchowska et al., 2018). While errors in the sequence will affect within-species sequence similarity negatively, the effect of taxonomic mislabeling is more diffuse. For example, in a group like springtails where species delimitation is tricky (Porco et al., 2012), the existence of cryptic species will decrease within-species sequence similarity while increasing over-splitting rates. In a group like Bacteria, type strains are sometimes entered at the species level in the NCBI (EMBL) taxonomy (Federhen 2015), leading to an inflation of within-genus similarity and over-merging rates. In any case, database errors will make within-species and within-genus similarities distributions more difficult to distinguish and clustering thresholds trickier to identify, thus the over-splitting or over-merging rates reported here could be artificially higher than in reality.

In this work, we came up with a global measure of the error associated with a given clustering threshold, that we called the “summed error”. We calculated it by summing over-splitting and over-merging rates, assuming both have the same cost for biodiversity studies. However, it is possible to assign a differential weight to over-splitting and over-merging. For instance, if the aim is to reach conservative estimated of alpha diversity (i.e. avoid over-splitting), more weight can be assigned to over-splitting rate. Conversely, if the aim is to tease apart closely related species, that differ in their sensitivity to environmental stressors or in threat levels, one may prefer to avoid over-merging, particularly when extensive reference databases are available (Roy et al., 2019, Lopes et al., 2021).

For most of the markers we examined, the summed error approach provided relatively clear results and identified a range of threshold values that minimized the summed error. For instance, for Euka02, the summed error was relatively low at thresholds between 0.96 and

0.99 (Figure 4), indicating a good trade-off between over-merging and over-splitting. Interestingly, this range of values was also highlighted by the analysis of probability distributions (Table 2). Indeed, 0.96 is the threshold minimizing over-splitting for Euka02 while 0.99 is the balance (midpoint) threshold. The consistency of values obtained with very different approaches supports the robustness of our conclusions.

However, for a few markers, the threshold values minimizing summed error yielded somewhat less clear patterns. For Fung02, the summed error rate was rather constant (36-37%) at all the thresholds between 0.91 and 0.98, while it quickly increased for higher clustering thresholds. For Coll01 and Oligo01, the summed error rate showed multiple minima, some of which at very low clustering thresholds (Figure 4). In principle, increasing the threshold value should determine a monotone decrease of over-merging, and a monotone increase of over-splitting (Figure 1B). However, at low similarity values this was not always the case (Figure 4). This probably occurs because, for these markers a large proportion of sequences have pairwise similarities of 0.80-0.85 (Figure 2), and this might affect the identification of clusters, with some sequences clustering together e.g. at 0.85 but not at 0.86 similarity values. We also note that these similarity values match the ones corresponding to the intersection between the within-genus and within-species similarities for these markers (Table 2). It is also possible that, at this level of sequence similarity, there is strong uncertainty between MOTUs representing different hierarchical levels of taxonomy.

Our results provide quantitative data that can help researchers set their optimal clustering thresholds and understand the consequences of choosing low or high threshold values. If a clear minimum exists for the summed error rate, it probably represents an excellent trade-off between over-merging and over-splitting. In this sense, a threshold value ranging from 0.97 to 0.99 is probably appropriate for both Bact02 and Euka02, while Arth02 should accommodate a slightly higher range (0.98-0.99) and a threshold of 0.97 seems to be more suitable for Sper01. For Inse01 and COI-BF1/BR2, lower threshold values (0.94-0.97 and 0.93-0.96, respectively) are more judicious. All these values match with those obtained on the basis of within-species and within-genus similarities (Table 2). However, for Coll01, Oligo01 and Fung02, the summed error rate does not provide clear indications, and within-species and within-genus similarity distributions (e.g. midpoint between modes) might be more informative to set the clustering threshold (Figure 2 and Table 2).

The selection of clustering thresholds can have strong effect in the estimates of MOTUs richness (Figure 3), still it is important to remember that it often does not have a tremendous effect on the ecological message conveyed by metabarcoding data. For instance, Clare et al. (2016) examined different clustering thresholds to analyze dietary overlap between skinks and shrews in Mauritius. Although high clustering thresholds yielded a larger number of MOTUs, ecological conclusions remained rather consistent overall. Therefore, provided that appropriate parameters are considered (e.g. alpha diversity measured using Hill's numbers with $q > 0$ instead of richness, beta diversity estimates), the interpretation of data can be relatively robust (Clare et al., 2016, Roy et al., 2019, Calderón-Sanou et al., 2020, Mächler et al., 2021). Nevertheless, we discourage the blind application of one single clustering threshold like the classical 0.97, as it can have very different meaning across markers, and can inflate MOTU richness for fast-evolving markers. Instead, we advocate the ad-hoc definition of the most appropriate thresholds, depending on the research aims, the potential costs of over-splitting and over-merging, and the features of the studied markers.

