Received: 31 August 2020 | Accepted: 23 December 2020

DOI: 10.1111/2041-210X.13552

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

metabaR: An R package for the evaluation and improvement of DNA metabarcoding data quality

Lucie Zinger[1] | Clément Lionnet[2] | Anne-Sophie Benoiston[1] | Julian Donald[3,4] | Céline Mercier[2] | Frédéric Boyer[2]

[1]Département de Biologie, Institut de Biologie de l'ENS (IBENS), École Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France

[2]Univ. Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, LECA, Laboratoire d'Écologie Alpine, Grenoble, France

[3]Evolution et Diversité Biologique (EDB UMR5174), Université Toulouse 3 Paul Sabatier, CNRS, IRD, Toulouse, France

[4]Centre for Ecology and Conservation, University of Exeter, Penryn, UK

Correspondence
Lucie Zinger
Email: lucie@zinger.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-10-LABX-0041, ANR-10-LABX-25-01, ANR-10-LABX-56, ANR-11-BSV7-0020 and ANR-16-CE02-0009

Handling Editor: Steven Kembel

Abstract

1. DNA metabarcoding is becoming the tool of choice for biodiversity assessment across taxa and environments. Yet, the artefacts present in metabarcoding datasets often preclude a proper interpretation of ecological patterns. Bioinformatic pipelines to remove experimental noise exist. However, these often only partially target produced artefacts, or are marker specific. In addition, assessments of data curation quality and chosen filtering thresholds are seldom available in existing pipelines, partly due to the lack of appropriate visualisation tools.
2. Here, we present **metabaR**, an R package that provides a comprehensive suite of tools to effectively curate DNA metabarcoding data after basic bioinformatic analyses. In particular, **metabaR** uses experimental negative or positive controls to identify different types of artefactual sequences, that is, contaminants and tag-jumps. It also flags potentially dysfunctional PCRs based on PCR replicate similarities when those are available. Finally, **metabaR** provides tools to visualise DNA metabarcoding data characteristics in their experimental context as well as their distribution, and facilitates assessment of the appropriateness of data curation filtering thresholds.
3. **metabaR** is applicable to any DNA metabarcoding experimental design but is most powerful when the design includes experimental controls and replicates. More generally, the simplicity and flexibility of the package makes it applicable any DNA marker, and data generated with any sequencing platform, and pre-analysed with any bioinformatic pipeline. Its outputs are easily usable for downstream analyses with any ecological R package.
4. **metabaR** complements existing bioinformatics pipelines by providing scientists with a variety of functions to effectively clean DNA metabarcoding data and avoid serious misinterpretations. It thus offers a promising platform for automatised data quality assessments of DNA metabarcoding data for environmental research and biomonitoring.

KEYWORDS

contaminations, data curation, data mining, environmental DNA, high-throughput, sequencing, tag-jumps

1 | INTRODUCTION

DNA metabarcoding coupled with high-throughput sequencing is currently revolutionising the way we assess and describe biodiversity across environments and taxa, and is therefore becoming a tool of choice for basic and applied research, as well as for biomonitoring applications (Cordier et al., 2020; Deiner et al., 2017; Taberlet et al., 2018). In recent years, various bioinformatic pipelines and tools have been developed to handle DNA metabarcoding data. These include, for example, **QIIME** (Caporaso et al., 2010; Estaki et al., 2020), **OBITools** (Boyer et al., 2016; Taberlet et al., 2018), **vsearch** (Rognes et al., 2016) or **dada2** (Callahan et al., 2016). These bioinformatic packages typically perform bioinformatic analyses such as sequence alignment, clustering into Molecular Operational Taxonomic Units (MOTUs), data denoising or taxonomic assignment and ultimately produce a MOTU-by-sample matrix. This matrix, similar to the community table of community ecologists, can then be used to reveal patterns of alpha and beta diversity with more classical ecological R packages such as **vegan** (Oksanen et al., 2019) or **adiv** (Pavoine, 2020), or with packages dedicated to microbiome analyses (e.g. **phyloseq**, McMurdie & Holmes, 2013).

