From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices?

Irene Calderón-sanou, Tamara Münkemüller, Frédéric Boyer, Lucie Zinger, Wilfried Thuiller

From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices?

Running title: Sensitivity of eDNA-based ecological results

Irene Calderón-Sanou[1]*, Tamara Münkemüller[1], Frederic Boyer[1], Lucie Zinger[2] & Wilfried Thuiller[1]

[1]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, Laboratoire d'Ecologie Alpine, F-38000, Grenoble, France

[2]Ecole Normale Superieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'Ecole Normale Superieure (IBENS), F-75005, Paris, France

*Correspondance: E-mail: irecalsa@gmail.com

ACKNOWLEDGMENTS: We thank the large team of researchers and students that helped collect the data. The research received funding from the French Agence Nationale de la Recherche (ANR) through the GlobNets (ANR-16-CE02-0009) project, and from "Investissement d'Avenir" grants managed by the ANR (Trajectories: ANR-15-IDEX-02; Montane: OSUG@2020: ANR-10-LAB-56). All computations were performed using the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr).

ABSTRACT

Aim

Environmental DNA (eDNA) is increasingly used for analysing and modelling all-inclusive biodiversity patterns. However, the reliability of eDNA-based diversity estimates is commonly compromised by arbitrary decisions for curating the data from molecular artefacts. Here, we test the sensitivity of common ecological analyses to these curation steps, and identify the crucial ones to draw sound ecological conclusions.

Location

Valloire, French Alps.

Taxon

Vascular plants and Fungi.

Methods

Using soil eDNA metabarcoding data for plants and fungi from twenty plots sampled along a 1000-m elevation gradient, we tested how the conclusions from three types of ecological analyses: (i) the spatial partitioning of diversity, (ii) the diversity-environment relationship, and (iii) the distance-decay relationship, are robust to data curation steps. Since eDNA metabarcoding data also comprise erroneous sequences with low frequencies, diversity estimates were further calculated using abundance-based Hill numbers, which penalize rare sequences through a scaling parameter, namely the order of diversity q (Richness with q=0, Shannon diversity with q~1, Simpson diversity with q=2).

Results

We showed that results from different ecological analyses had varying degrees of sensitivity to data curation strategies and that the use of Shannon and Simpson diversities led to more reliable results. We demonstrated that MOTU clustering, removal of PCR errors and of cross-sample contaminations had major impacts on ecological analyses.

Main conclusions

In the Era of Big Data, eDNA metabarcoding is going to be one of the major tools to describe, model and predict biodiversity in space and time. However, ignoring crucial data curation steps will impede the robustness of several ecological conclusions. Here, we propose a roadmap of crucial curation steps for different types of ecological analyses.

KEYWORDS: Data curation strategies, distance-decay, environmental DNA, Hill numbers, metabarcoding, sensitivity analysis, spatial partitioning of diversity.

1 INTRODUCTION

Understanding the structure and distribution of biodiversity across space and time is a critical goal in ecology. The development of environmental DNA (eDNA) metabarcoding approaches now facilitates the monitoring of species at biogeographical scales and across the whole tree of life (Drummond et al., 2015; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). It is now possible to tackle unresolved questions that could not be addressed with traditional biodiversity surveys so far. For example, eDNA-based biodiversity studies have enabled the spatial partitioning of diversity (i.e. gamma, alpha and beta diversity) of so far elusive taxa in both terrestrial and marine environments (e.g. marine viruses and protists, soil fungi and bacteria), thereby improving our understanding of their community assembly processes and of their role in structuring communities and networks at global scales (e.g. Lima-Mendez et al., 2015; Tedersoo et al., 2014). However, while the eDNA metabarcoding approach promises substantial advances in macroecology and multi-taxa studies, it requires an appropriate and careful processing of the tremendous amount of sequences generated to draw robust and ecologically meaningful conclusions.

Indeed, the analyses of diversity patterns (e.g. alpha- and beta-diversity; Whittaker, 1960) across space and of the processes generating these patterns are traditionally based on community matrices representing the presence/abundance of species across samples. In eDNA metabarcoding surveys, the data consist of hundreds to millions of DNA sequencing reads from the hundreds to thousands of species co-occurring within samples. Using bioinformatics, these data are then transformed in community matrices, but with species replaced by DNA sequences, and species abundance replaced by a number of sequencing reads. While, in an ideal world, one sequence should correspond to a single species, in practice, it can correspond to several species if the DNA region has a low taxonomic resolution, and more critically, one species can be represented by tens to thousands of variant sequences. Amongst those variants, a few are biologically meaningful (e.g. intraspecific variability), but the large majority of them are technical errors produced at the different stages of the lab treatments, from DNA extraction to sequencing (see Table 1 and Appendix S1, Bálint et al., 2016; Taberlet, Bonin, Zinger, & Coissac, 2018). These errors can represent more than 70% of the sequences in raw metabarcoding datasets, and have usually low frequencies (e.g. singletons; S. P. Brown et al., 2015). If interpreted as genuine, these sequences can therefore inflate diversity by several orders of magnitude and lead to flawed ecological interpretations (Kunin, Engelbrektson, Ochman, & Hugenholtz, 2010). Molecular protocols are thus applied to reduce and/or control specific technical errors accumulated during the data production. For example, replicated PCR amplification and use of negative controls allow identifying artefactual sequences resulting from random errors introduced by DNA polymerases or sequencers, as well as reagent contaminants (de Barba et al., 2014). However, error rates remain high even with the most stringent molecular protocols (Bálint et al., 2016; Taberlet et al., 2018), which has led to the development of bioinformatics algorithms aiming

at detecting errors known to occur during data generation (e.g. PCR errors or chimeric sequences). Also, most of these tools require specifying thresholds and parameter values, which are usually based on arbitrary decisions and visual assessments. An example is the classification of sequence variants into MOTUs (Molecular Operational Taxonomic Units) based on the similarity of sequences. While this step is critical because MOTUs are used as a proxy for species in the majority of DNA metabarcoding studies (Appendix S1), MOTUs are commonly defined using a 97% sequence similarity threshold, a value historically defined as the similarity level of full-length 16S rRNA barcodes below which bacterial strains necessarily belong to different species (Stackebrandt & Goebel, 1994). However, the optimal threshold value to define MOTUs depends on the focal taxa and polymorphism/length of the DNA marker used (e.g. E. A. Brown, Chain, Crease, MacIsaac, & Cristescu, 2015; Kunin et al., 2010). It also depends on the PCR/sequencing error rate, which varies across molecular protocols, and depends on the amount of target DNA: when it is low, each genuine DNA fragment has a higher probability of being amplified at each PCR cycle (Taberlet et al., 2018).

Hence, using DNA metabarcoding requires taking several methodological choices. Beyond those related to molecular protocols and bioinformatics software, one of the most critical choice is to decide which data curation steps to include in the curation procedure. Indeed, each step directly affects the community matrix obtained, by influencing the final list of MOTUs and/or their frequencies within samples. Previous methodological studies have thus underlined the importance of data curation steps on the reliability of ecological analyses and provided guidelines for bioinformatics decision-making (e.g. Alberdi, Aizpurua, Gilbert, & Bohmann, 2018; Schloss, 2010). However, most of these studies tested the influence of data curation procedures on a single metric or ecological question. However, questions related to local community richness can be very sensitive to errors (Flynn, Brown, Chain, MacIsaac, & Cristescu, 2015), while comparisons of communities' composition might be less affected (Leray & Knowlton, 2015; Taberlet et al., 2018). In addition, most studies have focused on microbial communities (bacteria or fungi), and few have addressed such questions to macro-organisms. Finally, most published tests have so far relied on mock communities (i.e. positive controls) usually made of DNA extracts for few known species. While mock communities are useful to identify errors and estimate error rates, the conclusions cannot easily be translated to realistic environments with rich and complex communities (Alberdi et al., 2018).