Acknowledgments

The authors wish to thank Clément Lionnet for help with the EMBL database release 140, and three anonymous reviewers for valuable suggestions on a first version of this work. This study was supported by the European Research Council under the European Community's Horizon 2020 Programme, Grant Agreement no. 772284 (IceCommunities).

References

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution, 9*, 134-147.

Antich, A., Palacin, C., Wangensteen, O. S., & Turon, X. (2021). To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics, 22*, 177.

Baamrane, M. A. A., Shehzad, W., Ouhammou, A., Abbad, A., Naimi, M., Coissac, E., … Znari, M. (2012). Assessment of the food habits of the Moroccan dorcas gazelle in M'Sabih Talaa, West Central Morocco, using the *trnL* approach. *PLoS ONE, 7*, e35643.

Bálint, M., Bahram, M., Eren, A. M., Faust, K., Fuhrman, J. A., Lindahl, B., … Tedersoo, L. (2016). Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiology Reviews, 40*, 686-700.

Bhat, A. H., Prabhu, P., & Balakrishnan, K. (2019). A critical analysis of state-of-the-art metagenomics OTU clustering algorithms. *Journal of Biosciences, 44*, 9.

Bidartondo, M. I. (2008). Preserving accuracy in GenBank. *Science, 319*, 1616.

Bienert, F., De Danieli, S., Miquel, C., Coissac, E., Poillot, C., Brun, J. J., & Taberlet, P. (2012). Tracking earthworm communities from soil DNA. *Molecular Ecology, 21*, 2017-2030.

Bohmann, K., Elbrecht,V., Carøe, C., Bista, L., Leese, F., Bunce, M., Yu, D. W., … Creer, S. (in press). Strategies for sample labelling and library preparation in DNA metabarcoding studies. *Molecular Ecology Resources*.

Botnen, S. S., Davey, M. L., Halvorsen, R., & Kauserud, H. (2018). Sequence clustering threshold has little effect on the recovery of microbial community structure. *Molecular Ecology Resources, 18*, 1064-1076.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). OBITOOLS: a Unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources, 16*, 176-182.

Bridge, P. D., Roberts, P. J., Spooner, B. M., & Panchal, G. (2003). On the unreliability of published DNA sequences. *New Phytologist, 160*, 43-48.

Brown, E. A., Chain, F. J. J., Crease, T. J., MacIsaac, H. J., & Cristescu, M. E. (2015). Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecology and Evolution, 5*, 2234-2251.

Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., & Thuiller, W. (2020). From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices? *Journal of Biogeography, 47,* 193–206.

Capo, E., Giguet-Covex, C., Rouillard, A., Nota, K., Heintzman, P., Vuillemin, A. …Parducci, L. (2021). Lake sedimentary DNA research on past terrestrial and aquatic biodiversity: Overview and recommendations. *Quaternary 4,* 6.

Chen, W., & Ficetola, G. F. (2020). Statistical and numerical methods for Sedimentary-ancient-DNA-based study on past biodiversity and ecosystem functioning. *Environmental DNA, 2,* 115–129.

Clare, E. L., Chain, F. J. J., Littlefair, J. E., & Cristescu, M. E. (2016). The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. *Genome, 59,* 981-990.

Couton, M., Baud, A., Daguin-Thiébaut, C., Corre, E., Comtet, T., & Viard, F. (2021). High-throughput sequencing on preservative ethanol is effective at jointly examining infraspecific and taxonomic diversity, although bioinformatics pipelines do not perform equally. *Ecology and Evolution, 11,* 5533-5546.

Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., … Weaver, L. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. *Molecular Ecology Resources, 18,* 940-952.

Eichmiller, J. J., Miller L. M., & Sorensen, P.W. (2016). Optimizing techniques to capture and extract environmental DNA for detection and quantification of fish. *Molecular Ecology Resources, 16,* 56-68.

Elbrecht, V., & Leese, F. (2017). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Frontiers in Environmental Science, 5,* 11.

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J. N., … Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ, 4,* 12.

Epp, L. S., Boessenkool, S., Bellemain, E. P., Haile, J., Esposito, A., Riaz, T., … Brochmann, C. (2012). New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology, 21,* 1821-1833.