While the aforementioned bioinformatic tools have been heavily used, they yet hold a certain number of limitations. DNA metabarcoding generates numerous experimental biases besides PCR/sequencing errors and chimeras, which range from field or laboratory contaminations through to tag-jumps (Table 1; reviewed in Taberlet et al., 2018; Zinger et al., 2019). The processing of these artefacts is often missing in many studies, even though they can substantially affect ecological inference (Calderón-Sanou et al., 2019; Frøslev et al., 2017; Sommeria-Klein et al., 2016). Such artefacts can only be flagged and corrected by including experimental controls and experimental replicates throughout the data production process. However, most existing bioinformatic pipelines only deal with PCR/sequencing errors, and do not make use of experimental controls to filter out potential contaminants or artefacts (but see Zepeda-Mendoza et al., 2016). Second, these bioinformatic pipelines often lack tools to monitor and evaluate the bioinformatic data filtering process. As a result, it can be difficult to tune data filtering parameters, often resulting in the use of suboptimal default settings. Finally, DNA metabarcoding data are in essence multidimensional, as they encompass MOTUs, PCR product and biological sample information. This multi-fold information, often stored in separate tables, is not easily handled by most R packages for data analyses (but see e.g. **phyloseq**). As such, we currently lack effective tools for the transition of DNA metabarcoding data produced by bioinformatic analysis pipelines to ecological R packages.

To bridge this gap, we developed **metabaR**, an R package that enables the post-processing and filtering of DNA metabarcoding data already processed through bioinformatic pipelines so as to improve downstream ecological inferences. It is designed to take advantage of negative controls, positive controls and PCR replicates when available to efficiently flag and remove artefactual MOTUs or dysfunctional PCRs. It is implemented in the R statistical programming environment (R Core Team, 2020), which provides flexible analytical tools coupled with powerful graphical capabilities. **metabaR** uses these properties to provide highly customisable functions, as well as effective visualisation of DNA metabarcoding data in their experimental context. Hence, it is of direct use for any practitioner of DNA metabarcoding techniques with basic skills in R programming.

TABLE 1 Overview of DNA metabarcoding experimental artefacts

Experimental bias	Description
PCR/sequencing errors	Any MOTU resulting from base misincorporation during PCR amplification or sequencing, or base miscalling during sequencing
Contaminants	Any MOTU not originally present in a biological sample. Such contamination can occur at all stages of data production, that is, field work, DNA extraction, PCR amplification and library preparation
Tag-jumps	MOTU of which presence is erroneous in a given sample/PCR product due to a switch of so called 'tag' or 'library index', that is, a characteristic nucleotide kmer inserted that assigns an amplicon to its original sample/PCR reaction
Artefactual sequences	Any sequence or MOTU originating from primer dimers, or chimeras from two or multiple original templates. These sequences usually largely differ from any known sequence
Failed PCRs	Any PCR product yielding a low amount of sequences or an irreproducible signal

2 | DATA STRUCTURE, IMPORT/EXPORT AND MANIPULATION

metabaR performs the analysis of DNA metabarcoding data while handling the multiple information it contains. The central object of the package is a metabarlist, an R list composed of four interconnected tables (Figure 1): (a) reads, a table that stores the read abundance of MOTUs in each PCR product, (b) motus, a table that stores any information relative to each MOTU in the dataset (e.g. taxonomic information), (iii) pcrs, a table that stores any information relative to each PCR reaction (e.g. if it is a sample or an experimental control, what are the primer used, etc.) and (d) samples, a table that contains any metadata relative to the biological sample from which the PCR reaction was obtained (e.g. geographic coordinates, abiotic parameters, etc.). A metabarlist can be generated from outputs of various bioinformatic pipelines such as **vsearch**, **qiime** or **OBITools** through a set of data-import functions. These include two generic functions, tabfiles_to_metabarlist and biomfiles_to_metabarlist that import files in csv or BIOM (Biological Observation Matrix) format, and the more specific obifiles_to_metabarlist function adapted for **OBITools** outputs. We also provide

