Towards quantitative DNA Metabarcoding: A method to overcome PCR amplification bias

Sylvain Moinard[1*], Didier Piau[2], Frédéric Laporte[1], Delphine Rioux[1], Pierre Taberlet[1], Christelle Gonindard-Melodelima[1*†] and Eric Coissac[1*†]

[1]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, FR-38000, Grenoble, France.
[2]Univ. Grenoble-Alpes, CNRS, Institut Fourier, FR-38000, Grenoble, France.

*Corresponding authors. E-mails: sylvain.moinard@univ-grenoble-alpes.fr; christelle.gonindard@univ-grenoble-alpes.fr; eric.coissac@metabarcoding.org;
Contributing authors: didier.piau@univ-grenoble-alpes.fr; frederic.laporte@univ-grenoble-alpes.fr; delphine.rioux@univ-grenoble-alpes.fr; pierre.taberlet@univ-grenoble-alpes.fr;
†These authors contributed equally to this work.

Abstract

Metabarcoding analyses have recently undergone significant development due to the power of this technique in biodiversity monitoring. However, it is still difficult to draw accurate quantitative conclusions about the ecosystems studied, mainly because of biases inherent in the environmental DNA or introduced during the experimental process. These biases alter the relationship between the amount of DNA observed and the biomass or number of individuals of the species detected. Two of the biases inherent in metabarcoding have been measured: the ratio between total DNA and target DNA concentrations, and the PCR amplification bias. A method for their correction is proposed. All experimental tests were performed on mock alpine plant communities using the marker *Sper01*, which is expected to have low amplification bias due to its highly conserved priming sites. Our approach combines standard quantitative PCR techniques (qPCR and digital droplet PCR) with a realistic stochastic model of PCR dynamics that accounts for PCR saturation. The model was used to estimate PCR efficiencies for each species and to infer the true species proportions of the mock communities from the read relative frequencies. The corrections are easy to implement and can be applied to previously generated DNA metabarcoding data. This work demonstrates the relative importance of the two biases considered and is an open door to quantitative metabarcoding data, although many other biases remain to be considered.

Keywords: Amplification bias, droplet digital PCR, PCR model, Quantitative metabarcoding, Taqman qPCR

Introduction

In the context of mass species extinction (Barnosky et al., 2011), biodiversity assessment is currently a major challenge. Classically, biodiversity inventories consist not only of a list of species occurring at a site, but also of quantitative data assessing the abundance of each species. Traditional approaches based on direct observation by taxonomists may be unrealistic in terms of available skills and costs, given the enormous effort required to conduct such a survey on a global scale and across the tree of life. Therefore, high-throughput methods, including DNA metabarcoding (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012), are the only chance to achieve such a goal. DNA metabarcoding has been used for more than a decade in many areas of ecology, such as biodiversity monitoring (*e.g.* Bohmann et al., 2014), detection of invasive species (*e.g.* Klymus, Marshall, & Stepien, 2017), or tracking animal diets (*e.g.* Pompanon et al., 2012). It is now part of the basic toolbox of ecologists, if we consider more than a thousand articles published annually based on this technique. While metabarcoding provides a not too much biased overview of biodiversity in terms of species detection (Beng & Corlett, 2020; Ficetola & Taberlet, 2023; Taberlet et al., 2012) with some insight into their relative abundance (Pornon et al., 2016), the quality of quantitative data produced is questionable (Krehenwinkel et al., 2017; Yang et al., 2021).

The relationship between the abundance of a species in the field and the number of sequence reads measured in a DNA metabarcoding experiment is far from straightforward. Many reasons can lead to biased abundance estimates. Biases arise from both natural properties and technical issues (Luo, Ji, Warton, & Yu, 2022; van der Loos & Nijland, 2021). At least three natural biases can be considered. First, if the amount of DNA shed into the environment depends on the biomass of individuals (Elbrecht & Leese, 2015; Elbrecht, Peinert, & Leese, 2017; Lamb et al., 2019), it is also a function of shedding rates specific to each DNA source (Wilder, Farrell, & Green, 2023). Second, the relationship between the eDNA sampled, and the DNA actually shed depends on its decay rate, which in turn depends on the ecosystem studied (Andruszkiewicz Allan, Zhang, Lavery, & Govindarajan, 2021; Krehenwinkel et al., 2018). Third, the number of copies of the DNA marker targeted by metabarcoding per unit of biomass or per individual varies from species to species (Garrido-Sanz, Senar, & Piñol, 2022; Krehenwinkel et al., 2017; Zoschke, Liere, & Börner, 2007), and may also vary among tissues, during development or according to phenology. Two main sources can be considered for technical biases. First, the DNA extraction method, whose efficiency depends on the extracted substrate and varies between taxonomic groups (Dopheide, Xie, Buckley, Drummond, & Newcomb, 2019). Second, the PCR amplification, which implies species-specific amplification biases (Pawluczyk et al., 2015) related to the annealing step (Piñol, Mir, Gomez-Polo, & Agustí, 2015) or to the PCR extension step, which may depend, among other things, on the GC content of the metabarcodes (Nichols et al., 2018). Thus, the sum of all these biases obscures the relationship between the abundance of the sequenced reads and the abundance of the species in terms of biomass or number of individuals.

Metabarcoding thus requires an appropriate pipeline to robustly estimate species abundances (Alberdi & Gilbert, 2019; Mächler, Walser, & Altermatt, 2021). For a long time, that quantification problem has been considered. Authors have proposed improvements by optimizing the choice of primers (Krehenwinkel et al., 2017), by varying the number of PCR cycles for different replicates (Silverman et al., 2021) or by creating mock communities to infer correction factors with one species of interest and one control species (Thomas, Deagle, Eveson, Harsch, & Trites, 2016), with two species of interest in different quantities (Matesanz et al., 2019) or by comparing several mock communities of more complex composition (Krehenwinkel et al., 2017); or to infer PCR efficiencies (Shelton et al., 2022). Internal controls can be used, but these do not allow measuring amplification bias (Smets et al., 2016; Ushio et al., 2018).

The present paper examines the biases introduced by the most commonly criticized step of DNA metabarcoding, the PCR amplification. The strength of the amplification bias and its impact on the estimated abundances of metabarcoding are assessed. This study is based on a new mathematical model of PCR amplification that is applicable to the simulation of DNA metabarcoding experiments. Several models exist to describe PCR dynamics (*e.g.* Carr & Moore, 2012; Hayward, 1998; Mehra & Hu, 2005) but have not been linked to metabarcoding. The model developed from existing models considers the amplification bias between species in conjunction with the saturation phase of PCR amplification, with a minimum number of parameters. A usual model in quantitative metabarcoding is the exponential model, also called log-ratio linear model (*e.g.* Gold et al., 2023; Kelly, Shelton, & Gallego, 2019; Shelton et al., 2022), where the abundance of each species increases geometrically during the PCR. The non-treatment of saturation is not a problem in quantitative real-time PCR (qPCR) because the amplification starts with an exponential phase, but is incompatible with metabarcoding PCR, which relies on the final state of the system.

The impact of low priming site conservation on species detection and quantification of COI markers has been widely discussed. These biases are related to the annealing phase of PCR cycles due to primer mismatches (Clarke, Soubrier, Weyrich, & Cooper, 2014; Piñol et al., 2015; Pompanon et al., 2012). To specifically target the biases induced by the extension step of PCR, we assessed them on three mock alpine plant communities using the *Sper01* marker (Taberlet et al., 2007). This marker is widely used in many ecological studies: soil biodiversity (Yoccoz et al., 2012), paleoecology based on ancient eDNA (Willerslev et al., 2014) or diet (Valentini et al., 2009). Although there is very little variation at the *Sper01* priming sites, no strong annealing bias can be assumed for this marker. However, the length of the metabarcodes and the complexity of its sequence (length and frequency of homopolymers) varies from species to species, making it an appropriate candidate to study extension bias. PCR efficiency for three species was accurately estimated using Taqman qPCR to calibrate our model and then to infer the pre-PCR eDNA proportions of each species. Combined with precise estimates of target DNA concentrations in each species by droplet digital PCR (ddPCR), the results of this experiment demonstrate the benefit of handling PCR extension bias and the variation of target DNA concentration among taxa to correctly estimate taxa abundance from DNA metabarcoding results. Although only a single marker was studied here on a limited number of species, the presented protocol is easily generalizable and opens perspectives for quantitative DNA metabarcoding (qMetabarcoding).