Fahner, N. A., Shokralla, S., Baird, D. J., & Hajibabaei, M. (2016). Large-scale monitoring of plants through environmental DNA metabarcoding of soil: Recovery, resolution, and annotation of four DNA markers. *PLoS ONE, 11,* e0157505.

Federhen, S. (2015). Type material in the NCBI Taxonomy Database. *Nucleic Acids Research, 43,* D1086-D1098.

Ficetola, G. F., Boyer, F, Valentini, A. Bonin, Meyer, A., Dejean, T., … Taberlet, P. (2021). Comparison of markers for the monitoring of freshwater benthic biodiversity through DNA metabarcoding. *Molecular Ecology, 30,* 3189–3202.

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., … Pompanon, F. (2010). An *in silico* approach for the evaluation of DNA barcodes. *BMC Genomics, 11,* 434.

Froslev, T. G., Kjoller, R., Bruun, H. H., Ejrnaes, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications, 8,* 11.

Guardiola, M., Uriz, M. J., Taberlet, P., Coissac, E., Wangensteen, O. S., & Turon, X. (2015). Deep-sea, deep-sequencing: metabarcoding extracellular DNA from sediments of marine canyons. *PLoS ONE, 10,* e0139633.

Guerrieri, A., Bonin, A., Münkemüller, T., Gielly, L., Thuiller, W., & Ficetola, G. F. (2021). Effects of soil preservation for biodiversity monitoring using environmental DNA. *Molecular Ecology, 30,* 3313-3325.

Hebert, P. D. N., Ratnasingham, S., & deWaard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B-Biological Sciences, 270,* S96-S99.

Jackson, M. A., Bell, J. T., Spector, T. D., & Steves, C. J. (2016). A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. *PeerJ, 4,* 19.

Janssen, P., Bec, S., Fuhr, M., Taberlet, P., Brun, J.-J., & Bouget, C. (2018). Present conditions may mediate the legacy effect of past land-use changes on species richness and composition of above- and below-ground assemblages. *Journal of Ecology, 106,* 306-318.

Kopylova, E., Navas-Molina, J. A., Mercier, C., Xu, Z. Z., Mahe, F., He, Y., … Knight, R. (2016). Open-source sequence clustering methods improve the state of the art. *mSystems, 1,* 16.

Kunin, V., Engelbrektson, A., Ochman, H., & Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology, 12,* 118-123.

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H. L., Buckley, T. R., … Holdaway, R. (2018). Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *New Zealand Journal of Ecology, 42,* 10.

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences of the United States of America, 116,* 22651-22656.

Lopes, C. M., Baêta, D., Valentini, A., Lyra, M. L., Sabbag, A. F., Gasparini, J. L., … Zamudio, R. K. (2021). Lost and found: Frogs in a biodiversity hotspot rediscovered with environmental DNA. *Molecular Ecology, 30,* 3289-3298.

Macher, T.-H., Beermann, A. J., & Leese, F. (2021). TaxonTableTools: A comprehensive, platform-independent graphical user interface software to explore and visualise DNA metabarcoding data. *Molecular Ecology Resources, 21,* 1705-1714.

Machida, R. J., Hashiguchi, Y., Nishida, M., & Nishida, S. (2009). Zooplankton diversity analysis through single-gene sequencing of a community sample. *BMC Genomics, 10,* 438.

Mächler, E., Walser, J.-C., & Altermatt, F. (2021). Decision-making and best practices for taxonomy-free environmental DNA metabarcoding in biomonitoring using Hill numbers. *Molecular Ecology, 30,* 3326-3339.

Mercier, C., Boyer, F., Bonin, A., & Coissac, E. (2013). SUMATRA and SUMACLUST: fast and exact comparison and clustering of sequences. *Programs and Abstracts of the SeqBio 2013 Workshop*, 27-29.

Meyer, C. P., & Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology, 3,* 2229-2238.

Mioduchowska, M., Czyz, M. J., Goldyn, B., Kur, J., & Sell, J. (2018). Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal *cox1* gene primers too "universal" ? *PLoS ONE, 13,* e0199609.

Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., … Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources, 18,* 927-939.

Paliy, O., & Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology, 25*, 1032-1057.

Porco, D., Bedos, A., Penelope, G., Janion, C., Skarżyński, D., Stevens, M. I., … Deharveng, L. (2012). Challenging species delimitation in Collembola: cryptic diversity among common springtails unveiled by DNA barcoding. *Invertebrate Systematics, 26*, 470-477.

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Roy, J., Mazel, F., Sosa-Hernández, M. A., Dueñas, J. F., Hempel, S., Zinger, L., & Rillig, M. C. (2019). The relative importance of ecological drivers of arbuscular mycorrhizal fungal distribution varies with taxon phylogenetic resolution. *New Phytologist, 224*, 936-948.