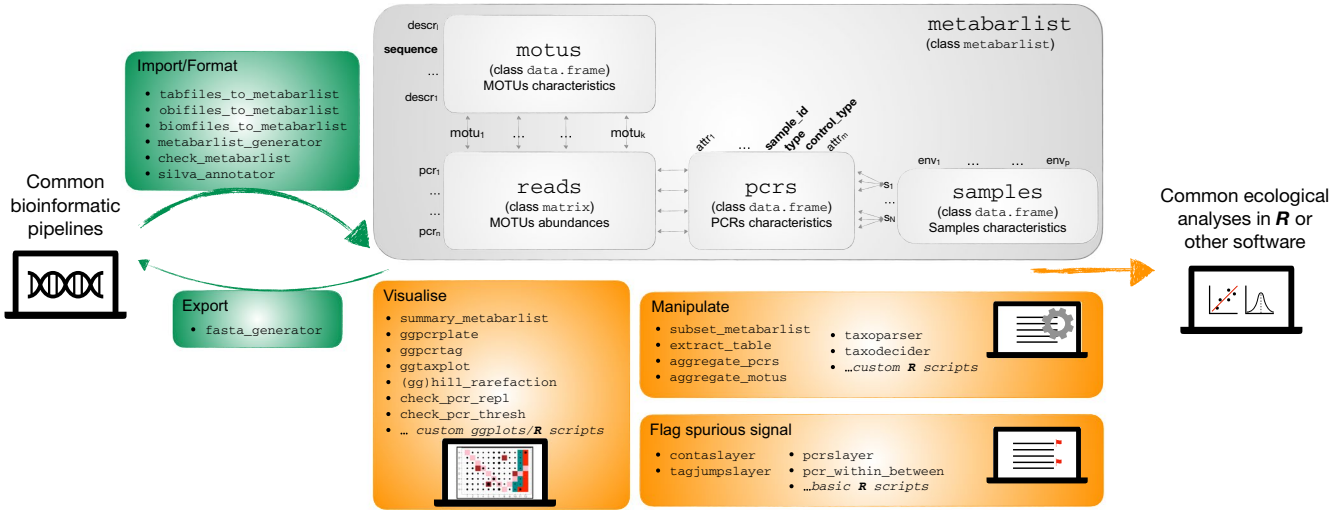

FIGURE 1 Overview of the **metabaR** package data structure (grey box) and functions (green and orange boxes). Mandatory fields in each table of the metabarlist are indicated in bold. More details are available in the help page of check_metabarlist

the metabarlist_generator function, which facilitates metabarlist building directly from objects in the R environment. The package also provides a tool, silva_annotator, which imports **SILVAngs** (Quast et al., 2012) taxonomic output files to complement the motus table for more specific applications.

All these import tools use the function check_metabarlist, which verifies whether the imported or created metabarlist fulfills a set of mandatory properties for the package to work. The function returns a warning message with guidance to the user when the format is incorrect.

Any table of the metabarlist can be amended easily with R commands non-specific to the **metabaR** package. For example, the reads matrix can be transformed into relative abundance data with rowSums or the decostand function of the **vegan** package. Likewise, any column can be added to the data frames pcrs, motus or samples by using basic R commands.

The metabarlist object can be manipulated for different purposes. It can be subsetted with subset_metabarlist with customisable criteria relating to any table of the metabarlist. The user can also aggregate read counts based on MOTU criteria with aggregate_motus, such as for obtaining community data at higher taxonomic ranks than the MOTU level. Similarly, aggregate_pcrs can be used to aggregate read counts based on PCR-related criteria, typically to aggregate technical PCR replicates at the sample level.

We also include two functions to facilitate the customisation of taxonomic information. The first, taxoparser, is a simple tool that parses full or partial taxonomic paths generated during upstream bioinformatic processing. The function taxodecider enables users to process taxonomic assignment for the same MOTU from multiple databases. For example, building a custom reference database is often recommended, since including species from the regional species pool increases the reliability of taxonomic assignments (Taberlet et al., 2018). However, these databases are often incomplete and it is common to run annotation tools with commonly available reference databases such as EMBL (https://www.ebi.ac.uk/ena/browser/home) or BOLD (Ratnasingham & Hebert, 2007). The taxodecider function allows users to merge different annotations based on assignment scores and by priotirising assignments from the user's preferred reference database, usually the one with the most reliable taxonomic and sequence information.