Material and Methods

Metabarcoding experiment

Quantification biases were investigated using three mock communities composed of thirteen alpine plants belonging to the *Spermatophyta* clade (Supplementary Table 1), using the *Sper01* primer (Taberlet, Bonin, Zinger, & Coissac, 2018; Taberlet et al., 2007) targeting the P6 loop of the *trn*L of the chloroplast genome. Plant species were selected for having no mismatches at their priming sites with the *Sper01* primers.

Plant sampling

Plants leaves were collected in Chartreuse and Belledonne massif in the French Alps during Spring 2021 (Supplementary Table 1). Freshly collected material was stored in silica gel before DNA extraction.

DNA Extraction

Plant DNA was extracted using the CTAB protocol (Doyle, 1990), except for *Carpinus betulus*, for which a *DNeasy Plant Mini Kit* (Qiagen) was used after unsuccessful CTAB extractions.

Quantification of target DNA

The total DNA concentration for each plant sample was determined using Qubit (ThermoFisher). The amount of DNA targeted by the *Sper01* primer is not proportional to the total DNA concentration, as the number of chloroplasts per cell is expected to vary between different species and tissues and during plant development (Golczyk et al., 2014; Sakamoto & Takami, 2018; Zoschke et al., 2007). ddPCR was used to provide absolute quantification of the *Sper01* target DNA. ddPCR was preferred over qPCR because it is much less affected by inhibition than qPCR, which varies from sample to sample. (Sidstedt, Rådström, & Hedman, 2020). This quantification was performed using serial dilutions of total DNA concentrations ranging from 6.25×10^{-2} $ng/\mu l$ to 6.25×10^{-5} $ng/\mu l$ with one or two replicates for each condition. The reaction mixtures had a total volume of 20 μl (5 μl of DNA solution, 10 μl of Master Mix EvaGreen, 0.6 μl of primers (forward and reverse) at $10\mu M$, 4.4 μl of milliQ water). The *QX200 Droplet Digital System* (Bio-Rad) was used to generate droplets (*QX200 Droplet Generator*) and to analyze them after PCR amplification (*QX200 Droplet Reader* with the *QuantaSoft Software*). Thermocycler conditions with optimized annealing temperature for the *Sper01* primer (52°C) were set (30 seconds at 95°C, 30 seconds at 52°C, one minute at 72°C). Replicates identified as incorrect by the reader and the most diluted replicate in cases where this concentration was outside the expected detection range were removed.

The concentration index chosen to compare the samples is the expected number of target copies per ng of total DNA. It is calculated from each assay as in the equation 1. The number of copies per μl (in target DNA) is the value measured by ddPCR. C(Total DNA)$_{replicate}$ is the total DNA concentration of the sample in the reaction mix. The average concentration for each species is used for the rest of the protocol.

$$\text{Concentration}(\text{Copies}/ng) = \frac{(\text{Copies}/\mu\text{l})_{\text{ddPCR}}}{\text{C}(\text{Total DNA})_{replicate}} \qquad (1)$$

Mock communities

Three mock communities were constructed after the ddPCR assays: (i) a uniform community ($\boldsymbol{\mathcal{M}_U}$) where each plant has the same concentration of target DNA, (ii) a community where each plant has the same concentration of total DNA ($\boldsymbol{\mathcal{M}_T}$), and (iii) a community where the concentrations of target DNA are distributed according to a geometric sequence of common ratio 1/2 (concentrations of 1, 1/2, 1/4...) ($\boldsymbol{\mathcal{M}_G}$). The species used are described in Table 1. The metabarcode sequences are given in the Supplementary Table 1 and the exact composition of each community is given in the Supplementary Table 2. The comparison between $\boldsymbol{\mathcal{M}_U}$ and $\boldsymbol{\mathcal{M}_T}$ communities allows to determine the bias introduced by variation in the number of chloroplast genomes per unit of total DNA. The $\boldsymbol{\mathcal{M}_U}$ and $\boldsymbol{\mathcal{M}_G}$ comparison allows the estimation of relative PCR extension step efficiencies.

Species	Short form	Length	GC content (%)	Total DNA concentration ($ng/\mu l$)	Rank ($\boldsymbol{\mathcal{M}_G}$)
Briza media	Bme	53	39.6	183	1
Rosa canina	Rca	51	31.4	50.8	2
Lotus corniculatus	Lco	55	38.2	65.2	3
Populus tremula	Ptr	68	25.0	31.4	4
Salvia pratensis	Spr	46	26.1	24.4	5
Lonicera xylosteum	Lxy	46	32.6	45.8	6
Fraxinus excelsior	**Fex**	**39**	**33.3**	**22.4**	**7**
Acer campestre	Aca	56	39.3	12.2	8
Capsella bursa-pastoris	**Cbp**	**48**	**45.8**	**38.8**	**9**
Geranium robertianum	Gro	53	34.0	15.0	10
Carpinus betulus	**Cbe**	**61**	**27.9**	**9.14**	**11**
Abies alba	Aal	47	44.7	3.58	12
Rhododendron ferrugineum	Rfe	46	30.4	3.90	13

Table 1: Plants used for the three mock communities and their characteristics for the *Sper01* marker. Total DNA concentrations are assayed in the samples after extraction by Qubit. Rank stands for decreasing abundance in the $\boldsymbol{\mathcal{M}_G}$ community.

DNA metabarcoding PCR amplification

For each community, 20 replicates ($2\mu l$ of DNA) and one PCR negative control ($2\mu l$ of milliQ water) are made. Three wells are left blank (sequencing controls). Each well was individually tagged. 40 PCR cycles were run with an optimized annealing temperature for *Sper01* (30 seconds at 95°C, 30 seconds at 52°C, one minute at 72°C).

Metabarcoding DNA Sequencing

High-throughput sequencing was performed on NextSeq (Illumina) by Fasteris (Plan-les-Ouates, Switzerland; https://www.fasteris.com/). One library was constructed per community following the Metafast protocol (as proposed by Fasteris).

Bioinformatic pipeline

All the bioinformatic work was performed on a laptop MacBook Air (2017, 2.2 GHz Intel Core i7 Dual Core Processor). The data and analysis scripts are available on the project's git page, https://github.com/LECA-MALBIO/metabar-bias. Raw data was processed with OBITools (version 4 aka OBITools4; Boyer et al., 2016, https://metabarcoding.org/obitools4). Unless otherwise stated, the further analyses were carried out using R.

A DNA metabarcoding experiment model

The goal of the model is to estimate the initial relative abundances of each species s, p_s, from the number of reads R_s among the S different species in the considered environmental sample.

The model integrates the three steps involved in the production of a DNA metabarcoding result from a DNA extract, as in Gold et al. (2023): i) the sampling of a portion of the DNA extract, ii) the PCR amplification, iii) the sampling of a portion of the PCR reaction for sequencing.

Sampling of a portion of the DNA extract

The initial number of molecules in a replicate r, $M_0^s(r)$, is modeled by a Poisson distribution with expectation m_0^s. It is more realistic to represent this variability by a negative binomial distribution with a larger variance as the standard deviation of the final observed proportions is approximately 25 times larger than in the simulations with the Poisson distribution, but this choice simplifies the model and the mean value remains unchanged.

$$\begin{gathered} M_0^s \sim Poisson\left(m_0^s\right) \qquad (2) \\ \text{so that } \mathbb{E}[M_0^s] = m_0^s \text{ and } Var(M_0^s) = m_0^s \end{gathered}$$

The total number of DNA molecules initially present is needed for the inference, for technical reasons. It is known in the mock communities thanks to absolute quantification by ddPCR, but this is not the case in practice. Based on the ddPCR measurements, the order of magnitude of $m_0^{\text{total}} = \sum_s m_0^s$ was set to 10^5 molecules.

PCR amplification

The used PCR model, here called logistic model, accounts for the different amplification efficiencies and the saturation phase. It is related to Hayward (1998) or Carr and Moore (2012) but uses fewer parameters and explicitly incorporates different species. Compared with a conventional exponential model, the logistic model accounts for saturation phase at the end of the PCR (Figure 1). Both are parametric stochastic models.