Here, we address how methodological choices related to the DNA metabarcoding data curation strategy influence the results for different types of ecological analyses and their related diversity metrics. We used soil eDNA data from an elevation gradient in the French Alps, and focused on plants and soil fungi to represent both macro- and microorganisms, as well as DNA markers with different length (Table 2). Patterns of plant diversity have been extensively studied in this area (e.g. Chalmandrier, Münkemüller, Lavergne, & Thuiller, 2015) and serve as a good reference to evaluate

the results estimated from eDNA metabarcoding data. We subjected these data to 256 different data curation strategies, which correspond to all possible combinations of seven critical data curation steps. We then tested how the curation strategies influence the inferences drawn from three different ecological analyses: 1) a spatial partitioning of diversity (i.e. gamma, alpha and beta diversities) to estimate the regional and local diversity of the gradient, 2) a diversity-environment relationship, to analyse the influence of environment on the local community diversity (alpha), and 3) a distance-decay analysis, to evaluate if similarities between communities (beta) decrease with increasing geographic distances. To this end, we first checked the accuracy of eDNA metabarcoding data in detecting ecological patterns by comparing the eDNA-based diversity patterns with the expected values based on mock communities and traditional botanical surveys (only available for plants). Second, we did an overall sensitivity analysis to test the sensitivity of ecological results to the data curation strategy. Finally, with a variance partitioning analysis we identified the crucial curation steps (i.e. those that introduced more variance to the results) to include or consider in the curation procedure.

To achieve these objectives, we built on Hill numbers (Hill, 1973) to estimate diversity, which unifies mathematically the best known diversity measures in ecology through a unique parameter q (i.e. Richness at q=0, the exponential of Shannon entropy at q~1 and the inverse of Simpson at q=2). In this framework, the weight of rare species decreases when increasing the value of the parameter q. This feature is particularly relevant for DNA metabarcoding data, since artefactual sequences are usually rare compared to the genuine ones (Bálint et al., 2016; Taberlet et al., 2018). Hill numbers can thus penalize these rare sequences at different degrees: q=1 is the order of diversity that levels the MOTUs exactly according to their relative abundances, while q<1 overweigh rare MOTUs and q>1 overweight abundant MOTUs. As a result, we could expect that diversity measures that give less importance to rare sequences (i.e. q>0) are less sensitive to the data curation strategy, because they penalize the artefactual sequences targeted by the curation steps.

2 MATERIAL AND METHODS

2.1 Sample data

Soil cores were sampled at 10 different elevations equally distributed across an elevation gradient in the northern French Alps (from 1748 m to 2725 m a.s.l.) in 2012. At each elevation, two 10m × 10m plots were selected (20 plots in total). In each plot, 21 soil cores distributed along the two diagonals were sampled. Soil corers were cleaned and sterilized between each sample collection. Extracellular DNA was then extracted twice, from 15 g as described in Taberlet, Prud'Homme, et al., (2012). Aboveground plant community information (hereafter observed plant diversity) was obtained in each plot with a botanical survey conducted during the annual productivity peak (mid-July) using the Braun-Blanquet cover-abundance scale (Braun-Blanquet, 1946).

2.2 Molecular analyses

eDNA-based plant diversity was estimated by targeting a vascular plant specific marker (P6 loop of chloroplast trnL, Table 2). It targets highly conserved priming sites across vascular plants and amplifies a short region, which is desired when working with degraded DNA. eDNA-based fungal diversity was assessed using the nuclear ribosomal Internal Transcribed Spacer 1 (ITS1, Table 2). For each DNA extract, PCRs were run in duplicate leading to four technical replicates per core sample and DNA marker. PCR thermocycling conditions and mixes composition and purification can be found in Table S2.1 in Appendix S2. To control for potential contaminants, extraction and PCR blank controls were included in the experiment. To control for false positives caused by tag-switching events, we also defined "sequencing blank controls", i.e. tag combinations not used in our experimental design, but that could be formed at the library preparation or sequencing stage (See Appendix S1). We also included positive controls in this experiment, which consisted of a mix of DNA extracted from 16 plant species. Fir this, genomic DNA was extracted from leaf tissue using the DNeasy Plant Kit (Qiagen GmbH, Hilden, Germany), quantified, diluted at different concentrations for each species and mixed to form a mock community (species composition provided in Table S2.2, Appendix S2). Positive controls allow for quantification of technical biases introduced by PCR and sequencing. Illumina sequencing was performed on a HiSeq platform (2x100bp paired-end reads) for plant amplicons and on a MiSeq (2x250bp paired-end reads) for fungi amplicons, both using the paired-end technology.

2.3 Bioinformatics analyses

The Illumina sequencing paired-end reads (Table S2.3) were pre-processed for each marker with three procedures: (i) assembling forward and reverse paired-end reads based on their overlapping 3'-end sequences, (ii) assigning each read to its respective sample (demultiplexing) and (iii) combining strictly identical sequences into unique DNA sequences while keeping information on their abundance (number of sequencing reads) in each sample (dereplication). Then we systematically processed the dereplicated sequences following common data curation procedures that included removal of sequences with low paired-end alignment scores, removal of singletons, removal of short sequences and removal of sequences containing ambiguous bases (not to be confounded with a phred-quality filtering; Fig.1a, Table 1 and Table S2.4). Singletons are sequences that occur only once in the whole dataset and many studies agree that their removal is necessary to reduce data complexity/computational time and because they mostly correspond to molecular artefacts that may inflate disproportionately diversity indices (S. P. Brown et al., 2015; Kunin et al., 2010). In our data, they represented 70-80% of the total number of sequences but only 1-15% of the total number of sequencing reads for plants and fungi respectively (Table S2.3 in Appendix S2). We finally assigned each remaining sequence to a taxonomic clade with the *ecotag* command from the OBITOOLS

200 software package (Boyer et al., 2016) that uses a lowest common ancestor algorithm for the
201 assignment, and the EMBL database version 133 as a reference.
202
203 Next, data from each marker were processed following a range of different data curation
204 strategies to test the sensitivity of ecological analyses to different methodological choices (Fig. 1b).
205 To do so, we selected seven important steps: (i) removal of PCR errors, (ii) filtering of highly
206 spurious sequences, (iii) removal of chimeras, (iv) sequence classification into MOTUs (MOTU
207 clustering), (v) removal of reagent contaminants, (vi) cross-sample contamination cleaning and (vii)
208 dysfunctional PCRs filtering (see Table 1, Appendix S1 and Table S2.4 in Appendix S2 for target
209 errors and step descriptions). Curation steps were either kept or excluded, and were always performed
210 in the same order in each data curation strategy. For the MOTU clustering step, when kept, three
211 clustering thresholds were tested (1, 2 or 3 mismatches allowed between pairwise aligned sequences).
212 We used here raw mismatches rather than percentages of dissimilarities because the DNA markers
213 used are short (<100 bp) and/or highly polymorphic in length. Using the percentages of dissimilarity
214 in this case would penalize more little differences when alignments are short than when they are long.
215
216 All different possible combinations of these curation strategies were implemented (Fig.1b). Most
217 of the curation steps were done using the software OBITOOLS (Boyer et al., 2016). Chimera detection
218 was performed with UCHIME (Edgar, Haas, Clemente, Quince, & Knight, 2011) and we used
219 SUMACLUST (Mercier, Boyer, Bonin, & Coissac, 2013) for MOTU clustering due to its ability in
220 handling large datasets and its flexibility for defining the clustering threshold (see Table S2.4 for
221 more details on the algorithm). After data curation, PCR replicates were summed and standardized by
222 the total number of reads in each core sample. We then pooled the samples for each of the 20 plots to
223 obtain a single community per plot. For this, MOTUs abundance (already standardized by the number
224 of reads) were summed and standardized by the number of samples in each plot. For each of the data
225 curation strategies, we obtained a community matrix with rows representing plots and columns
226 representing all the MOTUs obtained after curation, which we used here as a proxy for species.
227 Therefore, our sensitivity analysis was conducted on a total of 256 matrices for each DNA marker
228 (Fig.1c).
229
230 **2.4 Ecological questions**
231 We tested the sensitivity of the results for three common ecological analyses to the above-
232 mentioned data curation strategies using MOTUs as equivalent of species:
233
234 *Spatial partitioning of diversity* - We used the multiplicative diversity partitioning approach
235 (Whittaker, 1960) to analyse gamma (here the diversity across the entire gradient), alpha (diversity of
236 local communities) and beta diversity (diversity between communities). In the Hill numbers