Schloss, P. D. (2021). Amplicon sequence variants artificially split bacterial genomes into separate clusters. *mSphere, 6*, e00191-00121.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). Environmental DNA for biodiversity research and monitoring. Oxford University Press, Oxford.

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., … Willerslev, E. (2007). Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Research, 35*, e14.

Taberlet, P., Prud'homme, S. M., Campione, E., Roy, J., Miquel, C., Shehzad, W., … Coissac, E. (2012). Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology, 21*, 1816-1820.

Tatangelo, V., Franzetti, A., Gandolfi, I., Bestetti, G., & Ambrosini, R. (2014). Effect of preservation method on the assessment of bacterial community structure in soil and water samples. *FEMS Microbiology Letters, 356*, 32-38.

Valkiūnas, G., Atkinson, C. T., Bensch, S., Sehgal, R. N., & Ricklefs, R. E. (2008). Parasite misidentifications in GenBank: how to minimize their number? *Trends in Parasitology, 24*, 247-248.

Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S. Fourth Edition. Springer, New York.

Wei, Z.-G., Zhang, X.-D., Cao, M., Liu, F., Qian, Y., & Zhang, S.-W. (2021). Comparison of methods for picking the operational taxonomic units from amplicon sequences. *Frontiers in Microbiology*, *12*, 644012.

Zinger, L., Bonin, A., Alsos, I., Bálint, M., Bik, H., Boyer, F., … Taberlet, P. (2019). DNA metabarcoding - need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology, 28*, 1857-1862.

Zinger, L., Chave, J., Coissac, E., Iribar, A., Louisanna, E., Manzi, S., … Taberlet, P. (2016). Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys based on soil DNA. *Soil Biology & Biochemistry, 96*, 16-19.

Data Accessibility

Raw data obtained from EMBL r140 (*ecopcr* files) and example scripts to prepare the datasets and perform the analyses are available on Dryad: https://doi.org/10.5061/dryad.crjdfn353.

Authors Contribution

All authors conceived the idea for the manuscript, AB and GFF designed the study, AB performed the analyses, AB and GFF generated the figures and drafted the manuscript, and all authors contributed with discussions and edits.

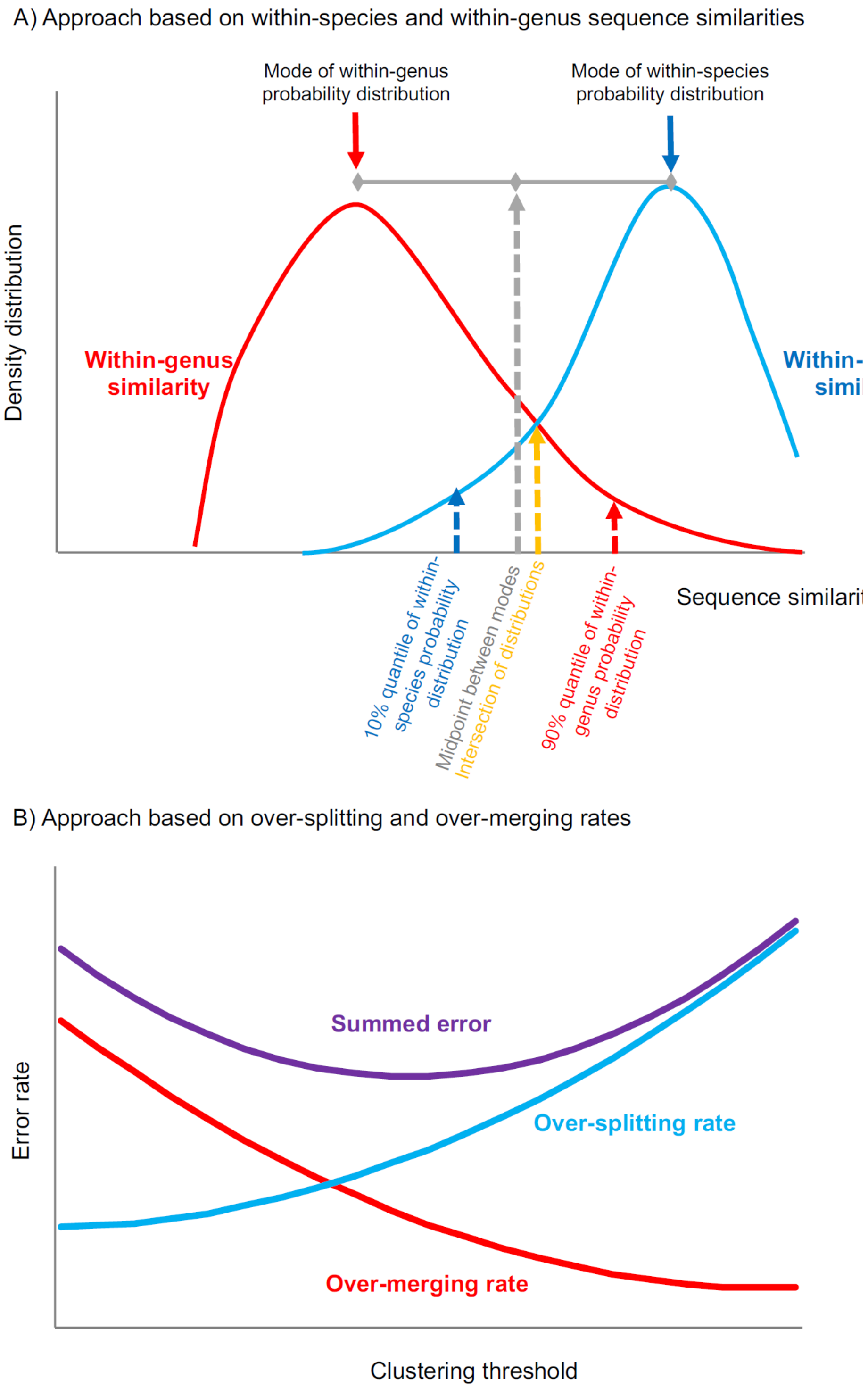
Table 1. Characteristics of the nine studied markers.