Finally, **metabaR** has different export tools. First, fasta_generator exports sequences in the fasta format where the user is free to add any information from the metabarlist to the sequence header. This function can be of use when during the **metabaR** workflow, it becomes apparent that specific bioinformatic procedures require retuning. Since the metabarlist is a simple R list, it can be directly exported with the R base functions saveRDS while individual table can be extracted using extract_table and R's write.table function.

3 | EXAMPLE DATASET

The package contains a dataset, named soil_euk, which is a typical output of a DNA metabarcoding experiment. It is used in the package help and the vignette to illustrate the functions of **metabaR**. soil_euk is a metabarlist and contains the abundance of 12,647 MOTUs obtained from 384 PCRs, corresponding to a total of 64 biological samples. The dataset also includes different information on MOTUs, PCRs and samples. The dataset was generated from soil and litter samples collected in two tropical forests in French Guiana, from which a variable region of the 18S rRNA gene was amplified by PCR and sequenced on an Illumina sequencing platform. The data were then processed with the **OBITools** bioinformatic pipeline. Detailed information regarding the generation of this dataset are available on the soil_euk help page.

4 | VISUALISATION

Appropriate visual representation of DNA metabarcoding data greatly facilitates the assessment of data quality and of the curation

process. In addition to representing dataset characteristics such as sample sequencing depth or richness in MOTUs using standard boxplots and histograms, we developed two functions, ggpcrplate and ggpcrtag, to represent dataset characteristics in their experimental context, that is, the PCR plate. Their input consists of a metabarlist and a function pre-encoded in **metabaR** or designed by the user to be applied to the input metabarlist so as to enable the plotting of numerous dataset characteristics. Such visualisation can enable the identification of potential experimental problems, such as pipetting or tag/primer issues as exemplified in Figure 2.

The taxonomic composition of DNA metabarcoding data is also often difficult to represent because taxonomic assignments are seldom available at a uniform taxonomic level. This problem usually results from either the incompleteness of reference databases, or as a result of the inherent variation of DNA markers in taxonomic/phylogenetic resolution across lineages. To facilitate the visualisation