framework, gamma diversity is the effective number of species in the pooled meta-community (i.e. across all plots), alpha diversity is the effective number of species per community (i.e. plot), and beta diversity is the effective number of communities, calculated as the ratio of gamma diversity to alpha diversity. We followed Chao, Chiu, & Jost, (2014)'s definition where beta diversity is independent of alpha and ranges from 1 (all communities are identical) to the total number of communities N (when N=20 all communities are different). We limited our study to taxonomic diversity, because the DNA markers we used here are rather short (Table 2) and are highly variable in length, which make them not suitable for inferring accurate phylogenetic relationships at the scale of the community.

Diversity-environment relationship (*Alpha~SOM content*) – Diversity is often linked to abiotic drivers, and a common ecological research question is how alpha diversity changes along an environmental gradient. Here, we fitted a linear model to determine changes in alpha diversity along a gradient of soil organic matter content (SOM content), known to be a strong predictor of diversity changes in the study site (Ohlmann et al., 2018).

Distance-decay relationship (*Similarity~geographic distance*) – Species' distributions and resulting diversity patterns are controlled by both species dispersal abilities and spatial turnover of environmental conditions (Tuomisto, 2003). One hypothesis is thus that spatially distant communities are more different than close communities ("distance-decay", Green et al., 2004; Tuomisto, 2003). We used the Jaccard-type overlap (U_{qN}) as a measure of similarity (Chao et al., 2014) and we fitted a linear model using the log transformation of similarity against the geographic distance to evaluate the distance-decay. The geographic distance between plots was calculated with Euclidean distances using the elevation values of the plots.

For each DNA marker (plant and fungi), we calculated the gamma, alpha and beta diversities (spatial partitioning of diversity) for each of the 256 community matrices obtained from the different metabarcoding data curation strategies using Hill numbers with values of q={0,0.5,1,2}. For the diversity-environment and the distance-decay relationships, we fitted our models to each community matrix and extracted the slopes and the R-squares of the models. Alpha diversity and community similarity were calculated using Hill numbers with values of q={0,1,2}.

2.5 Sensitivity analyses

Detectability of ecological patterns - To test the ability of eDNA metabarcoding data and of the different data curation strategies to detect ecological patterns we (1) evaluated the completeness of the sampling unit (plot), and (2) used the observed plant diversity and positive controls as references to evaluate the accuracy of the ecological results. We acknowledge that eDNA-based diversity is expected to slightly diverge from observed diversity (see discussion) but they should follow similar

trends (Hiiesalu et al., 2012; Träger, Öpik, Vasar, & Wilson, 2019; Yoccoz et al., 2012). The sampling completeness of each plot was evaluated with rarefaction curves for the different orders of diversity q={0,1,2} and for three data curation strategies with varying filtering stringency: a "no data curation" strategy with no curation step at all; a "basic curation" strategy including only the chimera removal and a traditional clustering threshold allowing three mismatches between clustered sequences and, a "rigorous curation" strategy, including all the curation steps considered here and a clustering threshold allowing two mismatches.

Overall sensitivity analyses - To test the sensitivity of the results for the different ecological analyses and their related diversity metrics to the data curation strategy, we used the variance of each diversity estimate, obtained across the 256 community matrices and for each marker (Fig. 1c). For the diversity-environment and the distance-decay relationships, we looked at the variance in the slope and the R-square of the linear regression across the 256 models for each marker. In addition, we used 'the rigorous' and 'the basic' curation strategies explained above, that correspond to commonly used pipelines, to exemplify how results can differ between studies.

Identifying the crucial steps of the curation procedure - To identify the crucial steps we did a variance partitioning analysis for each diversity metric. In respect to spatial partitioning of diversity, the diversity metrics (gamma, alpha and beta diversities) were used as the response variable in function of the curation steps. For the diversity-environment and the distance-decay relationships we used the slope and the R-square of the models as the response variable in function of the curation steps. Variance partitioning analyses were done with the R package RELAIMPO (Grömping, 2006).

3 RESULTS

3.1 Detectability of ecological patterns with eDNA metabarcoding data

Sampling completeness of the plots – For both markers/taxa, the total diversity was well represented by the number of reads sequenced, when considering the diversity at q={1,2} (Fig.S2.1 and Fig.S2.2 in Appendix S2). At q={0}, the rarefaction curve rarely saturated, but we obtained more asymptotic curves when increasing the stringency of the data curation strategy.

Spatial partitioning of diversity – Overall, we found that alpha diversity estimates at q={1,2} were closer to the observed plant diversity (Fig.2b) and to the positive controls composition (Fig. 3) than at q={0,0.5}. However, diversity at q={1} slightly underestimated gamma (Fig.2a) and beta (Fig.2c) while all diversity components were underestimated for most curation strategies at q={2} (Fig.2a-c).

Richness (q=0) was always overestimated. While we obtained very accurate results for diversity at q={0.5} when using a rigorous pipeline, a basic pipeline led to a substantial overestimation.

Diversity-environment relationship – While the expected positive slope was in most cases detected (Fig.2g) and its value was on average very similar to the one observed for plant diversity, especially when using a rigorous pipeline, it was highly overestimated for some data curation strategies at q={0,1}.

Distance-decay relationship – The expected negative slope of the distance-decay curve was always detected (Fig.2k). However, independently of the data curation strategy, the slope was always underestimated compared to the curve calculated with observed plant diversity. Also, the R-square of the distance-decay relationship was reduced at q={2} (Fig.2l).

3.2 Overall sensitivity of ecological questions and diversity metrics

The results of different ecological questions had varying degrees of sensitivity to the data curation strategies. While the estimates in all ecological questions were highly sensitive (width of the boxplots in Fig. 2), the main signal of the diversity-environment and the distance-decay relationships was consistent across most curation strategies.

Spatial partitioning of diversity - Sensitivity of gamma, alpha and beta diversity decreased for higher values of q, i.e. weighing down rare MOTUs (Fig.2a-f). Diversity estimates at q={0} were the most sensitive, with more than two orders of magnitude for both gamma and alpha (Fig.2a & b) diversities of plants. Likewise, the rigorous and basic curation strategies (circles and triangles in Fig.2) exhibited a steep difference at q={0}, which decreased when using higher values of q in the majority of cases.

Diversity-environment relationship - The interpretation of the alpha-SOM content relationship could change depending on the data curation strategy used. However, the alpha-SOM content relationship was more robust when using q={1,2}, i.e. a positive relation between alpha diversity and SOM content was detected independently of the data curation strategy used (Fig.2g,h). Patterns in fungi diversity were more robust, i.e. no relation between fungi diversity and SOM content was detected across the different pipelines. A very weak positive relation between fungi diversity and SOM content was observed for q={1,2}. The rigorous and the basic strategies led to very similar results for both DNA markers/taxa.

Distance-decay relationship – In contrast, a significant distance-decay relationship was always detected from eDNA metabarcoding data independently of the data curation strategy, but the rate at which similarity decays with increasing distance between plots (i.e. slope) slightly changed across

strategies. While very similar results were found between the rigorous and the basic strategies for the distance-decay curve of plants, the slope of the distance-decay curve for fungi was very low when using a basic instead of a rigorous strategy.