Marker	**Target gene**	**Target group**	**Taxonomic level**	**Taxonomic resolution ***				**Reference(s)**
				Species level	**Genus level**	**Family level**	**Order level**	
Bact02	V4 region of the 16S rDNA gene	Bacteria	Superkingdom	19.6%	55.7%	55.1%	60.2%	Taberlet et al., (2018)
Euka02	V7 region of the 18S rDNA gene	Eukaryota	Superkingdom	47.0%	59.5%	68.3%	67.1%	Guardiola et al., (2015)
Fung02	ITS1 nuclear rDNA gene	Fungi	Kingdom	72.5%	90.2%	87.7%	85.5%	Epp et al., (2012), Taberlet et al., (2018)
Sper01	P6 loop of the intron of the chloroplastic *trnL* gene	Spermatophyta	Clade < kingdom	21.5%	36.9%	77.4%	89.6%	Taberlet et al., (2007)
Arth02	16S mitochondrial rDNA gene	Arthropoda	Phylum	68.6%	89.6%	97.5%	100.0%	Taberlet et al., (2018)
COI-BF1/BR2	Cytochrome c oxidase subunit 1 mitochondrial gene	Arthropoda	Phylum	85.6%	97.0%	95.1%	93.5%	Elbrecht & Leese (2017)
Coll01	16S mitochondrial rDNA gene	Collembola	Class	80.5%	87.2%	75.0%	NA	Janssen et al., (2018)
Inse01	16S mitochondrial rDNA gene	Insecta	Class	87.8%	96.8%	95.4%	79.3%	Taberlet et al., (2018)
Olig01	16S mitochondrial rDNA gene	Oligochaeta	Subclass	89.3%	95.7%	100.0%	100.0%	Bienert et al., (2012), Taberlet et al., (2018)

*Percentage of discriminated taxa among taxa amplified *in silico*, as calculated by the *ecotaxspecificity* program from the OBITools. Reported from Taberlet et al., (2018) for all markers, except for COI-BF1/BR2 for which these values were determined using the sequences amplified *in silico* from EMBL r140.

Table 2. Values of the different thresholds estimated for the nine studied markers on the basis of within-species and within-genus sequence similarities.

Target	Bact02	Euka02	Fung02	Sper01	Arth02	COI-BF1/BR2	Coll01	Inse01	Olig01
Criterion *i*: Avoid over-splitting (10% quantile of within-species probability distribution)	0.961	0.962	0.885	0.967	0.986	0.937	0.739	0.944	0.855
Criterion *ii*: Avoid over-merging (90% quantile of within-genus probability distribution)	1.000	1.000	0.986	1.000	1.000	0.975	0.765	0.981	0.944
Criterion *iii*-a: Balance-a (intersection of within-species and within-genus probability distributions)	0.982	0.976	0.949	0.980	0.989	0,955	0.849	0.964	0.920
Criterion *iii*-b: Balance-b (midpoint between modes)	0.997	0.995	0.972	0.997	0.996	0.936	0.856	0.948	0.880