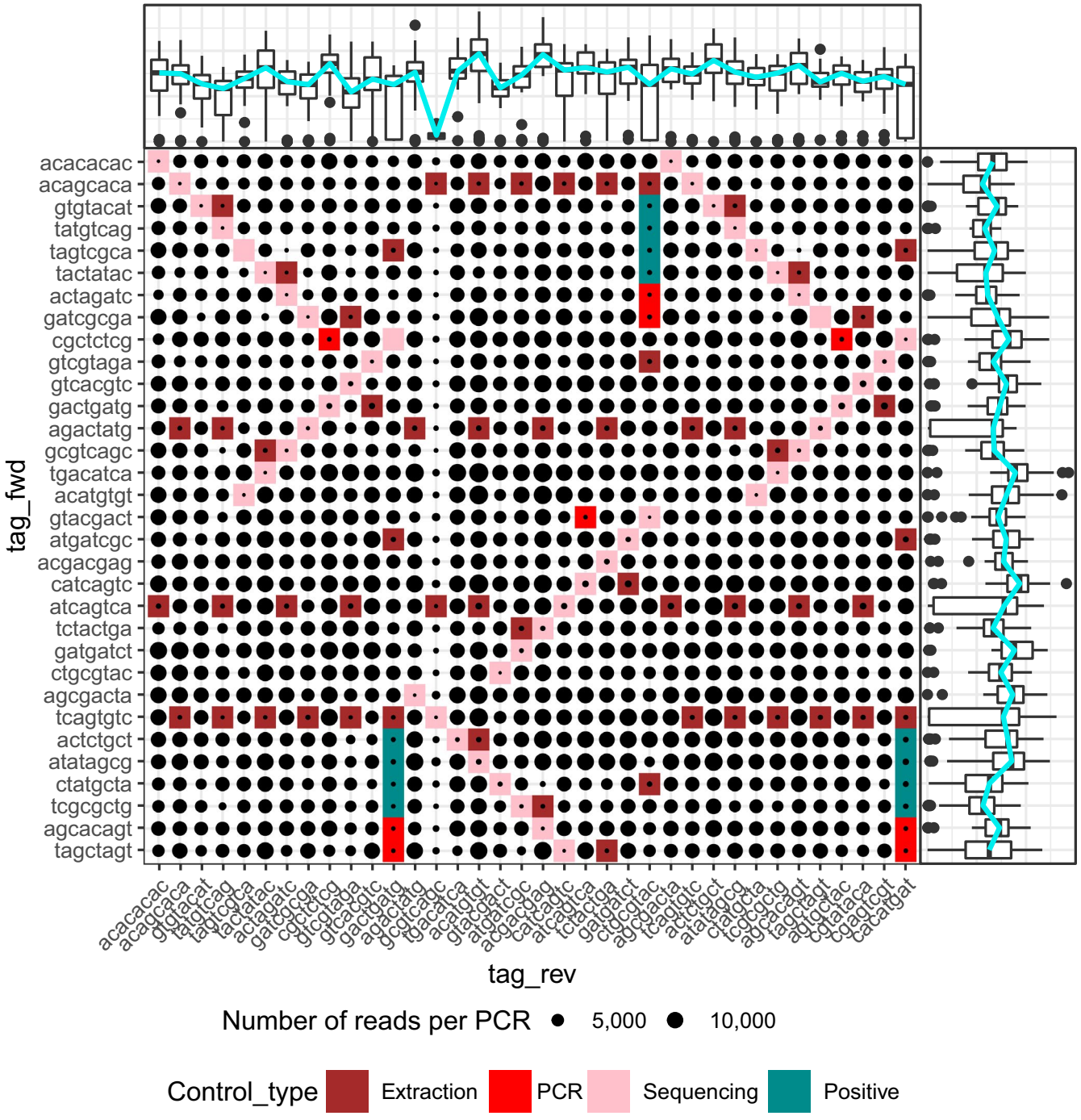

FIGURE 2 Example of an output from ggpcrtag with a problematic DNA metabarcoding dataset exhibiting low amounts of reads for all PCR reactions conducted with the reverse primer including the tag 'gcgtcagc'. Upper and right boxplots show the total value of the variable of interest (here number of reads) across all PCRs using a primer with the same tag. The figure also shows what experimental design was used for this particular dataset (controls type and locations in a 4 × 3 PCR plate set up)

