



HAL
open science

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

Lucie Zinger, Clément Lionnet, Anne-Sophie Benoiston, Julian Donald, Céline Mercier, Frédéric Boyer

► To cite this version:

Lucie Zinger, Clément Lionnet, Anne-Sophie Benoiston, Julian Donald, Céline Mercier, et al.. metabaR: An r package for the evaluation and improvement of DNA metabarcoding data quality. *Methods in Ecology and Evolution*, 2021, 12 (4), pp.586–592. 10.1111/2041-210X.13552 . hal-04717372

HAL Id: hal-04717372

<https://cnrs.hal.science/hal-04717372v1>

Submitted on 2 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 metabarR : an R package for the evaluation and
2 improvement of DNA metabarcoding data quality

3 Lucie Zinger ^{*1}, Clément Lionnet², Anne-Sophie Benoiston¹, Julian Donald^{3,4}, Céline
4 Mercier², and Frédéric Boyer²

5 ¹*Institut de Biologie de l'ENS (IBENS), Département de biologie, École Normale
6 Supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France*

7 ²*Univ. Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, LECA, Laboratoire d'Écologie
8 Alpine, F-38000 Grenoble, France*

9 ³*Evolution et Diversité Biologique (EDB UMR5174), Université Toulouse 3 Paul Sabatier,
10 CNRS, IRD - Toulouse, France*

11 ⁴*Centre for Ecology and Conservation, University of Exeter, Penryn TR10 9FE, UK*

*Corresponding author : lucie@zinger.fr

12 Abstract

- 13 1. DNA metabarcoding is becoming the tool of choice for biodiversity studies across taxa
14 and large-scale environmental gradients. Yet, the artefacts present in metabarcoding
15 datasets often preclude a proper interpretation of ecological patterns. Bioinformatic
16 pipelines removing experimental noise have been designed to address this issue. How-
17 ever, these often only partially target produced artefacts, or are marker specific. In
18 addition, assessments of data curation quality and the appropriateness of filtering
19 thresholds are seldom available in existing pipelines, partly due to the lack of appro-
20 priate visualisation tools.
- 21 2. Here, we present **metabaR**, an R package that provides a comprehensive suite of
22 tools to effectively curate DNA metabarcoding data after basic bioinformatic analyses.
23 In particular, **metabaR** uses experimental negative or positive controls to identify
24 different types of artefactual sequences, i.e. reagent contaminants and tag-jumps. It
25 also flags potentially dysfunctional PCRs based on PCR replicate similarities when
26 those are available. Finally, **metabaR** provides tools to visualise DNA metabarcoding
27 data characteristics in their experimental context as well as their distribution, and
28 facilitate assessment of the appropriateness of data curation filtering thresholds.
- 29 3. **metabaR** is applicable to any DNA metabarcoding experimental design but is most
30 powerful when the design includes experimental controls and replicates. More gener-
31 ally, the simplicity and flexibility of the package makes it applicable any DNA marker,
32 and data generated with any sequencing platform, and pre-analysed with any bioin-

33 formatic pipeline. Its outputs are easily usable for downstream analyses with any
34 ecological R package.

35 4. **metabaR** complements existing bioinformatics pipelines by providing scientists with
36 a variety of functions with customisable methods that will allow the user to effectively
37 clean DNA metabarcoding data and avoid serious misinterpretations. It thus offers a
38 promising platform for automatised data quality assessments of DNA metabarcoding
39 data for environmental research and biomonitoring.