The models considered describe the evolution of the number of DNA molecules of each species cycle by cycle, denoted M_k^s for each species s at PCR cycle k. Each molecule already present is maintained and has a probability λ_k^s of being replicated again, modeled by a binomial distribution (equation 3) depending on the state of the system after cycle $k-1$, described by the filtration $\mathcal{F}_{k-1}$.

$$M_k^s | \mathcal{F}_{k-1} \sim M_{k-1}^s + Bin(M_{k-1}^s, \lambda_k^s) \qquad (3)$$

Let $X_k = \frac{\sum_{s=1}^{S}\left(M_{k-1}^t - M_0^t\right)}{K}$ be the total number of molecules created prior to the k cycle divided by a charge capacity K, *ie* the total number of DNA molecules that can be created during the amplification. Due to saturation, the effective PCR efficiency of each species, λ_k^s, decreases during the PCR. The logistic saturation has been chosen for its simple shape (equation 4).

$$\lambda_k^s = \begin{cases} \Lambda_s\left(1 - X_k\right) & \text{if } X_k \leq 1 \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

The purely exponential model is a special case with no saturation where $\lambda_k^s = \Lambda_s$ at each cycle k. In this exponential model, the usual quantification formula (equation 5), is in our framework the expected value of M_n^s.

$$M_n^s = M_0^s(1 + \Lambda_s)^n \tag{5}$$

Sampling of a portion of the PCR reaction for sequencing

All the molecules created by the PCR are not sequenced: only a fraction constitutes the observed data, denoted R_s for each species s. At the end of n cycles, the sequencing step is described as a sub-sampling step (equation 6).

$$R_s | M_n^s \sim Bin\left(K.d, \frac{M_n^s}{K}\right) \tag{6}$$

The sub-sampling factor $d = \frac{R_{\text{total}}}{K}$ is computed from the estimated value of K and the known value of $R_{\text{total}} = \sum_{s=1}^{S} R_s$.

A typical result of simulations performed with the two models is shown in Figure 1.

Measure of the amplification efficiencies

Using Taqman qPCR assay

PCR amplification efficiencies Λ_s were measured by qPCR for three of the plant species present in our mock communities: *Carpinus betulus*, *Capsella bursa-pastoris* and *Fraxinus excelsior*. These three species were chosen because their metabarcodes differ widely in sequence length and GC content. This makes it possible to expect different amplification efficiencies and to design specific Taqman internal probes that allow individual PCR efficiency measurements within a mixture of the three plant DNAs. Two different probes were designed for *Carpinus betulus* to evaluate the influence of the probe itself on the measurement. The four probes used are described in the Supplementary Table 3. The assay was performed using Taqman qPCR on a uniform community composed of these three species. A 5-fold serial dilution from 1.05 to 654 copies/μl

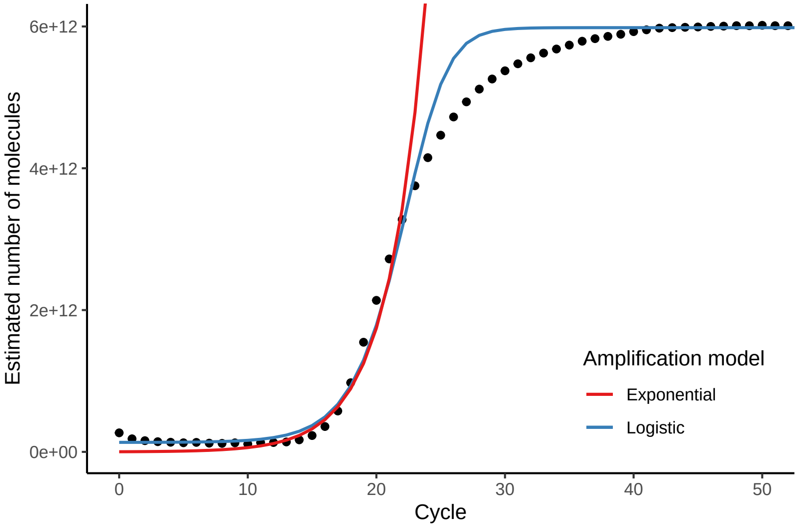

Fig. 1: Observed qPCR kinetics for a sample of *Capsella bursa-pastoris* (black dots) compared to two PCR models fitted to the data. Blue curve: logistic model; red curve: exponential model. An asymmetry of amplification is observed around the inflection point, which creates here a gap between the 25^{th} cycle and the 40^{th} cycle for the logistic model (Gottschalk & Dunn, 2005).

in the reaction mix (25 μl with $5 \mu l$ of DNA) was performed for each probe, with three replicates per concentration. Taqman qPCR was chosen to measure PCR efficiency because it allows measurement from a mixture of the three plant DNAs. This ensures the same inhibitory effect for each species. Since each individual DNA extract has its own pool of inhibitors that interfere with qPCR assays, independent measurement on pure extract would not be realistic (Svec, Tichopad, Novosadova, Pfaffl, & Kubista, 2015).

The exponential model (equation 5), which is valid before the PCR saturation phase, can be used to estimate apparent PCR efficiencies. Estimated efficiencies are referred to as apparent efficiencies because inhibition is always present. For this study, however, only the relative values of the efficiencies are important. A commonly used formula (equation 7, Gill, Bleka, & Fonneløp, 2022) can be derived from the exponential model to estimate amplification efficiencies from a series of qPCRs performed on successive dilutions. However, a major limitation of this formula that has been identified here is that the estimation of the slope is very sensitive to small variations in C_t, resulting in a large variance of the estimator.

$$\text{Linear regression: } C_t = -\frac{log_{10}(m_0)}{log_{10}(1+\Lambda)} + \frac{log_{10}(M_{C_t})}{log_{10}(1+\Lambda)}$$

$$\begin{aligned} &= a\log_{10}(m_0) + b + \epsilon,\ \epsilon \sim \mathcal{N}(0,\sigma^2)\ iid \\ \Lambda &= 10^{-1/a} - 1 \\ \text{and } M_{C_t} &= 10^{-b/a} \end{aligned} \tag{7}$$

To estimate efficiencies more precisely, this approach is adapted. Linear regression is used to estimate the constant value $M_{C_t} \simeq 1.5 : 10^{11}$ (number of molecules present at C_t) by averaging the results for the three species. Then $K \simeq 7.6 : 10^{12}$ (equation 8) is inferred from observed relative fluorescence unit (RFU) values, assuming within-replicate proportionality between RFU and DNA copy number (Gill et al., 2022), although RFU values are not standardized and depend on many experimental factors (Svec et al., 2015).

$$K \simeq M_{C_t} \times \sum_{s=1}^{3} \frac{RFU_{End}}{RFU_{C_t}}(s) \tag{8}$$

Then, the efficiencies Λ_s were estimated for each replicate from this constant value of M_{C_t} (equation 9). For subsequent analyses, the average Λ_s over all replicates is used.

$$\begin{aligned} M_{C_t} &= M_0^s(1+\Lambda_s)^{C_t(s)} \\ \text{so} \quad \Lambda_s &= \left(\frac{M_{C_t}}{M_0^s}\right)^{1/C_t(s)} - 1 \end{aligned} \tag{9}$$

The extreme estimates of M_{C_t} vary by a factor of 2.1, which implies a low potential factor, applied equally to all Λ_s, of the order of 1.03.

Using the $\mathcal{M}_U$ community

PCR efficiencies were also inferred by optimizing the logistic PCR model presented above to fit experimental data, using known initial quantities of the $\mathcal{M}_U$ community. The Fixed Landscape Inference MethOd (*flimo*, Moinard, Oudet, Piau, Coissac, & Gonindard-Melodelima, 2022) implemented in Julia was used for this purpose. The *flimo* method minimizes an objective function in the form of a χ^2 statistic (equation 10).

$$\begin{aligned} &\underset{m_0^1,\ldots,m_0^s>0}{\operatorname{argmin}}\ J((m_0^s)_s) \\ \text{with} \quad &J(m_0^1,\ldots,m_0^s) = \sum_{s=1}^{S} \frac{\left(\overline{p_s}(\text{data}) - \widehat{p}_s\right)^2}{\overline{p_s}(\text{data})} \end{aligned} \tag{10}$$

where $\widehat{p}_s$ is the average proportion of species s in a replicate, estimated over $n_{sim} = 1900$ simulations knowing the $(m_0^s)_s$, and $\overline{p_s}(\text{data})$ is the average proportion of species s in the data.

However, the inferred efficiencies are relative, as the model can produce similar results for different ranges of Λ_s. The maximum efficiency value has been set at 1. These efficiencies are then reused to infer the proportions of the thirteen species.

Correction of relative abundances of a MOTU

Figure 2 summarizes the additional pipeline recommended for correcting amplification bias in a metabarcoding experiment. The PCR amplification efficiency of each species is estimated from samples of species characteristic of the ecosystem studied that are assayed by ddPCR. There are two ways of doing this: Taqman qPCR or a mock community study. These efficiencies are then used to infer the initial proportions of each species.