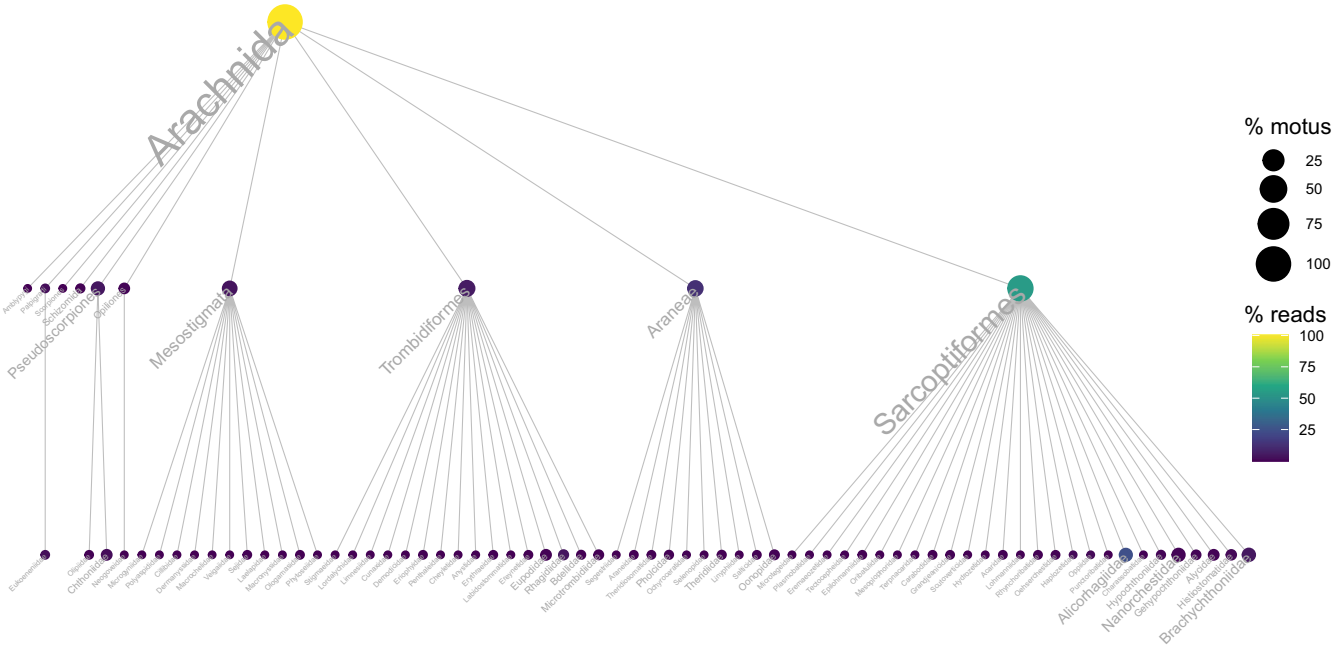

FIGURE 3 Example of an output from ggtaxplot using the soil_euk dataset, focusing on Arachnida MOTUs. Each node corresponds to a taxon, node size to the proportion of MOTUs and node colour to the proportion of read counts

of the sample or experiment's community composition in this context, we developed the function ggtaxplot, dependent on the **igraph** R package (Csardi & Nepusz, 2006). This function plots taxonomic trees where each node corresponds to a taxon, with node size and colour corresponding to the taxon number of reads and diversity in MOTUs (Figure 3).

Finally, rarefaction curves are routinely used with DNA metabarcoding data to assess whether the MOTU diversity of each PCR reaction or sample is appropriately covered by sequencing depth. The hill_rarefation function and its plotting complement gghill_rarefaction build rarefaction curves using three indices included in the Hill numbers framework (Chao et al., 2014; Hill, 1973), which have been shown to provide good estimates of alpha diversity for DNA metabarcoding data (Alberdi & Gilbert, 2019; Calderón-Sanou et al., 2019). More specifically, these functions estimate MOTU richness, the exponential of the Shannon index, the inverse of the Simpson index, as well as Good's coverage index (Good, 1953) at different sequencing depths chosen by the user.

All visualisation tools used in **metabaR** are based on ggplot2 and **cowplot** R packages (Wickham, 2016; Wilke, 2019) for greater flexibility.

5 | DATA CURATION TOOLS

Numerous bioinformatic tools allow the curation of DNA metabarcoding data to account for PCR and sequencing errors. By contrast, only a few (e.g. **LULU**, Frølev et al., 2017) deal with other types of artefactual MOTUs (Table 1). **metabaR** includes three functions which each target a particular type of noise data. To allow users to evaluate the downstream impacts of removing identified noise data, two of these only flag potential spurious objects in the output rather than removing them directly.

The tagjumpslayer function targets artefacts called 'tag-jumps', 'tag-switches' or 'cross-talks' (Table 1; Edgar, 2017; Esling et al., 2015; Schnell et al., 2015), which generate a noise similar to cross-sample contaminations but at the scale of the whole sequencing library, hence homogenising the data. The tagjumpslayer function aims to reduce this noise by removing a MOTU in a given PCR product when its relative abundance over the entire dataset is below a given threshold. This threshold can be empirically chosen by testing the effect of varying curation thresholds on the MOTU and read counts in the dataset in general and, when available, in the sequencing negative controls (i.e. unused tag or library index combinations) in particular.