40 **Keywords**

41 data mining, environmental DNA, high-throughput, sequencing, data curation, contamina-
42 tions, tag-jumps

43 1 | Introduction

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

58 While the aforementioned bioinformatic tools have been heavily used for analysing the
59 ever-expanding number of DNA metabarcoding data, they also present a certain num-
60 ber of limitations. DNA metabarcoding generates numerous experimental biases besides
61 PCR/sequencing errors and chimeras, which range from field or laboratory contaminations
62 through to tag-jumps (Table 1; reviewed in Taberlet et al., 2018; Zinger et al., 2019). The

63 treatment of these artefact is often missing in DNA metabarcoding studies, even though
64 they can substantially affect ecological inferences ([Sommeria-Klein et al., 2016](#); [Frøslev et
65 al., 2017](#); [Calderón-Sanou et al., 2019](#)). Such artefacts can only be flagged and corrected
66 by including experimental controls and experimental replicates throughout the data pro-
67 duction process. However, existing bioinformatic pipelines only deal with PCR/sequencing
68 errors, and do not make use of experimental controls to filter out potential contaminants
69 or artefacts. Second, these bioinformatic pipelines often lack tools to monitor and evalu-
70 ate the bioinformatic data filtering process. As a result, it is often difficult to tune data
71 filtering parameters, and users are therefore led to using default settings even when these
72 are suboptimal. Finally, DNA metabarcoding data are in essence multidimensional, as they
73 encompass MOTUs, PCR product, and biological sample information. This multi-fold in-
74 formation, often stored in separate tables, is not easily handled by most R packages for data
75 analyses (but see e.g. **phyloseq**). As such, we currently lack effective tools to transition
76 from bioinformatic pipelines to ecological R packages.

77 To bridge this gap, we developed **metabaR**, an R package that enables the post-processing
78 and filtering of DNA metabarcoding data already processed through bioinformatic pipelines
79 so as to improve downstream ecological inferences. It is designed to take advantage of neg-
80 ative controls, positive controls and PCR replicates when available to efficiently flag and
81 remove artefactual MOTUs or dysfunctional PCRs. It is implemented in the R statistical
82 programming environment ([R Core Team, 2020](#)) which provides flexible analytical tools cou-
83 pled with powerful graphical capabilities. **metabaR** uses these properties to provide highly

Experimental bias	Description
PCR/sequencing errors	Any MOTU resulting from base misincorporation by DNA polymerase during PCR amplification or sequencing, or base miscalling during sequencing.
Contaminants	Any MOTU external to the biological sample, typically from lab reagents. Such contamination can occur at all stages of the data production, i.e. field work, DNA extraction, PCR amplification, and library preparation.
Tag-jumps	MOTU of which presence is genuine at the sampling sites, but erroneous in a given sample/PCR product due to a switch of so called "tag" or "library index", i.e. a nucleotidic kmer allowing to reassign the amplicon to the amplicon to its sample/PCR reaction of origin.
Degraded sequences	Any sequence or MOTU such as primer dimers, or chimeras from two or multiple parents. Usually largely differ from any known sequence.
Dysfunctional PCRs	Any PCR product yielding a low amount of sequences or an irreproducible signal.

Table 1 – Overview of DNA metabarcoding experimental artefacts

84 customisable functions, as well as effective visualisation of DNA metabarcoding data in their
85 experimental context. Hence, it is of direct use for any practitioner of DNA metabarcoding
86 techniques with basic skills in R programming.

87 **2 | Data structure, import/export, and manipulation**

88 **metabaR** performs the analysis of DNA metabarcoding data by taking into account its
89 multidimensional nature. The central object of the package is a `metabarlist`, an R list
90 composed of 4 interconnected tables (Fig. 1): (i) `reads`, a table that stores the read abun-
91 dance of MOTUs in each PCR product, (ii) `motus`, a table which stores any information
92 relative to each MOTU in the dataset (e.g. taxonomic information), (iii) `pcrs`, a table
93 which stores any information relative to each PCR reaction (e.g. if it is a sample or an