3.3 Crucial steps of the curation procedure

Overall, we found that two curation steps, the removal of PCR error and the clustering to define MOTUs, explained most of the variation in diversity estimates across data curation strategies (more than 15% each and usually more than 40% in total) for most of the diversity metrics in the ecological analyses and for both markers/taxa (Fig.4 and Fig.S2.3 in Appendix S2). Also, cross-sample contamination removal explained large parts of the variance of beta diversity in the spatial partitioning of diversity analyses (Fig.4a,b) and of R-squares and slopes in the diversity-environment (Fig.4c,d) and distance-decay (Fig.4e,f) relationships analyses.

4 DISCUSSION

Ecologists do now increasingly rely on DNA metabarcoding to measure biodiversity as this approach holds the promise of allowing testing long-standing hypotheses at spatial, temporal and taxonomic scales that were hitherto inaccessible with traditional approaches. However, the technique is still hampered by a substantial amount of technical errors (Table 1, Appendix S1; Bálint et al., 2016; Taberlet et al., 2018). Here, we sought at testing the sensitivity of the conclusions drawn from different ecological analyses and diversity metrics to the steps commonly used to curate DNA metabarcoding data from such errors. We show that ecological conclusions had varying degrees of sensitivity to the data curation strategies and that the use of metrics that are less sensitive to rare species/MOTUs (i.e. Shannon and Simpson diversity) leads to more robust diversity estimates. Also, we demonstrated that MOTU clustering, removal of PCR errors and removal of cross-sample contaminations have a major influence on ecological results, and must always be carefully included when curating DNA metabarcoding data.

The breadth of our study makes our findings generalizable to other systems. Indeed, we found similar trends in the sensitivity of gamma and alpha diversity estimates for both our observed plant diversity and the mock community (Figure 2 vs Figure 3). Second, our study focus on both plants and fungi, that widely differ in their ecological properties and the length of their markers (on average 50 bp for plants vs. 225 bp for fungi). Still, while they do not share the same diversity patterns, their sensitivity to data curation strategies were comparable. Further, we expect that our study and the experimental testing design we developed will stimulate further methodological studies (e.g. for

tropical or aquatic systems and other markers/taxa) and that they will serve as a guide to prioritize some curation steps when deciding for a curation strategy.

4.1 Linking methodological choices with ecological questions

The ecological question(s) underlying a study should lead the prioritization of the curation steps to be included in the data curation procedure, as well as the selection of appropriate diversity metrics (Fig. 5). If the aim of the study is to estimate the spatial partitioning of diversity (Fig.5a), it is important to keep in mind that all diversity components are biased by the data curation steps. Richness is highly sensitive to error accumulation, and was hence the metric responding the strongest to the data curation strategy. Consequently, if measuring richness is crucial for the study, and, thus, rare species are important, the reliability of the results must be confirmed with additional analyses. For example, a more conservative strategy (i.e. keeping only MOTUs present in more than a certain number of PCR replicates) can improve the reliability of final results, but with the risk of missing species represented by few sequences in only a few samples due to the sampling process occurring when preparing aliquots of one DNA extract (Alberdi et al., 2018). Verifying the pertinence of species detected by looking in detail into the taxonomic assignments can also improve the reliability of results, even though this could be problematic for poorly known taxa with incomplete reference databases (Cristescu, 2014). Also, positive controls (with mock communities) and numerous negative controls (extraction, PCR) must be included in all the phases of sequence generations to ensure the accuracy of richness estimates (Bálint et al., 2016). In any cases, a certain degree of uncertainty will always remain because of the complexity of deciding objectively which sequences are genuine and which are artefactual.

We corroborated that richness is a very sensitive metric and is always overestimated (Fig.2a-c). The intrinsic properties of eDNA can inflate the diversity compared to traditional surveys because eDNA can persist in the environment or be transported through space depending on the abiotic conditions (e.g. water transport, temperature, UV, or microbial activity; Barnes & Turner, 2016). This means that the diversity eDNA estimates not only encompass local and current species, but also species that are dormant (Hiiesalu et al., 2012), that were present in the recent past (Yoccoz, 2012) or that are present in the vicinity of the studied area (Taberlet et al., 2018). In other words, the spatiotemporal window captured by local eDNA diversity estimates may be larger than that captured by traditional approaches, a property that can be desirable or not depending on the question addressed. Distinguishing this feature from methodological bias remains at this stage difficult, as it may look like cross-contamination, and also because the cycle of eDNA in the environment remains poorly understood (Barnes & Turner 2016). However, it is crucial to account for eDNA properties when interpreting richness-based studies to avoid meaningless conclusions.

When the detection of rare species is not of importance, Hill numbers are a promising solution to increase the robustness of results and to avoid the inflation of diversity estimates. The Hill numbers approach has been already proposed to better estimate microbial diversity (e.g. Bálint et al., 2016; Chiu & Chao, 2016), and we corroborate its efficiency for estimating plant diversity and potentially other macro-organisms from metabarcoding data. Both, Shannon and Simpson diversity measures led to a satisfying representativeness of the sampling unit diversity and were robust to the different data curation strategies tested here, but Shannon diversity was less biased. In the same way that richness overestimated diversity, Simpson diversity tended to underestimate diversity. Diversity measures, other than richness (i.e. q>0), account for species/MOTUs abundance structure. The factors determining species' abundances in a community are not the only factors determining the MOTUs' abundances. These correspond to a pool of DNA fragments from current, dormant, or past populations (e.g. microbes) down to one (or part of one) single multicellular individual that are besides amplified by PCR. Consequently, a highly abundant MOTU does not necessarily imply that more individuals of the corresponding taxon were present, it could also be due to e.g. higher body mass, larger root systems, or slower DNA decomposition. Besides, given the exponential nature of the PCR amplification, abundant taxa become even more abundant in this step and this could lead to an underestimation of Simpson diversity. Hence, interpreting MOTUs frequency directly as species abundance can be highly misleading, and estimating species abundance in terms of number of individuals or biomass from eDNA is still a major challenge in the field (Deiner et al., 2017). However, MOTUs frequency correlates to a certain extent to species relative abundance, and more importantly, errors are usually rarer than genuine sequences (reviewed in Taberlet et al., 2018). Accordingly, Shannon diversity from eDNA samples appears here as a balanced diversity measure, robust to the data curation strategy, and hence, to rare errors. This can be generalized to all ecological analyses tested in this study. Given these results, we argue that using a complete diversity profile (for example, with q values between 0 and 2) may allow improving confidence in diversity estimates from eDNA data while getting information about MOTUs structure of abundances.

Another important outcome of our assessment is that despite the above-mentioned limits, robust conclusions can be obtained from eDNA metabarcoding data if the aim is to link local diversity (alpha) or community similarity (beta) to environmental or geographic gradients (Fig.5b). Changes in local diversity across an environmental gradient were more sensitive to the data curation strategies than the distance-decay relationship. Our results thus corroborate other studies that demonstrated the robustness of beta diversity to bioinformatics analyses (Botnen, Davey, Halvorsen, & Kauserud, 2018; Deiner et al., 2017). However, the slope of the distance-decay was always underestimated compared to that obtained from observed plant diversity. On one hand, this could result from a lack of phylogenetic resolution of the genetic marker used here, which is relatively short. In alpine ecosystems, it is common to see abundant species replaced by closely related species across an

elevation gradient (Chalmandrier et al., 2015). A genetic marker with a low phylogenetic resolution would not detect these changes and as a consequence, gamma and beta diversities would be underestimated. However, the underestimation of gamma diversity relative to alpha diversity is not strong enough, suggesting that other reasons may also explain the lower slope of the distance-decay curve for eDNA-based plant diversity. Botanical surveys used in this study represent just a local snapshot of the visible plant diversity at the sampling time, and, unlike the eDNA approach, may miss species with an offset phenology or present only in the vicinity of the sampling area (Hiiesalu et al., 2012). We can expect that the larger spatiotemporal window captured by the eDNA metabarcoding approach would thus result in higher similarity among the sites, which could be tested by increasing the botanical sampling effort across seasons and years to reduce botanical surveys biases related to the differentiated phenology of the species.