The effect of these tag-jumps can complicate the detection of external contaminants, such as those occurring in laboratory reagents (Salter et al., 2014). An approach which only consists of the detection of MOTUs present in experimental negative controls would ignore tag-jumps and can result in the removal of the most abundant genuine MOTUs from the dataset. However, in negative controls, contaminants should be preferentially amplified in the absence of competing DNA, which is unlikely to be the case in biological samples. The contaslayer function relies on this assumption and detects MOTUs whose relative abundance across the whole dataset is highest in negative controls.

Finally, the pcrslayer function aims to identify potentially failed PCR reactions by comparing the dissimilarities in MOTU composition within a biological sample (i.e. between PCR replicates, hereafter *dw*) versus between biological samples (hereafter *db*). It relies on the assumption that PCR replicates from a same biological sample should be more similar than two different biological samples

(*dw* < *db*). A PCR replicate having *dw* above a given dissimilarity threshold, defined automatically by the function based on the distribution of *dw* and *db*, is considered to be an outlier. The function can be run with any dissimilarity index. Several functions are provided along with pcrslayer, such as check_pcr_repl, which draws an ordination of PCR replicates, as well as pcr_within_between and check_pcr_thresh which compute and represent the distribution of *dw* and *db*.

In addition to the identification and flagging of artefacts provided by these functions, other issues such as PCRs with shallow sequencing depths, MOTUs that are not targeted by the primers or those with too low taxonomic assignment scores, can also be flagged with R base functions (detailed in the vignette accompanying package).

6 | CONCLUSIONS

The **metabaR** package provides much needed tools to evaluate the quality of metabarcoding data and curate commonly overlooked artefacts. It is currently most adapted to users that have already basic R scripting knowledge. We also provide a vignette along the package that constitutes a good starting point for new users to build their own quality assessment and filtering of DNA metabarcoding data: it highlights all recommended steps and possible uses of experimental controls to clean the data. The **metabaR** package and its vignette will contribute to improving data quality standards in the field, ease the analysis of DNA metabarcoding data and will therefore help to broaden the use of environmental DNA-based analyses of biodiversity.

ACKNOWLEDGEMENTS

We are deeply indebted to Eric Coissac for stimulating discussions that led to the development of this package, and are also grateful to Jerôme Chave and Wilfried Thuiller for supporting this work and for their comments on an earlier version of this note. We thank Douglas Yu and Florian Leese for their assessment and constructive comments on this work, Pierre Taberlet and Heidy Schimann for providing data, as well as Irene Calderón-Sanou, Camille Martinez-Almoyna, Jerôme Murienne and Renato A. Ferreira de Lima for practical discussions on—and/or testing of—earlier versions of the package. We also thank Chris Bowler for providing informatics equipment to ASB. The work was funded by the METABAR (ANR-11-BSV7-0020) and GlobNets (ANR-16-CE02-0009) projects, and has benefitted from 'Investissement d'Avenir' grants managed by Agence Nationale de la Recherche (CEBA: ANR-10-LABX-25-01; TULIP: ANR-10-LABX-0041; OSUG@2020: ANR-10-LABX-56).

AUTHORS' CONTRIBUTIONS

L.Z., F.B. and C.L. conceived and wrote the package; A.-S.B. and C.M. contributed to the writing of functions and A.-S.B. and J.D. to the writing of the documentation and vignette; L.Z. wrote the manuscript with inputs from all co-authors.

DATA AVAILABILITY STATEMENT

The **metabaR** package is available on GitHub at https://github.com/metabaRfactory/metabaR. Its first version (v1.0.0, Zinger et al., 2021a) is available on Zenodo at https://doi.org/10.5281/zenodo.4419791. We also provide a full description of the package functions, as well as a step by step tutorial (R vignette) describing the package basic use at https://metabarfactory.github.io/metabaR. The example dataset is provided in .biom, and .txt formats within a companion package available at https://github.com/metabaRfactory/metabaR_external_data (first version available at https://doi.org/10.5281/zenodo.4419778 (Zinger et al., 2021b)). We also provide other example datasets for more tests.