94 experimental control, what are the primer used, etc.), and (iv) `samples`, a table which
95 contains any metadata relative to the biological sample from which the PCR reaction was
96 obtained (e.g. geographic coordinates, abiotic parameters, etc.). This object can be gener-
97 ated from outputs of various bioinformatic pipelines such as `vsearch`, `qiime` or `OBITools`
98 through a set of data-import functions. These include two generic functions, `tabfiles_-`
99 `to_metabarlist` and `biomfiles_to_metabarlist` that import files in csv or BIOM (Biolog-
100 ical Observation Matrix) format, and the more specific `obifiles_to_metabarlist` function
101 adapted for `OBITools` outputs. We also provide the `metabarlist_generator` function,
102 which facilitates `metabarlist` building directly from objects in the R environment. The
103 package also provide a tool, `silva_annotator`, which imports `SILVAngs` ([Quast et al.,](#)
104 [2012](#)) taxonomic output files to complement the `motus` table for more specific applications.

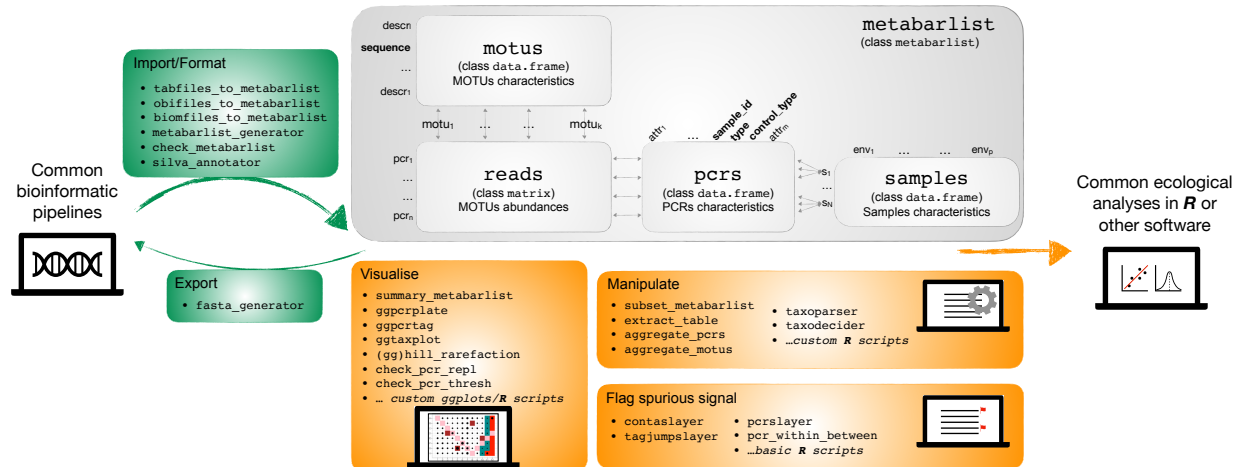


Figure 1 – Overview of the **metabarR** package data structure (grey box) and functions (green and orange boxes). Mandatory fields in each table of the **metabarlist** are indicated in bold. More details are available in the help page of `check_metabarlist`.

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

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

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

120 We also include two functions to facilitate the customisation of taxonomic information. The
121 first, `taxoparser` is a simple tool that parses full or partial taxonomic paths generated dur-
122 ing upstream bioinformatic processing. The function `taxodecider` enables users to process
123 taxonomic assignment for the same MOTU from multiple databases. For example, building
124 a custom reference database is often recommended, since including species from the regional
125 species pool increases the reliability of taxonomic assignments (Taberlet et al., 2018). How-
126 ever, these databases are often incomplete and it is common to run in parallel annotation
127 tools with more generalist reference databases such as EMBL (<https://www.embl.org/>).
128 The `taxodecider` function allows users to merge different annotations based on assignment
129 scores and by giving priority to assignments from the user's preferred reference database,
130 usually the one for which taxonomic and sequence information is the most reliable.