4.2 Crucial steps for designing a careful curation protocol

While we included here curation steps that are common to most bioinformatic tools (e.g. QIIME, USEARCH), we acknowledge that algorithms within OBITOOLS have their own particularities, as each of the other packages, and that the results obtained here may not be directly transferable. However, we expect that the differences from a specific software are minor compared to the differences caused by the choice of specific curation steps (Bonder, Abeln, Zaura, & Brandt, 2012). In general, we corroborate past studies concluding that the clustering threshold used for defining MOTUs leads to significant changes in diversity estimates and that this is especially important for alpha and gamma diversities, but less so for beta diversity (Botnen et al., 2018; E. A. Brown et al., 2015; Kunin et al., 2010). Additionally, we found that PCR errors and cross-sample contaminations are critical steps and that including them leads to more realistic spatial diversity patterns and estimates of diversity components. These two steps correct the diversity at local levels (i.e. sample level) and are especially important when comparing communities. To our knowledge, this is the first study testing in a systematic way the effect of these curation steps on results across different types of ecological analyses. We recommend carefully choosing the MOTU clustering threshold, e.g. empirical means can be estimated for each marker or targeted taxa using *in silico* methods with reference databases (Taberlet et al., 2018) or experimentally, using mock communities (E. A. Brown et al., 2015), and considering removing PCR errors and cross-sample contaminations when designing a curation protocol to study biodiversity patterns. Furthermore, a rigorous data curation strategy including all the curation steps of the present study allowed obtaining accurate diversity estimates and diversity-environment and distance-decay relationships. This demonstrates that the other curation steps should not be neglected.

5 DATA AVAILABILITY STATMENT

Pre-filtered sequencing data as well as associated metadata are available on the Dryad Digital Repository (doi:10.5061/dryad.0t39970).

6 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*(1), 134–147. https://doi.org/10.1111/2041-210X.12849

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*(5), 686–700. https://doi.org/10.1093/femsre/fuw017

Barnes, M. A., & Turner, C. R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, *17*(1), 1–17. https://doi.org/10.1007/s10592-015-0775-4

Bonder, M. J., Abeln, S., Zaura, E., & Brandt, B. W. (2012). Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics*, *28*(22), 2891–2897. https://doi.org/10.1093/bioinformatics/bts552

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*. https://doi.org/10.1111/1755-0998.12894

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*(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Braun-Blanquet, J. (1946). Über den Deckungswert der Arten in den Pflanzengesellschaften der Ordnung Vaccinio-Piceetalia. *Jahresbericht Der Naturforschenden Gesellschaft Graubünden*, *130*, 115–119.

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*(11), 2234–2251. https://doi.org/10.1002/ece3.1485

Brown, S. P., Veach, A. M., Rigdon-Huss, A. R., Grond, K., Lickteig, S. K., Lothamer, K., … Jumpponen, A. (2015). Scraping the bottom of the barrel: are rare high throughput sequences artifacts? *Fungal Ecology*, *13*, 221–225. https://doi.org/10.1016/j.funeco.2014.08.006

Chalmandrier, L., Münkemüller, T., Lavergne, S., & Thuiller, W. (2015). Effects of species' similarity and dominance on the functional and phylogenetic structure of a plant meta-community. *Ecology*, *96*(1), 143–153. https://doi.org/10.1890/13-2153.1

Chao, A., Chiu, C.-H., & Jost, L. (2014). Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 297–324. https://doi.org/10.1146/annurev-ecolsys-120213-091540

Chiu, C.-H., & Chao, A. (2016). Estimating and cmparing microbial diversity in the presence of sequencing errors. *PeerJ*, *4*, e1634. https://doi.org/10.7717/peerj.1634

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, *29*(10), 566–571. https://doi.org/10.1016/j.tree.2014.08.001

de Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources*, *14*(2), 306–323. https://doi.org/10.1111/1755-0998.12188

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., … Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C., … Nelson, N. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*, *4*(1). https://doi.org/10.1186/s13742-015-0086-1

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194–2200. https://doi.org/10.1093/bioinformatics/btr381

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

Green, J. L., Holmes, A. J., Westoby, M., Oliver, I., Briscoe, D., Dangerfield, M., … Beattie, A. J. (2004). Spatial scaling of microbial eukaryote diversity. *Nature*, *432*(7018), 747–750. https://doi.org/10.1038/nature03034

Grömping, U. (2006). Relative Importance for Linear Regression in *R* : The Package **relaimpo**. *Journal of Statistical Software*, *17*(1). https://doi.org/10.18637/jss.v017.i01

Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., & Weitz, J. S. (2013). Robust estimation of microbial diversity in theory and in practice. *The ISME Journal*, *7*(6), 1092–1101. https://doi.org/10.1038/ismej.2013.10

Hiiesalu, I., Öpik, M., Metsis, M., Lilje, L., Davison, J., Vasar, M., … Pärtel, M. (2012). Plant species richness belowground: higher richness and new patterns revealed by next-generation sequencing. *Molecular Ecology*, *21*(8), 2004–2016. https://doi.org/10.1111/j.1365-294X.2011.05390.x

Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, *54*(2), 427–432. https://doi.org/10.2307/1934352

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*(1), 118–123. https://doi.org/10.1111/j.1462-2920.2009.02051.x

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*, *112*(7), 2076–2081. https://doi.org/10.1073/pnas.1424997112

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., … Raes, J. (2015). Determinants of community structure in the global plankton interactome. *Science*, *348*(6237), 1262073–1262073. https://doi.org/10.1126/science.1262073

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. Abstract*, 27–29. Citeseer.

Ohlmann, M., Mazel, F., Chalmandrier, L., Bec, S., Coissac, E., Gielly, L., … Thuiller, W. (2018). Mapping the imprint of biotic interactions on β-diversity. *Ecology Letters*, *21*(11), 1660–1669. https://doi.org/10.1111/ele.13143

Schloss, P. D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Computational Biology*, *6*(7), e1000844. https://doi.org/10.1371/journal.pcbi.1000844

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, *44*(4), 846–849. https://doi.org/10.1099/00207713-44-4-846

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

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

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*(3), e14–e14. https://doi.org/10.1093/nar/gkl938

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*(8), 1816–1820. https://doi.org/10.1111/j.1365-294X.2011.05317.x

Tedersoo, L., Bahram, M., Polme, S., Koljalg, U., Yorou, N. S., Wijesundera, R., … Abarenkov, K. (2014). Global diversity and geography of soil fungi. *Science*, *346*(6213), 1256688–1256688. https://doi.org/10.1126/science.1256688

Träger, S., Öpik, M., Vasar, M., & Wilson, S. D. (2019). Belowground plant parts are crucial for comprehensively estimating total plant richness in herbaceous and woody habitats. *Ecology*, *100*(2), e02575. https://doi.org/10.1002/ecy.2575

Tuomisto, H. (2003). Dispersal, Environment, and Floristic Variation of Western Amazonian Forests. *Science*, *299*(5604), 241–244. https://doi.org/10.1126/science.1078037

White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In M. A. Innis, D. H. Gelfand, J. J. Sninsky, & T. J. White (Eds.), *PCR protocols a guide to methods and applications* (pp. 315–322). New York: Academic Press.

Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, *30*(3), 279–338. https://doi.org/10.2307/1943563

Yoccoz, N. G., BråThen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., … Taberlet, P. (2012). DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, *21*(15), 3647–3655. https://doi.org/10.1111/j.1365-294X.2012.05545.x

BIOSKETCH: Irene Calderon-Sanou is a PhD student aiming at a better understanding of multi-trophic assemblages through the use of environmental DNA. Author contributions: WT initiated the overall idea, and together with ICS, LZ and TM conceived the overall analyses. ICS, LZ and FB conceptualized the data curation strategies, ICS run the curation procedures and analysed all the results, and led the writing with the significant contribution of all co-authors.