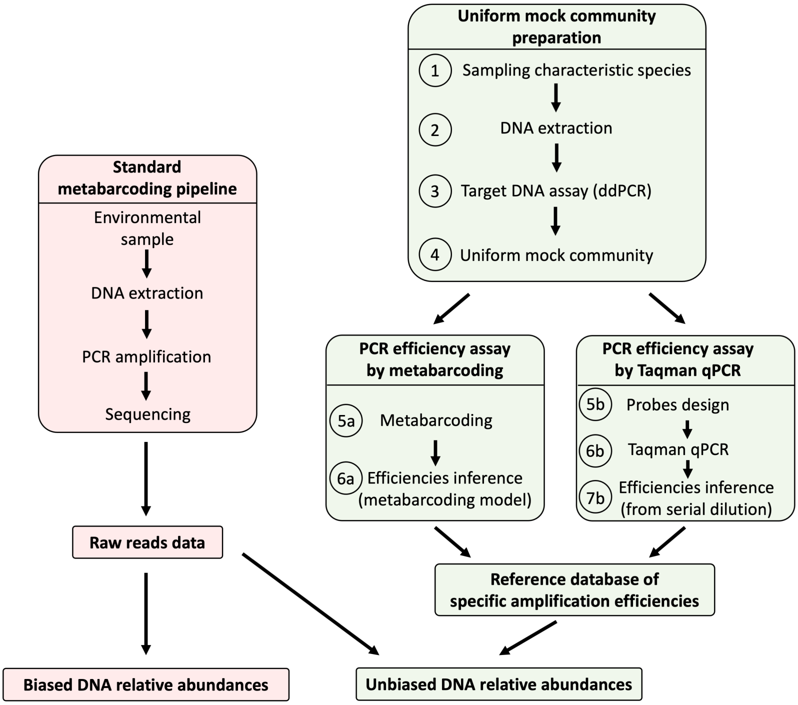
Uniform mock community preparation
1 Sampling characteristic species
2 DNA extraction
3 Target DNA assay (ddPCR)
4 Uniform mock community
Standard metabarcoding pipeline
Environmental sample
DNA extraction
PCR amplification
Sequencing
PCR efficiency assay by metabarcoding
5a Metabarcoding
6a Efficiencies inference (metabarcoding model)
PCR efficiency assay by Taqman qPCR
5b Probes design
6b Taqman qPCR
7b Efficiencies inference (from serial dilution)
Raw reads data
Reference database of specific amplification efficiencies
Biased DNA relative abundances
Unbiased DNA relative abundances

Fig. 2: Additional pipeline recommended for correcting amplification bias in a metabarcoding experiment as presented in this study.

Using the Ratio method

Previous works (*e.g.* Shelton et al., 2022; Silverman et al., 2021) showed that a reference mock community can be used to correct abundances in another community composed of the same species. Although this was not the main objective of our work, this result was verified using the three communities studied. The $\boldsymbol{\mathcal{M}_U}$ community was used as a reference to correct abundances in the $\boldsymbol{\mathcal{M}_T}$ and $\boldsymbol{\mathcal{M}_G}$ communities. In the $\boldsymbol{\mathcal{M}_U}$ community, each species had a starting relative frequency of $1/13 \simeq 7.7\%$, which should have been observed in the final read proportions in the absence of amplification bias. The correction factor for each species c_s is therefore simply the median ratio between the expected and the observed reads frequencies over all replicates in the $\boldsymbol{\mathcal{M}_U}$ community (equation 11).

$$c_s = \text{Median}\left(\frac{\text{Observed reads frequency}}{\text{Expected reads frequency}}(s)\right) \tag{11}$$

For the $\boldsymbol{\mathcal{M}_T}$ and $\boldsymbol{\mathcal{M}_G}$ communities, this correction factor is applied to estimate the initial proportions $\widehat{p_s}$ for each species s (equation 12).

$$\begin{aligned} R'_s &= \frac{\text{Reads}(s)}{c_s} \\ \widehat{p_s} &= \frac{R'_s}{\sum_t R'_t} \end{aligned} \tag{12}$$

Using the estimated amplification efficiencies

The inference of the actual proportions of eDNA from the relative read abundances (RRA) measured after DNA metabarcoding sequencing is achieved by the same algorithmic method presented above, but this time the Λ_s efficiencies are assumed to be known.

The efficiencies measured by Taqman qPCR or inferred from the model fit for the $\boldsymbol{\mathcal{M}_U}$ community can be used to infer the initial proportions of $\boldsymbol{\mathcal{M}_T}$ and $\boldsymbol{\mathcal{M}_G}$.

An estimate of these proportions can be obtained using the exponential model, but this requires knowledge of the PCR equivalent number of "exponential cycles". The result is then given by the $\widehat{m_0^s(k)}$ calculated at cycle k with the equation 13. The problem is that the relative frequencies vary by several points depending on the cycle chosen. This method has not been included in the following.

$$\widehat{m_0^s(k)} = \frac{K}{(1+\Lambda_s)^k} \times \frac{R_s}{\sum_{t=1}^{S} R_t} \tag{13}$$

Criteria for measuring quantification errors

The distance between the observed or corrected proportions $(\widehat{p_s})_s$, median over all the replicates) and the initial theoretical proportions (p_s^{th}) is measured by two RMSE (*Root-Mean-Square Error*) criteria. The error measured is either absolute (equation 14) or relative (normalized by the theoretical proportions, equation 15).

$$\text{Absolute Error: AbsErr}((\widehat{p_s})_s) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\widehat{p_s} - p_s^{\mathrm{th}})^2} \tag{14}$$

and

$$\text{Relative Error: RelErr}((\widehat{p_s})_s) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\frac{\widehat{p_s} - p_s^{\mathrm{th}}}{p_s^{\mathrm{th}}}\right)^2} \tag{15}$$

Ecological conclusions: biodiversity indices

To compare theoretical, observed and inferred compositions, biodiversity indices were computed for $\boldsymbol{\mathcal{M}_T}$ and $\boldsymbol{\mathcal{M}_G}$. Hill numbers (Hill, 1973) (equation 16), interpretable as an effective number of species in the community, were chosen with $q = 1$ (linked to Shannon entropy) and $q = 2$ (linked to Gini-Simpson index).

$$^{q}D = \left(\sum_{s=1}^{S} p_s^q\right)^{\frac{1}{1-q}} \tag{16}$$

Results

ddPCR assay

The concentrations of each plant sample measured by ddPCR are shown in Figure 3. For the same total DNA concentration, there was a wide variability in average target concentration, ranging from 3.7×10^4 copies per ng for *Rhododendron ferrugineum* to 2.5×10^5 copies per ng for *Populus tremula* with an average of 1.1×10^5 copies per ng among the thirteen species. The factor between the extremes is thus 6.6.

Metabarcoding experiment

Raw sequencing data

After processing with the OBITools, an average of 37,000 reads per non-negative replicate was obtained with a standard deviation of 27,000 reads (first

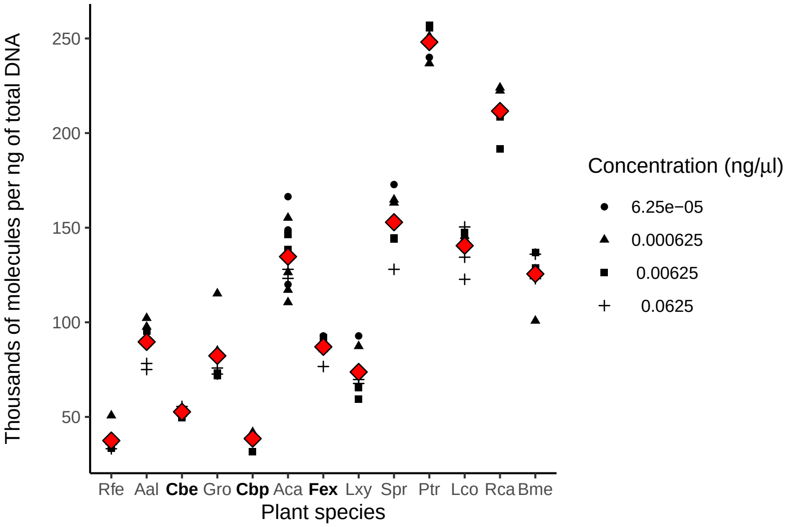

Fig. 3: Number of target DNA molecules (thousands) per ng of total DNA for thirteen alpine plants, computed with the index used in equation 1. Each black dot is a replicate, for different total DNA concentrations. The red diamonds correspond to the mean for each species.

and third quartiles : 14,000 and 56,000 reads). Negative controls showed negligible contamination. For each community out of the 20 PCR replicates, one replicate with fewer than 5,000 reads was discarded from further analysis.