131 Finally `metabaR` has different export tools. First, `fasta_generator` exports sequences
132 in the fasta format where the user is free to add any information from the `metabarlist`
133 to the sequence header. This function can be of use when following the implementation
134 of `metabaR` functions, it becomes apparent that specific bioinformatic data curation pro-

135 cedures require retuning. The package does not provide other export functions, since the
136 `metabarlist` is a simple R `list` that can be directly exported with the R base function
137 `saveRDS`. Alternatively, any table may be extracted from the `metabarlist` with `extract_-`
138 `table`, before R's `write.table` function is used for export.

139 **3 | Example dataset**

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

150 **4 | Visualisation**

151 Appropriately representing DNA metabarcoding data visually is a prerequisite to assessing
152 the quality of the data or of the curation process. Such assessments require going beyond
153 representing dataset characteristics such as sample sequencing depth or richness in MOTUs

154 using standard boxplots and histograms. Here, we developed two functions, `ggpcrplate` and
155 `ggpcrtag`, to represent data set characteristics in their experimental context, i.e. the PCR
156 plate. Their input consists of a `metabarlist` and a function pre-encoded in `metabar` or
157 designed by the user to be applied to the input `metabarlist` so as to enable the plotting of
158 numerous dataset characteristics. Such visualisation can enable the identification of potential
159 experimental problems, such as pipetting or tag/primer issues as exemplified in Fig. 2.

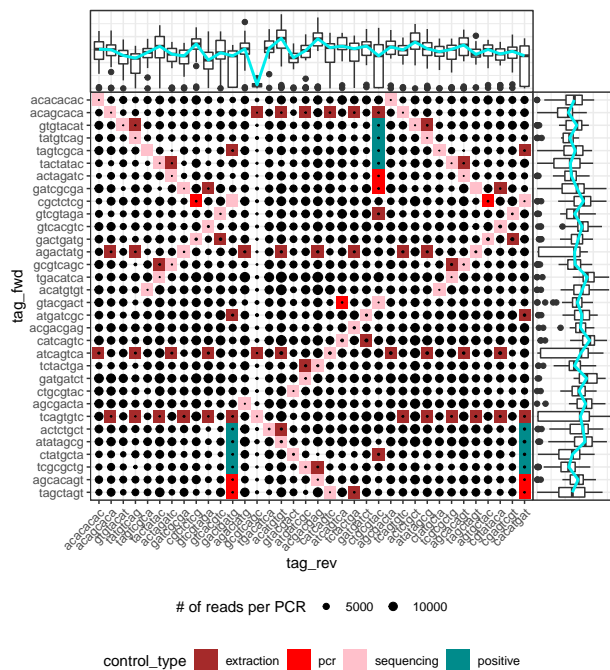


Figure 2 – Example of an output from `ggpcrtag` with a problematic DNA metabarcoding dataset exhibiting low amounts of reads for all PCR reactions conducted with the reverse primer including the tag “gcgtcagc”. Upper and right boxplot show the total value of the variable of interest (here number of reads) 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 x 3 PCR plate set up).

160 The taxonomic composition of DNA metabarcoding data is also often difficult to represent
161 because taxonomic assignments are seldom available at a uniform taxonomic level. This
162 problem usually results from either the incompleteness of reference databases, or as a result of
163 the inherent variation of DNA markers in taxonomic/phylogenetic resolution across lineages.
164 To facilitate the visualisation of the sample or experiment's community composition in this
165 context, we developed the function `ggtaxplot`, dependant on the **igraph** R package ([Csardi](#)
166 [& Nepusz, 2006](#)) . This function plots taxonomic trees where each node corresponds to a
167 taxon, with node size and colour corresponding to the taxon number of reads and diversity
168 in MOTUs (Fig. 3).