ORCID

Lucie Zinger https://orcid.org/0000-0002-3400-5825
Anne-Sophie Benoiston https://orcid.org/0000-0001-9446-5703
Julian Donald https://orcid.org/0000-0001-6900-3777
Céline Mercier https://orcid.org/0000-0002-4782-1530
Frédéric Boyer https://orcid.org/0000-0003-0021-9590

REFERENCES

Alberdi, A., & Gilbert, M. T. P. (2019). A guide to the application of Hill numbers to DNA-based diversity analyses. *Molecular Ecology Resources*, *19*(4), 804–817. https://doi.org/10.1111/1755-0998.13014

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. https://doi.org/10.1111/1755-0998.12428

Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., & Thuiller, W. (2019). From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices? *Journal of Biogeography*, *47*(1), 193–206. https://doi.org/10.1111/jbi.13681

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*, 581–583. https://doi.org/10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*, 335–336. https://doi.org/10.1038/nmeth.f.303

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

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (in press). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, https://doi.org/10.1111/mec.15472

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, *1695*, 1–9.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA

metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*, 5872–5895. https://doi.org/10.1111/mec.14350

Edgar, C. (2017). UNCROSS: Filtering of high-frequency cross-talk in 16S amplicon reads. *bioRxiv*, https://doi.org/10.1101/088666

Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, *43*(5), 2513–2524. https://doi.org/10.1093/nar/gkv107

Estaki, M., Jiang, L., Bokulich, N. A., McDonald, D., González, A., Kosciolek, T., Martino, C., Zhu, Q., Birmingham, A., Vázquez-Baeza, Y., Dillon, M. R., Bolyen, E., Caporaso, J. G., & Knight, R. (2020). QIIME 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data. *Current Protocols Bioinformatics*, *70*, e100. https://doi.org/10.1002/cpbi.100

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communication*, *8*, 1188. https://doi.org/10.1038/s41467-017-01312-x

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3–4), 237–264. https://doi.org/10.1093/biomet/40.3-4.237

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*(2), 427–432. https://doi.org/10.2307/1934352

McMurdie, P. J., & Holmes, S. P. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2019). *vegan: Community ecology package*. R package version 2.5-6. Retrieved from https://CRAN.R-project.org/package=vegan

Pavoine, S. (2020). adiv: An R package to analyse biodiversity in ecology. *Methods in Ecology and Evolution*, *11*(9), 1106–1112. https://doi.org/10.1111/2041-210x.13430

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

R Core Team. (2020). *R: A language and environment for statistical computing*. R foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*(1), 87. https://doi.org/10.1186/s12915-014-0087-z

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated – Reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303. https://doi.org/10.1111/1755-0998.12402

Sommeria-Klein, G., Zinger, L., Taberlet, P., Coissac, E., & Chave, J. (2016). Inferring neutral biodiversity parameters using environmental DNA data sets. *Scientific Report*, *6*, 35644. https://doi.org/10.1038/srep35644

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

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wilke, C. O. (2019). *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. R package version 1.0.0. Retrieved from https://CRAN.R-project.org/package=cowplot

Zepeda Mendoza, M. L., Bohmann, K., Carmona Baez, A., & Gilbert, M. T. P. (2016). DAMe: A toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. *BMC Research Notes*, *9*, 255.

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding-Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*, 1857–1862. https://doi.org/10.1111/mec.15060

Zinger, L., Lionnet, C., Benoiston, A.-S., & Boyer, F. (2021). metabaRfactory/metabaR: metabaR first release (Version v1.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.4419791

Zinger, L., Mercier, C., & Boyer, F. (2021). metabaRfactory/metabaR_external_data: metabaR_external_data first release (Version v1.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.4419778

How to cite this article: Zinger L, Lionnet C, Benoiston A-S, Donald J, Mercier C, Boyer F. metabaR: An R package for the evaluation and improvement of DNA metabarcoding data quality. *Methods Ecol Evol*. 2021;12:586–592. https://doi.org/10.1111/2041-210X.13552