Reads proportions

The comparison of observed and expected read proportions is shown in Figure 4. Significant differences can be observed: at most, between the observed and expected proportions, there is a factor of 3.0 for *Geranium robertianum* in the $\boldsymbol{\mathcal{M}_U}$ community, 4.2 for *Abies alba* in $\boldsymbol{\mathcal{M}_T}$ and 9.0 for *Abies alba* in $\boldsymbol{\mathcal{M}_G}$.

Comparing the observed proportions with the expected proportions allows to visualize the two biases under study. For example, *Rosa canina* species has both good efficiency and a high target concentration: the two biases add up. Conversely, *Geranium robertianum* is penalized by both biases. *Salvia pratensis* has a higher-than-average concentration, but poor efficiency. *Capsella bursa-pastoris* is well amplified, but its target concentration is low.

The joint effect of the double bias is visible for $\boldsymbol{\mathcal{M}_T}$, with median proportions comprised between 1.5% and 26%, and between 2.6% and 17% for $\boldsymbol{\mathcal{M}_U}$.

Inter-replicate variability is significant in some species, such as *Populus tremula* (in $\boldsymbol{\mathcal{M}_U}$: mean proportion : 8.6%, varying from 3.3% to 14%, standard deviation of 2.7%).

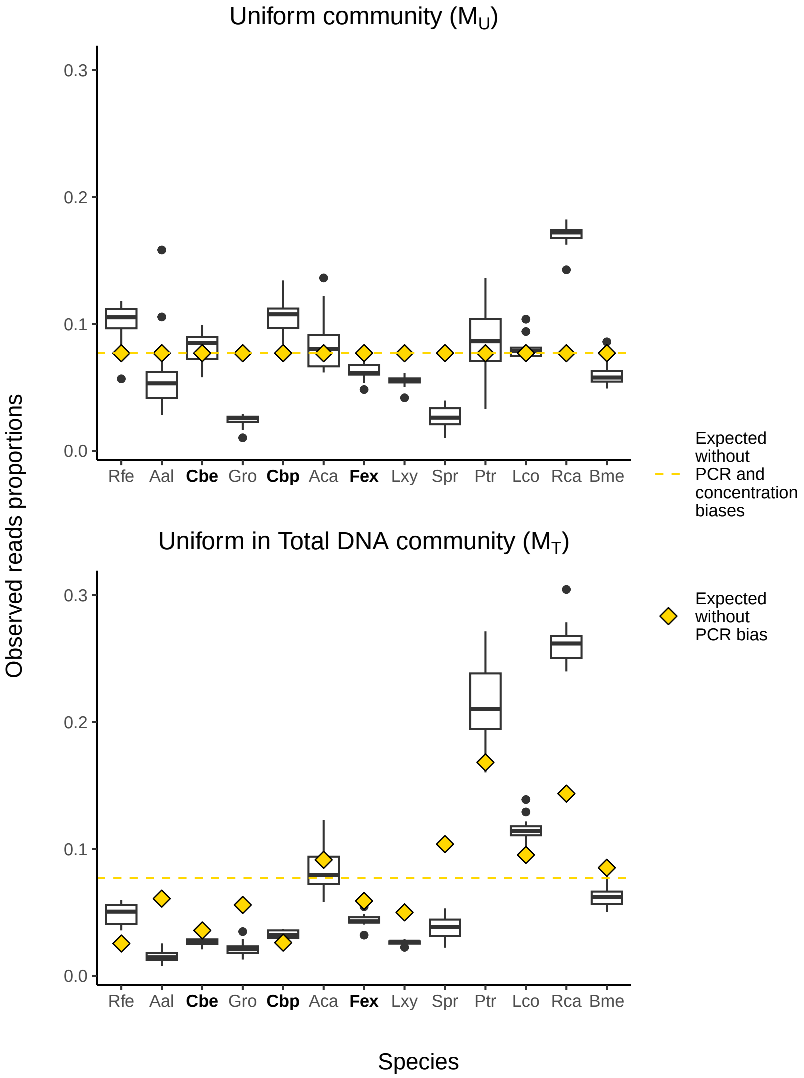

Fig. 4: Observed relative proportions of reads of thirteen plant species for the mock communities $\mathcal{M}_U$ and $\mathcal{M}_T$. Gold lines indicate proportions expected in the absence of target concentration and amplification bias. Gold diamonds are the proportions expected in the absence of amplification bias. For the $\mathcal{M}_U$ community, the deviation of the boxplots from the diamonds shows the amplification bias alone. For the $\mathcal{M}_T$ community, both biases are present. Concentration bias is visible as the difference between the diamond and the line.

Inferring PCR efficiencies and abundances 401

The apparent PCR efficiencies for the three species tested (**Fex**, **Cbe**, **Cbp**) 402
measured using the Taqman qPCR method for the four probes have a relative 403

differences of the order of 5%. That can be considered low, but due to the exponential nature of PCR, it has a real impact on the final proportions in the community due to the exponential nature of PCR amplification.

Table 2 shows the abundances in the reference mock community $\boldsymbol{\mathcal{M}_U}$ and the efficiencies inferred from the Taqman qPCR assay and from the model fit to the $\boldsymbol{\mathcal{M}_U}$ community, with the *flimo* method (around 100 seconds for the thirteen species). The lowest efficiency is around 15% lower than the maximum. The absolute values determined by Taqman qPCR are overestimated in relation to these values, but once normalized, they are broadly similar, even though more values would be required for a rigorous comparison. Because of this similarity and the fact that the assay involves only three species, the results are based on efficiencies measured in $\boldsymbol{\mathcal{M}_U}$.

Species	Average proportion in $\boldsymbol{\mathcal{M}_U}$ (%)		PCR Efficiency inferred from	
	Theoretical	Observed	Taqman	$\boldsymbol{\mathcal{M}_U}$
Bme	7.7	6.1		0.922
Rca	7.7	17		1.00
Lco	7.7	8.0		0.942
Ptr	7.7	8.6		0.948
Spr	7.7	2.7		0.862
Lxy	7.7	5.5		0.915
Fex	**7.7**	**6.3**	**0.924**	**0.924**
Aca	7.7	8.3		0.945
Cbp	**7.7**	**11**	**0.973**	**0.964**
Gro	7.7	2.4		0.855
Cbe	**7.7**	**8.1**	**0.956 (CbeA)** **0.931 (CbeB)**	**0.943**
Aal	7.7	5.8		0.918
Rfe	7.7	10		0.960

Table 2: Proportions in $\boldsymbol{\mathcal{M}_U}$ and relative PCR amplification efficiencies measured for the four Taqman qPCR probes and inferred from the $\boldsymbol{\mathcal{M}_U}$ community. The maximum efficiency was set at 1 for *Rosa canina*. Efficiencies inferred were normalized so that Fex has the same efficiencies with both methods.

Table 3 shows the proportions in the $\boldsymbol{\mathcal{M}_T}$ and $\boldsymbol{\mathcal{M}_G}$ communities, as well as the errors compared to the theoretical proportions and the biodiversity indices. The results of the two corrections are comparable and both improve the RMSE criteria, as expected. The corrected biodiversity indices also seem to better approximate the real biodiversity than the observed values.

PCR bias importance: comparison of model simulations and observed data

To illustrate the effect of small differences in efficiency, PCR kinetics was simulated for two species with equal initial quantities. Figure 5 shows the final proportions of the two species according to the difference in PCR efficiency. These simulations are compared with the proportions observed in the

Species	Average proportion in $\boldsymbol{\mathcal{M}_T}$ (%)				Average proportion in $\boldsymbol{\mathcal{M}_G}$ (%)			
	Theoretical	Observed	Inferred with $\boldsymbol{\mathcal{M}_U}$	Inferred with Λ_s	Theoretical	Observed	Inferred with $\boldsymbol{\mathcal{M}_U}$	Inferred with Λ_s
Bme	8.5	6.2	8.4	8.3	50	36	54	54
Rca	14	26	12	12	25	40	20	19
Lco	9.5	11	11	12	13	15	16	16
Ptr	17	21	20	20	6.3	5.2	5.5	5.5
Spr	10	3.9	12	12	3.1	0.63	2.0	2.2
Lxy	5.0	2.7	3.8	3.9	1.6	0.94	1.5	1.6
Fex	**5.9**	**4.3**	**5.7**	**5.6**	**0.78**	**0.68**	**0.96**	**0.96**
Aca	9.1	7.9	7.8	8.1	0.39	0.16	0.17	0.17
Cbp	**2.6**	**3.2**	**2.4**	**2.4**	**0.20**	**0.19**	**0.15**	**0.15**
Gro	5.6	2.1	6.5	7.4	0.098	0.019	0.064	0.091
Cbe	**3.6**	**2.8**	**2.6**	**2.7**	**0.049**	**0.030**	**0.031**	**0.032**
Aal	6.1	1.5	2.3	2.1	0.024	0.0045	0.0072	0.0045
Rfe	2.5	5.1	3.7	3.8	0.012	0.014	0.012	0.015
AbsErr		0.045	0.017	0.019		0.057	0.020	0.022
RelErr		0.53	0.26	0.28		0.50	0.34	0.34
^1D	11	9.8	11	11	4.0	3.7	3.7	3.8
^2D	10	6.7	9.1	9.1	3.0	3.1	2.8	2.8

Table 3: Proportions of species in $\boldsymbol{\mathcal{M}_T}$ and $\boldsymbol{\mathcal{M}_G}$. Inferred with $\boldsymbol{\mathcal{M}_U}$ means corrected by the ratios. Proportions inferred with Λ_s are obtained by fitting the PCR model using the efficiencies inferred previously.