A) Approach based on within-species and within-genus sequence similarities

B) Approach based on over-splitting and over-merging rates

Figure 1. Different approaches to identify the most appropriate clustering thresholds. A): approach based on similarities between sequences belonging to different individuals from the same species (blue curve), and similarities between sequences belonging to different species from the same genus (red curve). One can choose to minimize the risk that different sequences from the same species are split in different MOTUs (over-splitting risk; e.g. 10% quantile of the distribution of within-species similarities), the risk that sequences from different species belonging to the same genus are clustered in the same MOTU (over-merging risk; e.g. 90% quantile of within-genus similarities), or one can try to find a balance between the risks of over-splitting and over-merging (e.g. with the intersection between probability distributions, or the midpoint between the modes of both distributions). B) Approach based on rates of over-splitting and over-merging. One can compare the over-splitting (blue) and the over-merging (red) rates, and/or one can identify the thresholds minimizing the sum of these rates (violet).

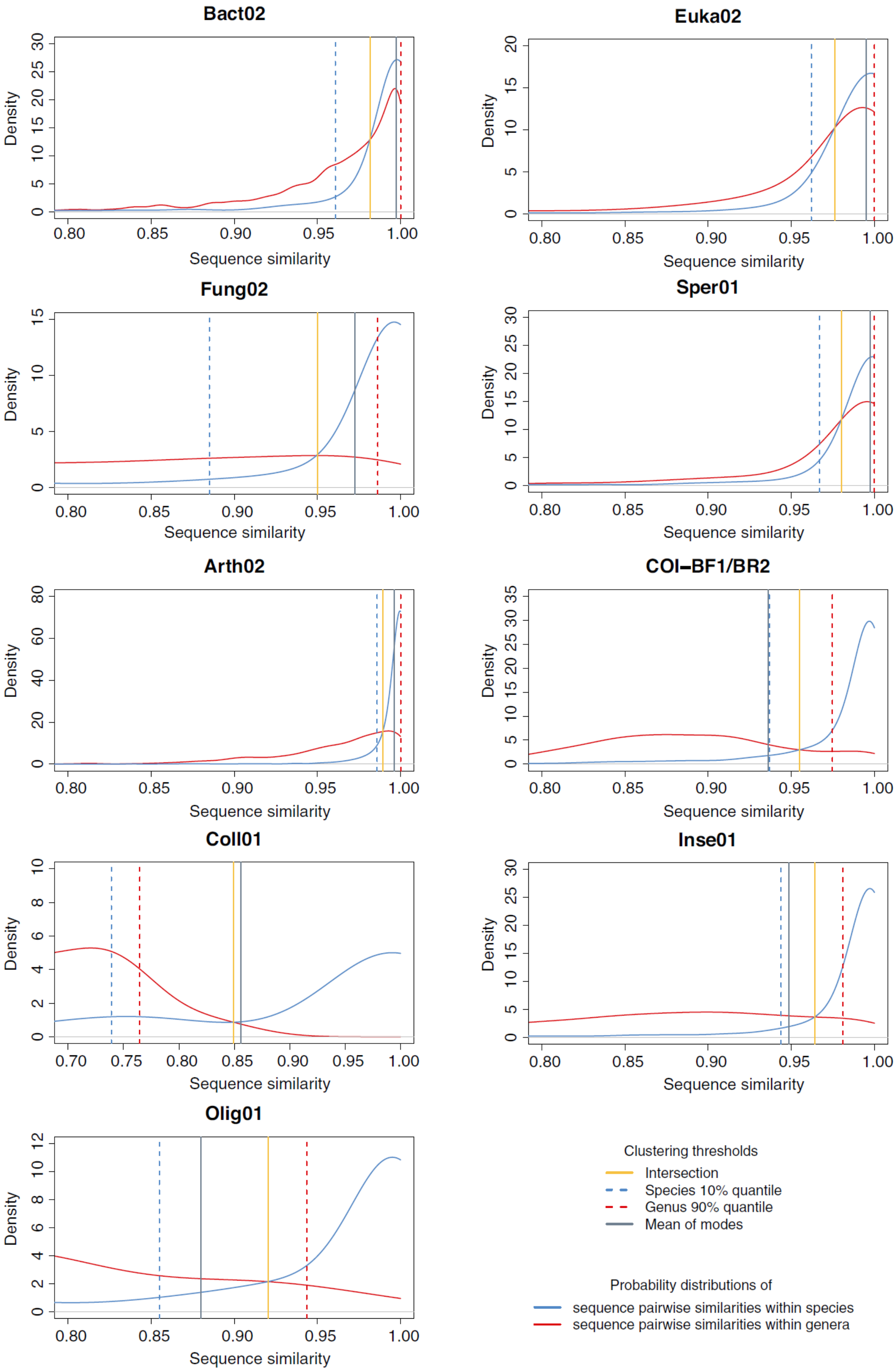


Figure 2. Density probability distributions of sequence pairwise similarities within species (blue lines) and within genera (red lines) for the nine studied markers. For each marker, vertical dotted lines represent the 10% quantile of the within-species probability distribution (blue; threshold limiting over-splitting) and the 90% quantile of the within-genus probability distribution (red; threshold limiting over-merging). Vertical full lines represent the intersection of the within-species and within-genus probability distributions (yellow, balance-a) and the midpoint between modes (grey, balance-b)

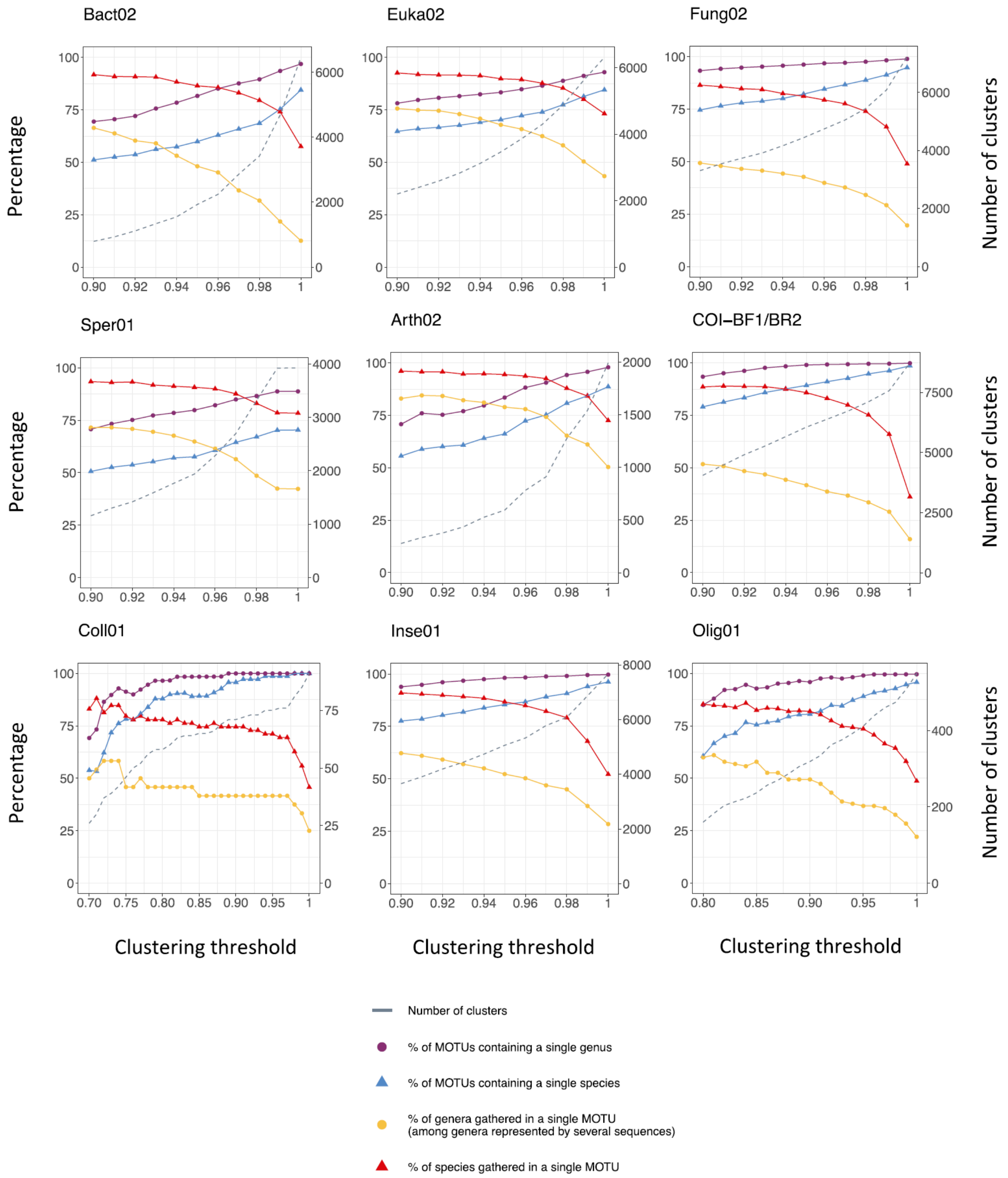


Figure 3. Evolution of over-splitting and over-merging rates for a range of clustering thresholds, for the nine studied markers.