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

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

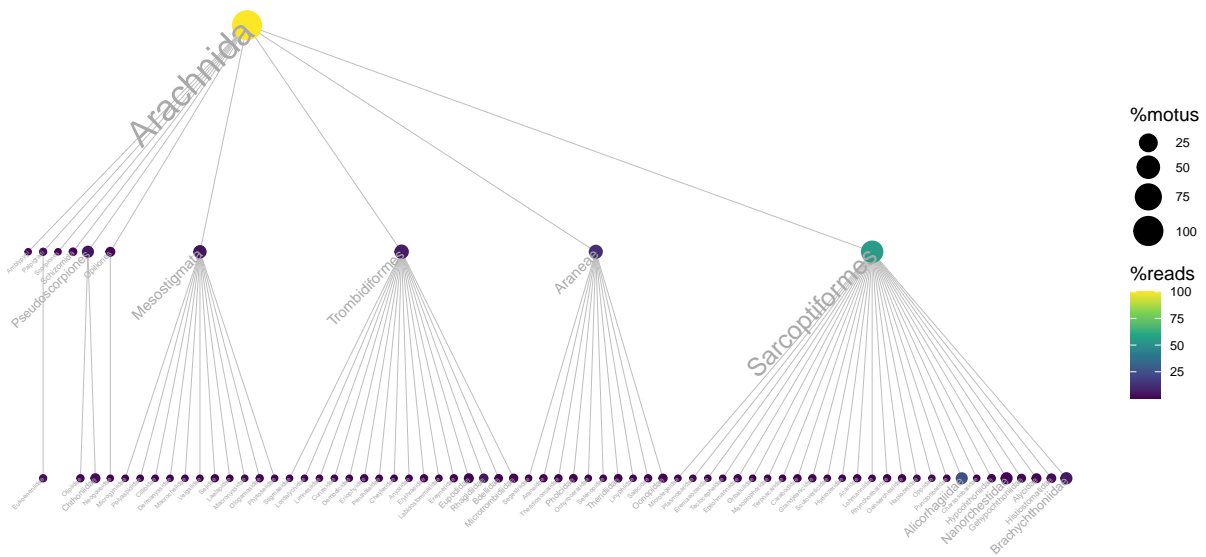


Figure 3 – Example of an output from `ggtaxplot` using the `soil_euk` dataset, focusing on Arachnida MOTUs. Each node corresponds to a taxon, node size to the proportion of MOTUs, and node color to the proportion of read counts.

180 4 | Data curation tools

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

187 The `tagjumpslayer` function targets artefacts called “tag-jumps”, “tag-switches” or “cross-

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

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

204 Finally, the `pcrs` function aims to identify potential failed PCR reactions by comparing
205 the dissimilarities in MOTU composition within a biological sample (i.e. between PCR
206 replicates, hereafter dw) vs. between biological samples (hereafter db). It relies on the
207 assumption that PCR replicates from a same biological sample should be more similar than
208 two different biological samples ($dw < db$). A PCR replicate having dw above a given

209 dissimilarity threshold, defined automatically by the function based on the distribution of
210 dw and db , is considered to be an outlier. The function can be run with any dissimilarity
211 index. Several functions are provided along with `pcrslayer`, such as `check_pcr_repl`, which
212 draw an ordination of PCR replicates; as well as `pcr_within_between` and `check_pcr_thresh`
213 which compute and represent the distribution of dw and db .

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

218 **5 | Conclusions**

219 The **metabaR** package provides a much needed tool at the interface between bioinformatic
220 pipelines and ecological analyses to evaluate the quality of data and of the curation process,
221 prior to offering further curation of commonly overlooked artefacts. We also provide a
222 vignette along the package that constitutes for new users a good starting point to build
223 their own quality assessment and filtering of DNA metabarcoding data: it highlights all
224 the recommended steps and possible uses of experimental controls to clean the data. The
225 **metabaR** package and its vignette will contribute in improving data quality standards in
226 the field, ease the analysis of DNA metabarcoding data and will therefore help to broaden
227 the use of eDNA-based analyses of biodiversity.