$\boldsymbol{\mathcal{M}_U}$ community when comparing *Rosa canina* (the most efficiently amplified species) and the other species individually. These two proportion series are very close to each other.

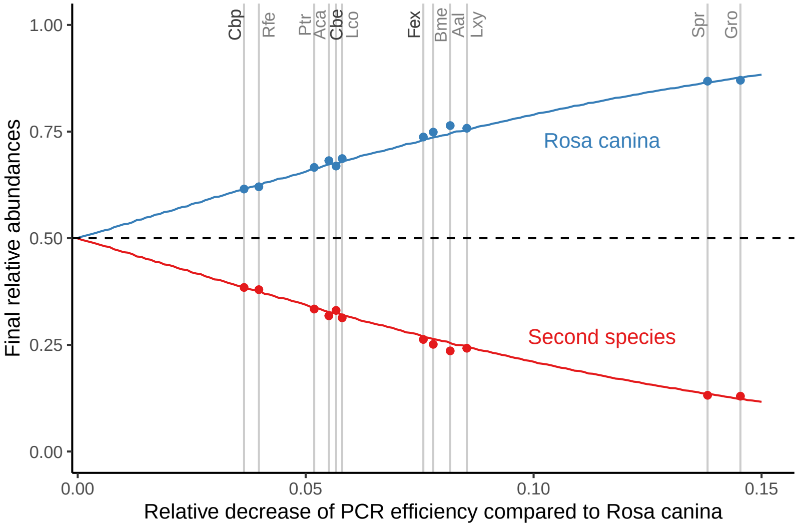

Fig. 5: Relative abundances in a mock community of two initially evenly distributed species simulated with the logistic model (lines) and observed in the $\boldsymbol{\mathcal{M}_U}$ community (dots) considering only *Rosa canina* and the other species individually. The first species has an efficiency of $\Lambda_1 = 1$. The second has a variable efficiency, of value $\Lambda_2 = \Lambda_1(1 - x)$ along the x-axis ($\Lambda_2 \in [0.85, 1.0]$).

Discussion

The quantitative aspect of DNA metabarcoding is regularly questioned by ecologists. Here, two potential biases were considered and their relative effects quantified.

The first is well known. It has long been discussed by microbial ecologists (Kembel, Wu, Eisen, & Green, 2012; Milivojević et al., 2021) and has been identified for macroorganisms (Garrido-Sanz et al., 2022; Krehenwinkel et al., 2017). It can be summarized by a simple question: how many copies of the target gene marker are present per genome in each species under consideration? In macro-organisms such as plants and animals, most of the targeted markers are carried by the chloroplast or mitochondrial genome, but the same question remains: how many copies of the organelle genome are there per cell? When the genome size of a species is unknown, the best proxy of this number of copies is the number of marker copies per weight unit of total DNA. This amount can be estimated by ddPCR. Among the 13 plants tested, the one more concentrated in chloroplast DNA, *Populus tremula* (Ptr), has 6.6 times more copies per unit of nuclear DNA than the one less concentrated, *Rhododendron ferrugineum* (Rfe). According to the Kew C-value database (https://cvalues.science.kew.org/), the 1C value of Ptr is 0.45 pg (Siljak-Yakovlev et al., 2010) and that of *Rhododendron ponticum*, the only *Rhododendron* measured, is 0.74 pg (Bou Dagher-Kharrat et al., 2013). Both together allow to estimate that the bias in chloroplast abundance (in copies per genome) can lead to a 4-fold overestimation of Ptr abundances relative to Rfe.

The second type of bias is an amplification bias, which has never been quantified. The amplification efficiency of a marker for the species s (Λ_s) is an intrinsic property of the sequence. It does not depend on co-amplified sequences. It can be measured by either of the two methods proposed in this study. Both methods provide similar values, and the choice between them depends on practical convenience. The values obtained can be used to correct the composition of any community, as long as differences in amplifiability between the species present do not cause one or more to disappear. The proposed correction method combines the generation of a reference base for the amplifiability and a mathematical model of the PCR. It does not require any modification of the metabarcoding protocol. Therefore, it can be applied to already generated results and is easy to implement.

The amplification bias is accumulated over each PCR cycle. Thus, the final bias on the observed read relative frequencies is a function of the amplifiability per cycle and the number of amplification cycles. In PCR, the actual number of amplification cycles is not necessarily the number of cycles programmed into the PCR instrument. This number may be lower because the total amount of DNA that can be synthesized is limited by the nucleotide concentration. It is therefore possible that the plateau will be reached before the programmed number of cycles has been reached, with the last cycles not corresponding to any amplification (Figure 1). Correcting for bias using the ratio method (*e.g.* Shelton et al., 2022; Silverman et al., 2021) requires that each sample, including the reference mock community used to estimate it, be amplified with the same effective number of PCR cycles. This means that each sample must contain the same total number of target DNA molecules at the start of the PCR. In our study, each mock community was prepared with close total amounts of target DNA, thus respecting the ideal condition for using the ratio method. Therefore, as shown in Table 3, the corrections made by the ratio method and our PCR model-based approach are strictly equivalent. When samples contain different amounts of target DNA, the efficiency of the ratio method should decrease because the number of effective PCR cycles varies from sample to sample. Fortunately, our PCR model-based correction method allows us to estimate the effective number of PCR cycles for each sample, thereby accounting for sample heterogeneity. Without performing ddPCR on each sample and diluting to equilibrate the amount of target DNA between samples, our model-based method results in a correction that is more robust to expected inter-sample variability than the ratio method.

When the two species with the most different amplifiability, *Rosa canina* ($\Lambda_{Rca} = 1.000$) and *Geranium robertianum* ($\Lambda_{Gro} = 0.855$) are co-amplified, with equal amounts of initial target DNA in the extract, the ratio between the RRA observed after sequencing can be up to 6.7 (Figure 5), leading to a strong overestimation of Rca abundance relative to Gro. This initial assessment shows that due to the exponential nature of PCR, even a small difference in amplifiability, as little as 15% between Rca and Gro, the two extreme species tested, can have as strong an effect on the observed RRA as the bias observed due to chloroplast richness. Sometimes the two biases studied push in the same direction, as in *Populus tremula* (Ptr), which has a high chloroplast concentration and a high amplifiability, or *Capsella bursa-pastoris* (Cbe), which combines both a low chloroplast concentration and a low amplifiability (Fig. 4). Sometimes, by chance, both biases partially compensate, as in *Salvia pratensis* (Spr).

Even if the abundances observed by traditional surveys and those of metabarcoding reads are correlated (Yoccoz et al., 2012), it is necessary to be cautious when analyzing DNA metabarcoding data in terms of quantitative information. If we consider the estimation of biodiversity indices, the worst situation is the estimation of α-diversity. Because of all the biases acting simultaneously on DNA metabarcoding measures, but their good reproducibility, the information they provide is inherently relative. Relative in terms of abundances, DNA metabarcoding can at best provide relative abundances, but also relative because the values provided are biased. Therefore, only changes between measures are truly meaningful. Although it has been shown that α-diversity of plant communities can be correctly estimated from DNA metabarcoding data (Calderón-Sanou, Münkemüller, Boyer, Zinger, & Thuiller, 2020), the limited condition under which this is true, Hill numbers computed for $q = 1$, indicates that this is because at this level of weighting of rare versus abundant species by chance most of the biases are compensated.