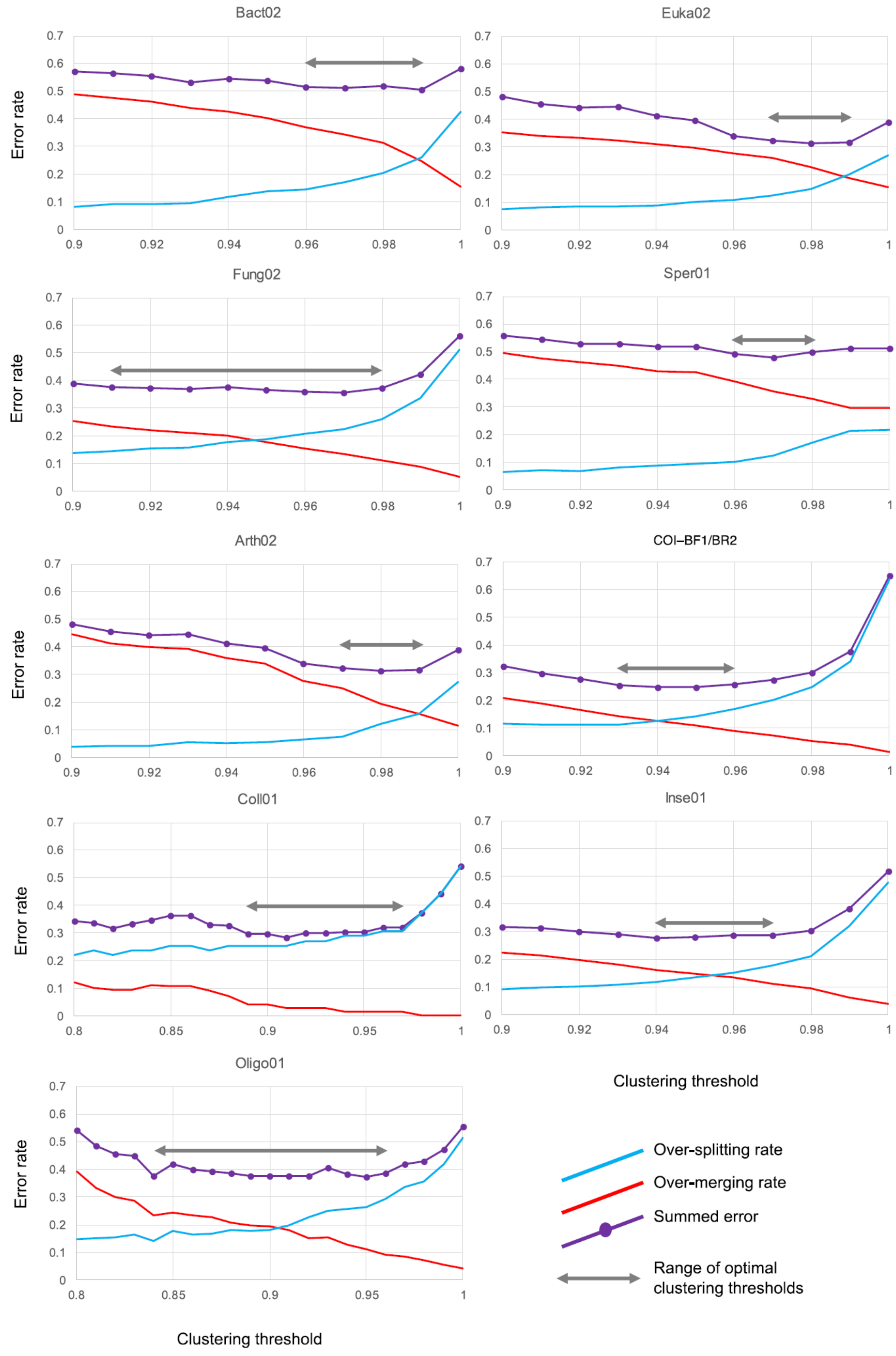


Figure 4. Over-splitting (blue) and over-merging (red) rates, as well as the summed error rate (i.e. over-splitting rate + over-merging rate; violet), for the nine studied markers across a range of clustering thresholds. Horizontal grey arrows indicate the range for which the summed error rate is minimal.