TABLE CAPTION

TABLE 1 Brief description of classical technical errors occurring in DNA metabarcoding data, the associated data curation steps tested in the present study and the curation methodology. Target errors make reference to the errors described further in Appendix S1. See also Table S2.4 for more details on the curation steps used in this study.

TABLE 2 Characteristics of the DNA markers used to estimate eDNA-based diversity in this study.

TABLES

TABLE 1

Target error	Definition	Curation step (Abbreviation) and methodology
Mixed	Common obvious molecular/sequencing errors such as mis-paired reads, sequences with ambiguous bases, that are too short or singletons.	Common basic filtering : Removal of sequences meeting these criteria. This step is not tested here and has been applied systematically.
PCR error	Base mis-incorporation by the DNA polymerase during the PCR amplification.	PCR errors removal (PCR error): Identification of PCR errors using a model-based classification of sequences based on their similarities and abundances. The model reflects the accumulation of base mis-incorporation across PCR cycles, where genuine sequences remain more abundant than their respective errors.
Highly spurious sequences	Chimeras from multiple parents, primers dimers, etc. or sequences from highly degraded DNA fragments that largely differ from any known sequence.	Highly spurious sequences removal (Spurious): Removal of sequences of whose similarity with their closest match in public reference databases is below 70% (plants) or 50% (fungi).
Chimeras	Sequences obtained from the recombination of two or more parent sequences	Chimera detection and removal (Chimeras): Removal of sequences that have a high probability to be a subsequence from other, more abundant sequences in the dataset.
Remaining PCR errors/ biological variation	Sequences from the same species either resulting from a PCR error that could not be filtered above, or from intraspecific variability	MOTU clustering (Clustering): Clustering of sequences into MOTUs on the basis of their pairwise similarity. Here done at different sequence similarity thresholds.
External contaminants	DNA coming from an external source other than the biological sample	Reagent contaminants cleaning (Reagent): Removal of sequences that are more abundant in negative controls relative to biological samples because of the absence of other competing DNA fragments during the amplification process.
Cross-contaminations or Tag-jumps	Genuine sequences present in a sample where actually absent, either due to cross-contaminations at the bench, or due to tag-jumps occurring during the library preparation or the sequencing , i.e. switches of nucleotidic labels used to assign the sequencing reads to their samples. These contaminants are usually of much lower abundance than in their sample of origin.	Cross-sample contamination curation (Cross): If the abundance of a given MOTU in a given sample is below 0.03% of the total MOTU abundance in the entire dataset, it is considered as absent in this sample.
Dysfunctional PCRs	PCRs that are too different in comparison with their technical replicates.	Dysfunctional PCR removal (DysPCR): Removal of PCR replicates from a single biological sample that are more dissimilar to each other in MOTUs composition and structure than are the PCR obtained from other biological sample.

TABLE 2

DNA Marker	Target taxa	Forward primer (5'-3')	Reverse primer (5'-3')	Length [range] (bp)	References
P6 loop ofthe chloroplast *trn*L intron	Vascular plants	g: GGGCAATCCTGAGCCAA	h: CCATTGAGTCTCTGCACCTATC	48 [10-220]	Taberlet et al., 2007
Nuclear ribosomal DNA Internal Transcribed Spacer 1 (ITS1)	Fungi	ITS5: GGAAGTAAAAGTCGTAACAAGG	Fung02: CCAAGAGATCCGTTGYTGAAAGTK	226 [68-919]	White, Bruns, Lee, & Taylor, 1990; Taberlet et al., 2018

FIGURE CAPTION

FIGURE 1 Workflow of the sensitivity analysis. (a) Raw data are curated with basic filtering steps for each DNA marker (plants: trnL-P6 loop, fungi: ITS1). (b) Filtered data are processed using seven curation steps that were varied or removed in each data curation strategy making a total of 256 possible combinations. As a result, 256 community matrices are obtained per DNA marker and used to (c) conduct three types of ecological analyses. The range of values obtained for each ecological analysis and diversity metric represents the variance due to the data curation strategy.

FIGURE 2 Estimated values of the spatial partitioning of diversity components (a-f), of the regression parameters from the diversity-environment (g-j), and of distance-decay (k-n) relationships across the 256 curation strategies for different diversity metrics (Hill numbers, q={0,0.5,1,2}). The top row (a-c, g-j, k-l) corresponds to the plant DNA marker (trnL-P6 loop) and bottom row (d-f, i-j, m-n) to the fungi DNA marker (ITS1). Size of each box (including whiskers) represents the sensitivity of the diversity metrics or the model parameters to the data curation strategy. The circle and the triangle symbols indicate the values obtained from a rigorous and a basic curation strategy, respectively. The star symbol indicates the values calculated from botanical survey (only represented for plants, top row).

FIGURE 3 Mean diversity estimated in positive controls across the 256 data curation strategies for different diversity metrics (Hill numbers, q={0,0.5,1,2}). Size of each box (including whiskers) represents the sensitivity of the diversity metrics to the data curation strategy. The star symbol indicates the values calculated from the known species composition in positive controls, the other symbols are as in Fig. 2.

FIGURE 4 Relative importance (% of variance explained) of the data curation steps on the variability of estimated values of the spatial partitioning of diversity components (a,b) and of the parameters from the diversity-environment (c,d) and distance-decay (e,f) relationships, using Hill numbers at q={1} (see Fig.S2.3 for the other q values). The top row (a,c,e) corresponds to the plant DNA marker (trnL-P6 loop) and bottom row (b,d,f) to the fungi DNA marker (ITS1). A model was fitted independently for each diversity component (a,b) or model parameter (c-f) as response variable, with curation steps as main effects.

FIGURE 5 Guidelines to improve the reliability of ecological results when analysing eDNA metabarcoding data.