This phenomenon can also be observed in our results (Table 3), where ${}^1D$ and ${}^2D$ values estimated from raw RRA and corrected abundances do not strongly differ, while the error between RRA and theoretical composition decreases by a factor of two when using corrected abundances. This discrepancy between the decrease in error due to the correction and the not so good increase in the quality of the α-diversity estimates can be at least partially explained in $\boldsymbol{\mathcal{M}_G}$ by the abundances of the two most abundant species, *Briza media* (Bme) and *Rosa canina* (Rca), which have inverted abundances when estimated from RRA. For any study analyzing changes in diversity across time or ecological gradients, because metabarcoding measures are biased but accurate, the true β-diversity patterns can be easily detected using metabarcoding. In fact, because the biases are repeatable between measures, they often amplify the pattern because the errors correlate with the ecological signal. The problem of all these biases only arises when trying to disentangle the observed pattern from changes in specific species. Therefore, we can strongly encourage people to be very cautious when interpreting the observed pattern, and to be careful not to over-interpret changes in the abundance of a few species in the community as an ecological cause.

Conclusion

We investigated two of the biases that prevent proper quantification of relative eDNA abundances in metabarcoding data. Despite their importance, these biases are far from being corrected or even considered in most current studies. In this study, we measure the two studied biases and propose a simple method to correct the amplification biases in the limit of extreme cases where some species are so strongly disadvantaged that they disappear from the raw results. The advantage of our method compared to the previous ones is that it is more robust to sample variability, while compared to the spiking-based method it does not require any change in metabarcoding protocols. This also allows the reanalysis of previously obtained results, providing the opportunity for a better ecological interpretation of them. By combining relative abundance correction and ddPCR to estimate the amount of target DNA in each sample, we can even consider the possibility of having access to an absolute quantification of DNA in the analyzed DNA extracts for each species instead of only relative abundances. This opens the possibility to increase the robustness of the quantitative interpretation of DNA metabarcoding results, although other biases still need to be assessed and modeled in a similar way to fully achieve the goal of truly quantitative metabarcoding.

Acknowledgments

The authors thank Christian Miquel for the logistic support, Frédéric Boyer and Clément Lionnet for their help with the bioinformatics pipeline. This work was supported by the Alpalga project (ANR-20-CE02-0020).

Data Accessibility and Benefit Sharing statement

Data Accessibility statement

The data and analysis scripts are available on the project's git page, https://github.com/LECA-MALBIO/metabar-bias.

Competing interests

The authors declare no competing interests.

Authors' contributions

SM, EC, CG and DP studied the PCR models. SM, EC, CG and PT designed the associated experimental protocol. SM, EC, CG and PT wrote the manuscript. DP contributed to the writing of the manuscript. EC and PT sampled the plants. DR and SM performed the extractions and metabarcoding PCRs. FL and SM performed the qPCR and ddPCR assays. SM wrote the analysis script. EC and CG supervised the project.

References

Alberdi, A., & Gilbert, M.T.P. (2019, July). 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

Andruszkiewicz Allan, E., Zhang, W.G., Lavery, A., Govindarajan, A. (2021, March). Environmental DNA shedding and decay rates from diverse animal forms and thermal regimes. *Environmental DNA*, *3*(2), 492–514, https://doi.org/10.1002/edn3.141

Barnosky, A.D., Matzke, N., Tomiya, S., Wogan, G.O.U., Swartz, B., Quental, T.B., ... Ferrer, E.A. (2011, March). Has the Earth's sixth mass extinction already arrived? *Nature*, *471*(7336), 51–57, https://doi.org/10.1038/nature09678 Retrieved from http://dx.doi.org/10.1038/nature09678

Beng, K.C., & Corlett, R.T. (2020, June). Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. *Biodiversity and Conservation*, *29*(7), 2089–2121, https://doi.org/10.1007/s10531-020-01980-0

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., ... de Bruyn, M. (2014, June). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, *29*(6), 358–367, https://doi.org/10.1016/j.tree.2014.04.003

Bou Dagher-Kharrat, M., Abdel-Samad, N., Douaihy, B., Bourge, M., Fridlender, A., Siljak-Yakovlev, S., Brown, S.C. (2013, December). Nuclear DNA C-values for biodiversity screening: Case of the Lebanese flora. *Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology*, *147*(4), 1228–1237, https://doi.org/10.1080/11263504.2013.861530 Retrieved from https://doi.org/10.1080/11263504.2013.861530

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., Coissac, E. (2016, January). obitools : a unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182, https://doi.org/10.1111/1755-0998.12428

Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., Thuiller, W. (2020, January). 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 Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/jbi.13681

Carr, A.C., & Moore, S.D. (2012, May). Robust Quantification of Polymerase Chain Reactions Using Global Fitting. *PLoS ONE*, *7*(5), e37640, https://doi.org/10.1371/journal.pone.0037640 Retrieved 2023-07-12, from https://dx.plos.org/10.1371/journal.pone.0037640

Clarke, L.J., Soubrier, J., Weyrich, L.S., Cooper, A. (2014, November). Environmental metabarcodes for insects: *in silico* PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, *14*(6), 1160–1170, https://doi.org/10.1111/1755-0998.12265

Dopheide, A., Xie, D., Buckley, T.R., Drummond, A.J., Newcomb, R.D. (2019, January). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods in Ecology and Evolution*, *10*(1), 120–133, https://doi.org/10.1111/2041-210X.13086 Retrieved 2023-07-19, from https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13086

Doyle, J.J. (1990). Isolation of plant dna from fresh tissue.. Retrieved from https://api.semanticscholar.org/CorpusID:85677467

Elbrecht, V., & Leese, F. (2015, July). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PLOS ONE*, *10*(7), e0130324, https://doi.org/10.1371/journal.pone.0130324

Elbrecht, V., Peinert, B., Leese, F. (2017, September). Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, *7*(17), 6918–6926, https://doi.org/10.1002/ece3.3192

Ficetola, G.F., & Taberlet, P. (2023, February). Towards exhaustive community ecology via dna metabarcoding. *Molecular Ecology*, mec.16881, https://doi.org/10.1111/mec.16881

Garrido-Sanz, L., Senar, M.A., Piñol, J. (2022, January). Relative species abundance estimation in artificial mixtures of insects using mito-metagenomics and a correction factor for the mitochondrial DNA copy number. *Molecular Ecology Resources*, *22*(1), 153–167, https://doi.org/10.1111/1755-0998.13464

Gill, P., Bleka, O., Fonneløp, A.E. (2022, November). Limitations of qPCR to estimate DNA quantity: An RFU method to facilitate inter-laboratory comparisons for activity level, and general applicability. *Forensic Science International: Genetics*, *61*, 102777, https://doi.org/10.1016/j.fsigen.2022.102777

Golczyk, H., Greiner, S., Wanner, G., Weihe, A., Bock, R., Börner, T., Herrmann, R.G. (2014, April). Chloroplast DNA in Mature and Senescing Leaves: A Reappraisal. *The Plant Cell*, *26*(3), 847–854, https://doi.org/10.1105/tpc.113.117465

Gold, Z., Shelton, A.O., Casendino, H.R., Duprey, J., Gallego, R., Van Cise, A., ... Kelly, R.P. (2023, May). Signal and noise in metabarcoding data. *PLOS ONE*, *18*(5), e0285674, https://doi.org/10.1371/journal.pone.0285674 Retrieved 2023-07-23, from https://dx.plos.org/10.1371/journal.pone.0285674

Gottschalk, P.G., & Dunn, J.R. (2005, August). The five-parameter logistic: A characterization and comparison with the four-parameter logistic. *Analytical Biochemistry*, *343*(1), 54–65, https://doi.org/10.1016/j.ab.2005.04.035 Retrieved 2023-07-12, from https://linkinghub.elsevier.com/retrieve/pii/S0003269705003313

Hayward, A. (1998, June). Modeling and analysis of competitive RT-PCR. *Nucleic Acids Research*, *26*(11), 2511–2518, https://doi.org/10.1093/nar/26.11.2511

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

Kelly, R.P., Shelton, A.O., Gallego, R. (2019, December). Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. *Scientific Reports*, *9*(1), 12133, https://doi.org/10.1038/s41598-019-48546-x

Kembel, S.W., Wu, M., Eisen, J.A., Green, J.L. (2012, October). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Computational Biology*, *8*(10), e1002743, https://doi.org/10.1371/journal.pcbi.1002743 Retrieved 2023-08-01, from https://dx.plos.org/10.1371/journal.pcbi.1002743

Klymus, K.E., Marshall, N.T., Stepien, C.A. (2017, May). Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. *PLOS ONE*, *12*(5), e0177643, https://doi.org/10.1371/journal.pone.0177643