228 **Acknowledgements**

229 We are deeply indebted to Eric Coissac for stimulating discussions that led to the devel-
230 opment of this package, and are also grateful to Jérôme Chave and Wilfried Thuiller for
231 supporting this work. We thank Pierre Taberlet and Heidy Schimann for providing data,
232 as well as Irene Calderón-Sanou, Camille Martinez-Almoyna, Jérôme Murienne and Re-
233 nato A. Ferreira de Lima for practical discussions on - and/or testing of - earlier versions
234 of the package. We also thank Chris Bowler for providing informatics equipment to ASB.
235 The work was funded by the METABAR (ANR-11-BSV7-0020) and GlobNets (ANR-16-
236 CE02-0009) projects, and has benefitted from “Investissement d’Avenir” grants managed by
237 Agence Nationale de la Recherche (CEBA: ANR-10-LABX-25-01; TULIP: ANR-10-LABX-
238 0041; OSUG@2020: ANR-10-LABX-56).

239 **Authors contribution**

240 LZ, FB, and CL conceived and wrote the package. ASB and CM contributed to the writing
241 of functions and ASB and JD to the writing of the documentation and vignette. LZ wrote
242 the manuscript with inputs from all co-authors.

243 **Data availability statement**

244 The **metabaR** package is available on GitHub at <https://github.com/metabaRfactory/>
245 **metabaR**. We also provide a full description of the package functions, as well as a step by

246 step tutorial (R vignette) describing the package basic use at [https://metabarfactory.](https://metabarfactory.github.io/metabaR)
247 github.io/metabaR. The example dataset is provided within the package in .rds, .biom,
248 and .txt formats.

249 **References**

250 Alberdi, A., & Gilbert, M. T. P. (2019). A guide to the application of Hill numbers to DNA-
251 based diversity analyses. *Molecular Ecology Resources* 19(4) 804-817. doi: 10.1111/1755-
252 0998.13014.

253 Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. and Coissac, E. (2016). OBITools:
254 a UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*
255 16, 176-182. doi: 10.1111/1755-0998.12428.

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

259 Callahan, B.J., McMurdie P.J., Rosen M.J., Han, A.W., Johnson, A.J.A, Holmes, S.P.
260 (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature*
261 *Methods* 13, 581-583. doi: 10.1038/nmeth.3869

262 Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D, Costello, E.K.,
263 ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing
264 data. *Nature Methods* 7, 335-336. doi: 10.1038/nmeth.f.303.

- 265 Chao, A., Chiu, C.-H., Jost, L. (2014). Unifying Species Diversity Phylogenetic Diversity,
266 Functional Diversity, and Related Similarity and Differentiation Measures Through Hill
267 Numbers. *Annual Review of Ecology, Evolution, and Systematics* 45(1), 297-324. doi:
268 10.1146/annurev-ecolsys-120213-091540.
- 269 Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D.A.,
270 Bouchez, A., Chariton, A., ... Lanzén, A. (in press). Ecosystems monitoring pow-
271 ered by environmental genomics: A review of current strategies with an implementation
272 roadmap. *Molecular Ecology*. doi: 10.1111/mec.15472.
- 273 Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research.
274 *InterJournal Complex Systems*.
- 275 Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F.,
276 ... Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how
277 we survey animal and plant communities. *Molecular Ecology* 26, 5872-5895. doi:
278 10.1111/mec.14350.
- 279 Edgar, C. (2017). UNCROSS: Filtering of high-frequency cross-talk in 16S amplicon reads.
280 *bioRxiv*. doi:10.1101/088666.
- 281 Esling, P., Lejzerowicz, F., Pawlowski, J. (2015). Accurate multiplexing and filtering for
282 high-throughput amplicon-sequencing. *Nucleic Acids Research* 43(5), 2513-2524. doi:
283 10.1093/nar/gkv107.