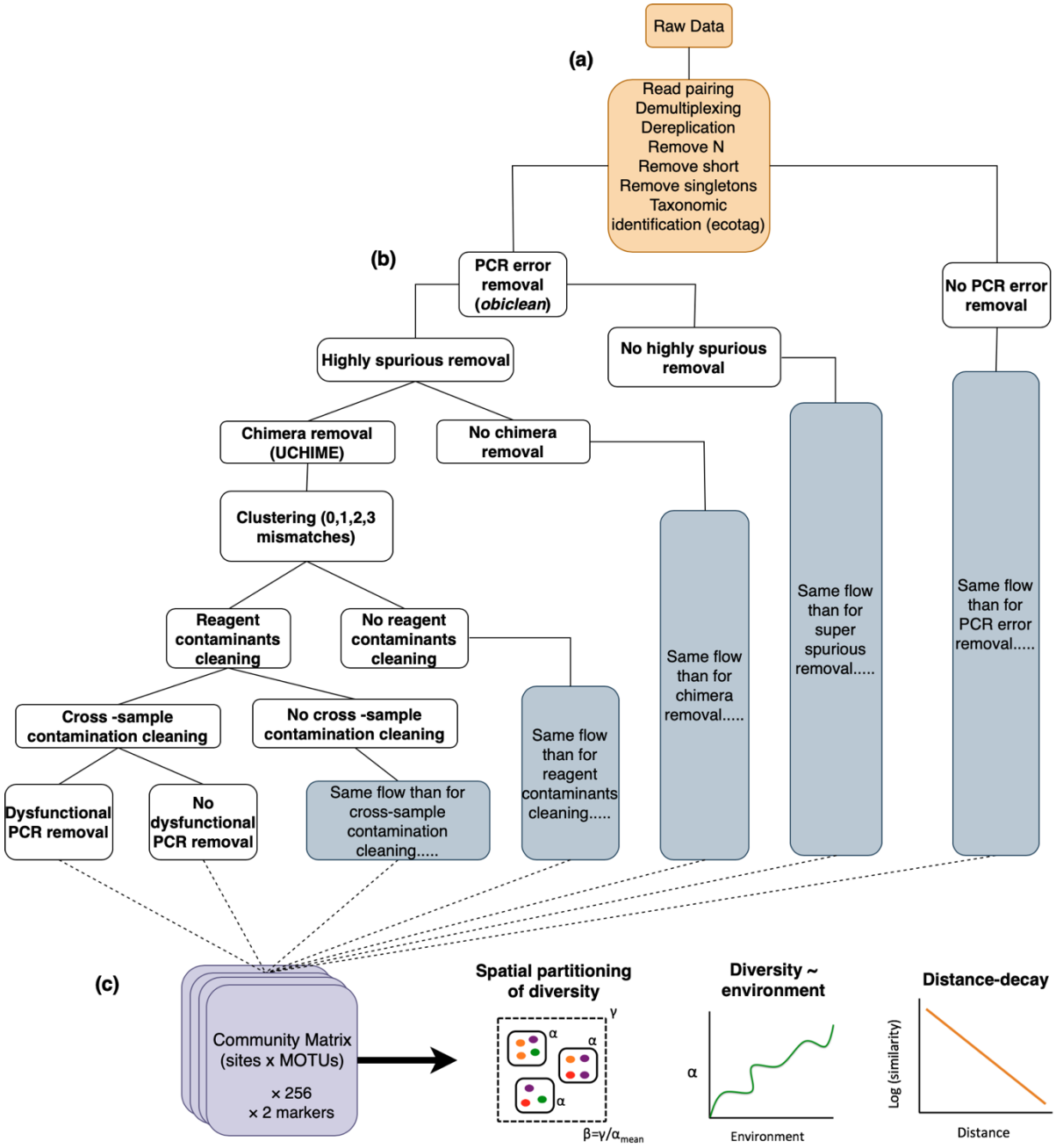

Raw Data
(a)
Read pairing
Demultiplexing
Dereplication
Remove N
Remove short
Remove singletons
Taxonomic identification (ecotag)
(b)
PCR error removal (obiclean)
No PCR error removal
Highly spurious removal
No highly spurious removal
Chimera removal (UCHIME)
No chimera removal
Clustering (0,1,2,3 mismatches)
Reagent contaminants cleaning
No reagent contaminants cleaning
Cross -sample contamination cleaning
No cross -sample contamination cleaning
Dysfunctional PCR removal
No dysfunctional PCR removal
Same flow than for cross-sample contamination cleaning.....
Same flow than for reagent contaminants cleaning.....
Same flow than for chimera removal.....
Same flow than for super spurious removal.....
Same flow than for PCR error removal.....
(c)
Community Matrix (sites x MOTUs)
× 256
× 2 markers
Spatial partitioning of diversity
γ
α
α
α
β=γ/α mean
Diversity ~ environment
α
Environment
Distance-decay
Log (similarity)
Distance

FIGURE 1

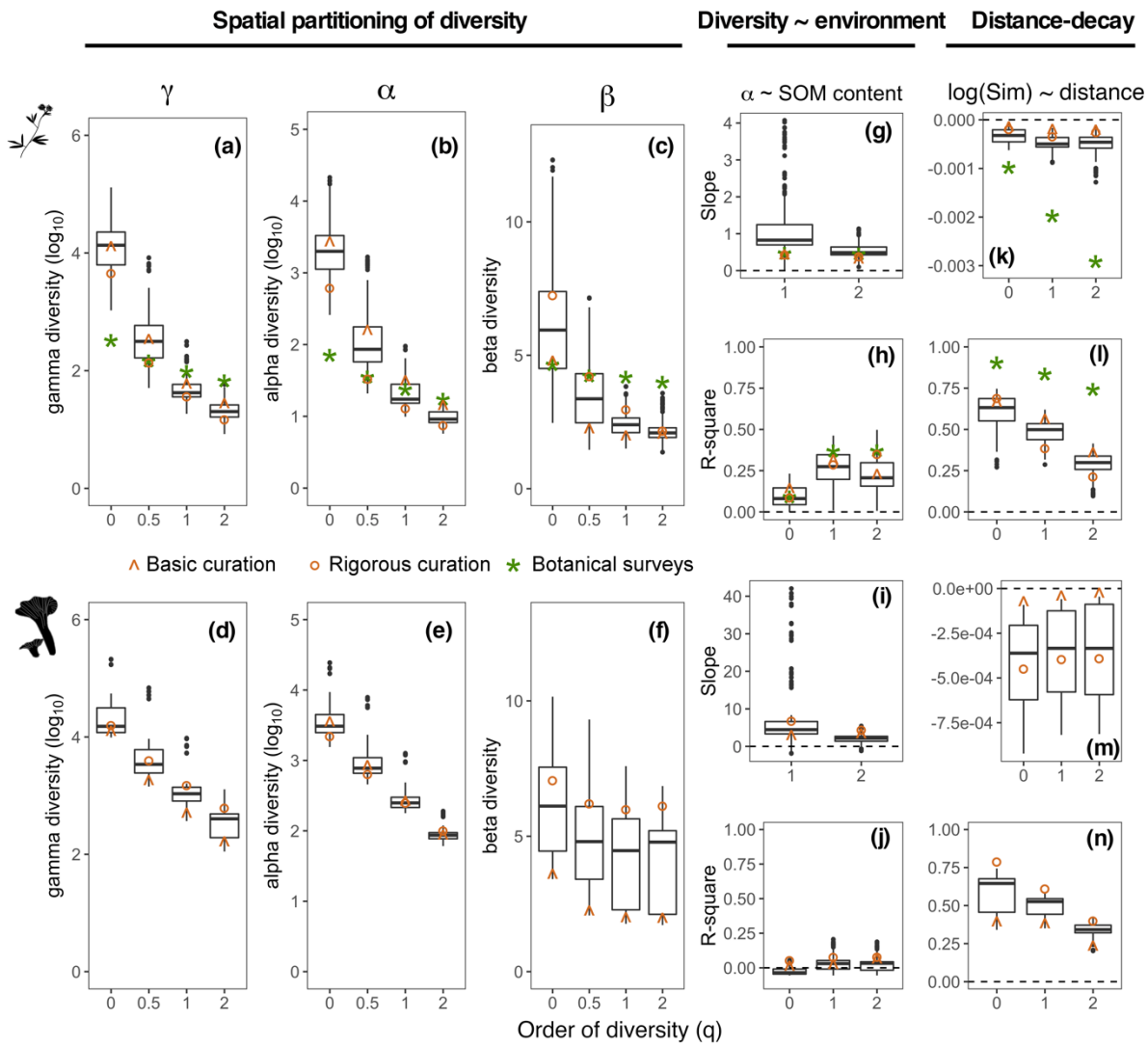

Spatial partitioning of diversity
Diversity ~ environment
Distance-decay
γ
α
β
α ~ SOM content
log(Sim) ~ distance
gamma diversity (log10)
alpha diversity (log10)
beta diversity
Slope
R-square
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
Basic curation
Rigorous curation
Botanical surveys
Order of diversity (q)