Krehenwinkel, H., Fong, M., Kennedy, S., Huang, E.G., Noriyuki, S., Cayetano, L., Gillespie, R. (2018, January). The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. *PLOS ONE*, *13*(1), e0189188, https://doi.org/10.1371/journal.pone.0189188

Krehenwinkel, H., Wolf, M., Lim, J.Y., Rominger, A.J., Simison, W.B., Gillespie, R.G. (2017, December). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, *7*(1), 17668, https://doi.org/10.1038/s41598-017-17333-x

Lamb, P.D., Hunter, E., Pinnegar, J.K., Creer, S., Davies, R.G., Taylor, M.I. (2019, January). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, *28*(2), 420–430, https://doi.org/10.1111/mec.14920

Luo, M., Ji, Y., Warton, D., Yu, D.W. (2022, August). Extracting abundance information from dna-based data. *Molecular Ecology Resources*, 1755–0998.13703, https://doi.org/10.1111/1755-0998.13703

Matesanz, S., Pescador, D.S., Pías, B., Sánchez, A.M., Chacón-Labella, J., Illuminati, A., ... Escudero, A. (2019, September). Estimating belowground plant abundance with DNA metabarcoding. *Molecular Ecology Resources*, *19*(5), 1265–1277, https://doi.org/10.1111/1755-0998.13049

Mehra, S., & Hu, W.-S. (2005, September). A kinetic model of quantitative real-time polymerase chain reaction. *Biotechnology and Bioengineering*, *91*(7), 848–860, https://doi.org/10.1002/bit.20555

Milivojević, T., Rahman, S.N., Raposo, D., Siccha, M., Kucera, M., Morard, R. (2021, October). High variability in SSU rDNA gene copy number among planktonic foraminifera revealed by single-cell qPCR. *ISME Communications*, *1*(1), 63, https://doi.org/10.1038/s43705-021-00067-3 Retrieved 2023-10-02, from https://www.nature.com/articles/s43705-021-00067-3

Moinard, S., Oudet, E., Piau, D., Coissac, E., Gonindard-Melodelima, C. (2022). The Fixed Landscape Inference MethOd (flimo): a versatile alternative to Approximate Bayesian Computation, faster by several orders of magnitude. *arXiv*, , https://doi.org/10.48550/ARXIV.2210.06520 (Publisher: arXiv Version Number: 3)

Mächler, E., Walser, J., Altermatt, F. (2021, July). Decision-making and best practices for taxonomy-free environmental DNA metabarcoding in biomonitoring using Hill numbers. *Molecular Ecology*, *30*(13), 3326–3339, https://doi.org/10.1111/mec.15725

Nichols, R.V., Vollmers, C., Newsom, L.A., Wang, Y., Heintzman, P.D., Leighton, M., ... Shapiro, B. (2018, September). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, *18*(5), 927–939, https://doi.org/10.1111/1755-0998.12895

Pawluczyk, M., Weiss, J., Links, M.G., Egaña Aranguren, M., Wilkinson, M.D., Egea-Cortines, M. (2015, March). Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Analytical and Bioanalytical Chemistry*, *407*(7), 1841–1848, https://doi.org/10.1007/s00216-014-8435-y

Piñol, J., Mir, G., Gomez-Polo, P., Agustí, N. (2015, July). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, *15*(4), 819–830, https://doi.org/10.1111/1755-0998.12355

Pompanon, F., Deagle, B.E., Symondson, W.O.C., Brown, D.S., Jarman, S.N., Taberlet, P. (2012, April). Who is eating what: diet assessment using next generation sequencing: NGS DIET ANALYSIS. *Molecular Ecology*, *21*(8), 1931–1950, https://doi.org/10.1111/j.1365-294X.2011.05403.x

Pornon, A., Escaravage, N., Burrus, M., Holota, H., Khimoun, A., Mariette, J., ... Andalo, C. (2016, June). Using metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific Reports*, *6*(1), 27282, https://doi.org/10.1038/srep27282

Sakamoto, W., & Takami, T. (2018, June). Chloroplast DNA Dynamics: Copy Number, Quality Control and Degradation. *Plant and Cell Physiology*, *59*(6), 1120–1127, https://doi.org/10.1093/pcp/pcy084

Shelton, A.O., Gold, Z.J., Jensen, A.J., D'Agnese, E., Andruszkiewicz Allan, E., Van Cise, A., ... Kelly, R.P. (2022, November). Toward quantitative metabarcoding. *Ecology*, , https://doi.org/10.1002/ecy.3906

Sidstedt, M., Rådström, P., Hedman, J. (2020, April). PCR inhibition in qPCR, dPCR and MPS—mechanisms and solutions. *Analytical and Bioanalytical Chemistry*, *412*(9), 2009–2023, https://doi.org/10.1007/s00216-020-02490-2

Siljak-Yakovlev, S., Pustahija, F., Oli, E.M., Boguni, F., Muratovi, E., Bai, N., . . . Brown, S.C. (2010). Towards a Genome Size and Chromosome Number Database of Balkan Flora: C-Values in 343 Taxa with Novel Values for 242. *Advanced science letters*, *3*(2), 190–213, https://doi.org/10.1166/asl.2010.1115 Retrieved from https://www.ingentaconnect.com/content/asp/asl/2010/00000003/00000002/art000

Silverman, J.D., Bloom, R.J., Jiang, S., Durand, H.K., Dallow, E., Mukherjee, S., David, L.A. (2021, July). Measuring and mitigating PCR bias in microbiota datasets. *PLOS Computational Biology*, *17*(7), e1009113, https://doi.org/10.1371/journal.pcbi.1009113

Smets, W., Leff, J.W., Bradford, M.A., McCulley, R.L., Lebeer, S., Fierer, N. (2016). A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biology and Biochemistry*, *96*, 145–151, https://doi.org/10.7287/peerj.preprints.1318v1 Retrieved from https://peerj.com/preprints/1318

Svec, D., Tichopad, A., Novosadova, V., Pfaffl, M.W., Kubista, M. (2015, March). How good is a PCR efficiency estimate: Recommendations for precise and robust qPCR efficiency assessments. *Biomolecular Detection and Quantification*, *3*, 9–16, https://doi.org/10.1016/j.bdq.2015.01.005

Taberlet, P., Bonin, A., Zinger, L., Coissac, E. (2018). *Environmental DNA* (Vol. 1). Oxford University Press.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., Willerslev, E. (2012, April). Towards next-generation biodiversity assessment using DNA metabarcoding: NEXT-GENERATION 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, January). 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

Thomas, A.C., Deagle, B.E., Eveson, J.P., Harsch, C.H., Trites, A.W. (2016, May). Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, *16*(3), 714–726, https://doi.org/10.1111/1755-0998.12490 Retrieved 2023-06-16, from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.12490

Ushio, M., Murakami, H., Masuda, R., Sado, T., Miya, M., Sakurai, S., . . . Kondoh, M. (2018). Quantitative monitoring of multispecies fish environmental dna using high-throughput sequencing. *Metabarcoding and Metagenomics*, *2*, e23297, https://doi.org/10.3897/mbmg.2.23297 Retrieved from https://doi.org/10.3897/mbmg.2.23297 https://arxiv.org/abs/https://doi.org/10.3897/mbmg.2.23297

Valentini, A., Miquel, C., Nawaz, M.A., Bellemain, E., Coissac, E., Pompanon, F., . . . Taberlet, P. (2009). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trn*L approach. *Molecular Ecology Resources*, *9*, 51–60,

van der Loos, L.M., & Nijland, R. (2021, July). Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecology*, *30*(13), 3270–3288, https://doi.org/10.1111/mec.15592 Retrieved 2022-08-31, from https://onlinelibrary.wiley.com/doi/10.1111/mec.15592

Wilder, M.L., Farrell, J.M., Green, H.C. (2023, March). Estimating edna shedding and decay rates for muskellunge in early stages of development. *Environmental DNA*, *5*(2), 251–263, https://doi.org/10.1002/edn3.349

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M.E., . . . Taberlet, P. (2014, February). Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, *506*(7486), 47–51, https://doi.org/10.1038/nature12921

Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., Ding, Z., . . . Yu, D.W. (2021, July). Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods in Ecology and Evolution*, *12*(7), 1252–1264, https://doi.org/10.1111/2041-210X.13602

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

Zoschke, R., Liere, K., Börner, T. (2007, April). From seedling to mature plant: Arabidopsis plastidial genome copy number, RNA accumulation and transcription are differentially regulated during leaf development: Plastome copy number in Arabidopsis leaf development. *The Plant Journal*, *50*(4), 710–722, https://doi.org/10.1111/j.1365-313X.2007.03084.x