- 284 Estaki, M., Jiang, L., Bokulich, N. A., McDonald, D., González, A., Kosciulek, T., ... Knight,
285 R. (2020). QIIME 2 enables comprehensive end-to-end analysis of diverse microbiome
286 data and comparative studies with publicly available data. *Current Protocols Bioinform-*
287 *matics* 70, e100. doi: 10.1002/cpbi.100.
- 288 Frøslev, T. G., Kjølner, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., Hansen,
289 A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable
290 biodiversity estimates. *Nature Communication* 8, 1188. doi: 10.1038/s41467-017-01312-
291 x.
- 292 Good, I. J. (1953). The population frequencies of species and the estimation of population
293 parameters. *Biometrika* 40(3-4), 237-264. doi: 10.1093/biomet/40.3-4.237.
- 294 Hill, M.O. (1973). Diversity and evenness: a unifying notation and its consequences, *Ecology*
295 54(2) 427-432. doi: 10.2307/1934352.
- 296 McMurdie, P.J., Holmes, S.P. (2013). phyloseq: An R Package for Reproducible Interactive
297 Analysis and Graphics of Microbiome Census Data. *PLoS ONE* 8(4), e61217. doi:
298 10.1371/journal.pone.0061217.
- 299 Oksanen, J., Blanchet F. G., Friendly, M., Kindt, R. Legendre, P., McGlinn, D., ... Wagner,
300 H. (2019). *vegan: Community Ecology Package*. R package version 2.5-6. Retrieved
301 from [https://CRAN.R-proje ct.org/package=vegan](https://CRAN.R-project.org/package=vegan).
- 302 Pavoine, S. (in press). adiv: An R package to analyse biodiversity in ecology. *Methods in*
303 *Ecology and Evolution*. doi: 10.1111/2041-210x.13430.

- 304 R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vi-
305 enna, Austria: R foundation for Statistical Computing. Retrieved from [https://www.R-](https://www.R-project.org/)
306 [project.org/](https://www.R-project.org/)
- 307 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner F. O.
308 (2012). The SILVA ribosomal RNA gene database project: improved data processing and
309 web-based tools. *Nucleic Acids Research* 41(D1), D590-D596. doi: 10.1093/nar/gks1219.
- 310 Rognes, T., Flourie, T., Nichols, B., Quince, C., Mahé, F. (2016). VSEARCH: a versatile
311 open source tool for metagenomics. *PeerJ* 4, e2584. doi:10.7717/peerj.2584.
- 312 Salter, s. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F.,
313 ... Walker, A. W. (2014). Reagent and laboratory contamination can critically impact
314 sequence-based microbiome analyses. *BMC Biology* 12(1), 87. doi: 10.1186/s12915-014-
315 0087-z.
- 316 Schnell, I. B., Bohmann, K., Gilbert, M. T. P. (2015). Tag jumps illuminated - reduc-
317 ing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology*
318 *Resources* 15(6), 1289-1303. doi: 10.1111/1755-0998.12402.
- 319 Sommeria-Klein, G., Zinger, L., Taberlet, P., Coissac, E., Chave, J. (2016). Inferring neutral
320 biodiversity parameters using environmental DNA data sets. *Scientific Report* 6, 35644.
321 doi: 10.1038/srep35644.
- 322 Taberlet, P., Bonin, A., Zinger, L., Coissac, E (2018). *Environmental DNA: For biodiversity*
323 *research and monitoring*. Oxford University Press.

324 Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New
325 York.

326 Wilke, C. O. (2019) *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R
327 package version 1.0.0. Retrieved from <https://CRAN.R-project.org/package=cowplot>.

328 Zinger, L., Bonin, A., Alsos, I.G., Bálint, M., Bik, H., Boyer, F., ... Taberlet, P. (2019).
329 DNA metabarcoding-Need for robust experimental designs to draw sound ecological con-
330 clusions. *Molecular Ecology* 28, 1857-1862. doi: 10.1111/mec.15060.