FIGURE 2

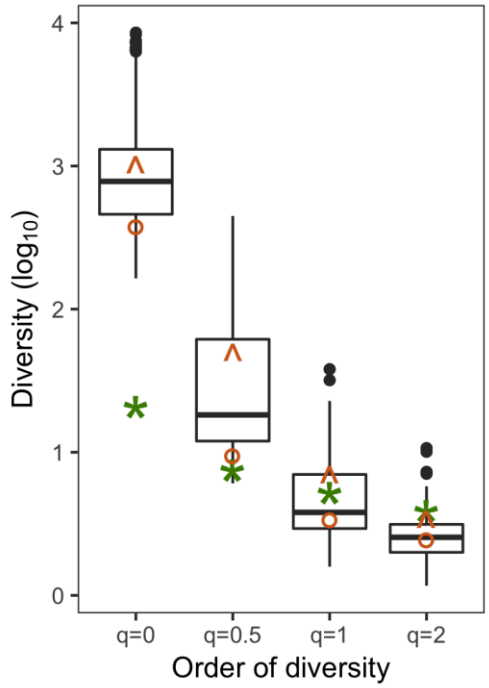


FIGURE 3

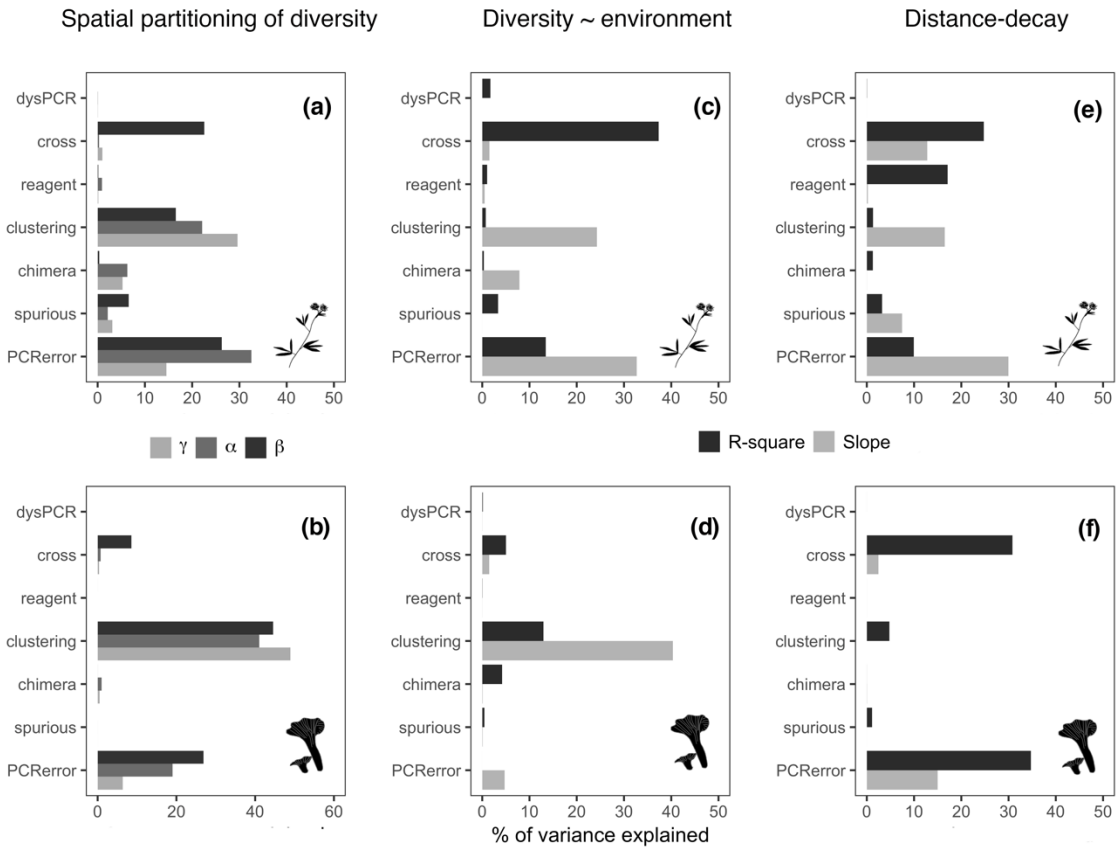
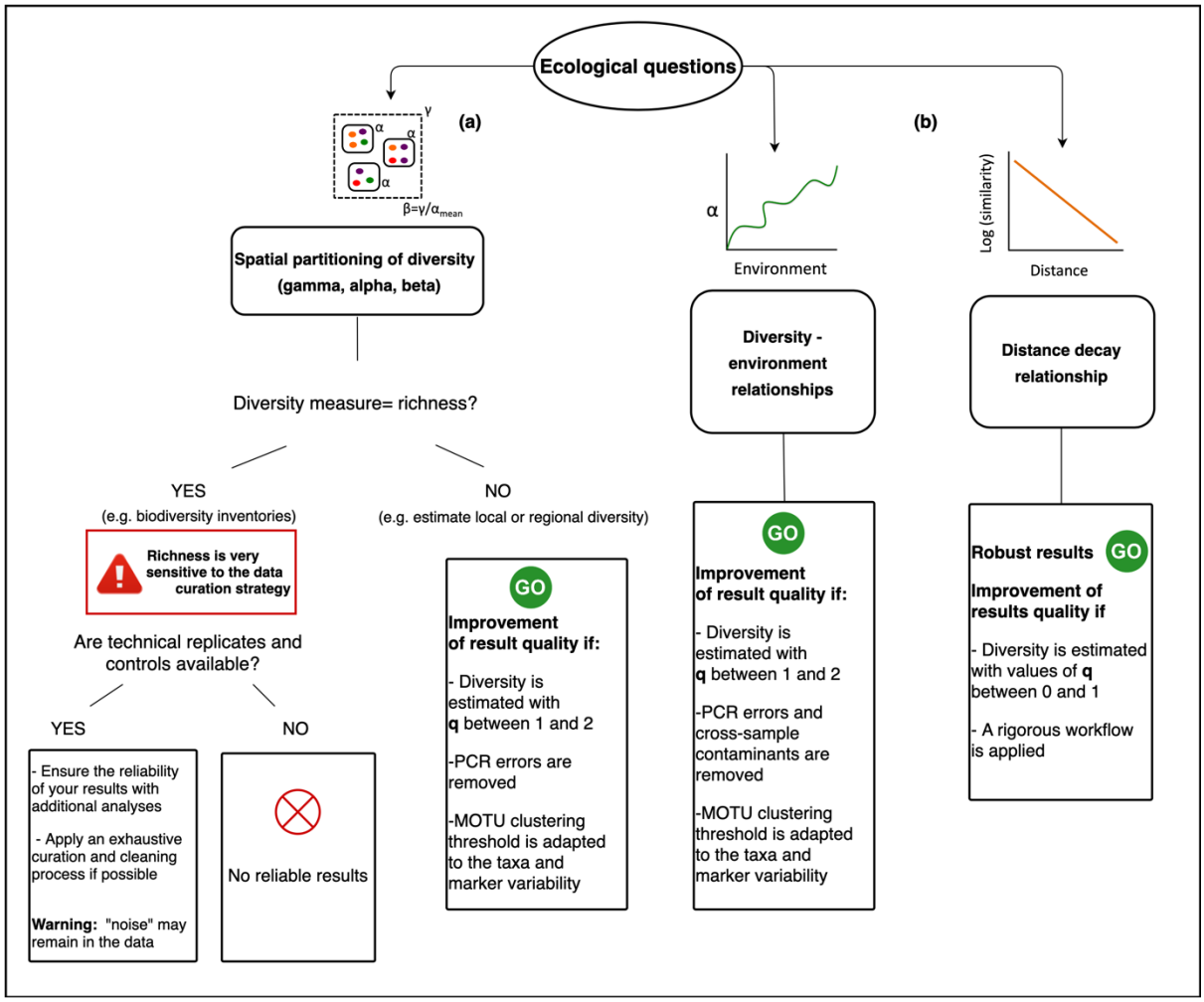

Spatial partitioning of diversity
Diversity ~ environment
Distance-decay
dysPCR
cross
reagent
clustering
chimera
spurious
PCRerror
(a)
(b)
(c)
(d)
(e)
(f)
γ α β
R-square Slope
% of variance explained

FIGURE 4

687



688
689 FIGURE 5