Piñol, Josep; Senar Rosell, Miquel; Symondson, William O.C. «The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative». *Molecular ecology*, Vol. 28, Num. 2 (January 2019), p. 407-419 DOI 10.1111/mec.14776, PMID 29939447

The choice of universal primers and the characteristics of the species mixture determines when DNA metabarcoding can be quantitative.

Josep Piñol[1,2], Miquel A Senar[1], William O C Symondson[3]

(1) Univ. Autònoma Barcelona, Cerdanyola del Vallès, 08193, Spain

(2) CREAF, Cerdanyola del Vallès, 08193, Spain

(3) Cardiff School of Biosciences, Cardiff University, Sir Martin Evans Building, Museum Avenue, Cardiff CF10 3AX, UK

Keywords: COI, diet analysis, environmental DNA, insects, *in silico* PCR, primer bias

Corresponding Author:

Josep Piñol

CREAF, Univ. Autònoma Barcelona, Cerdanyola del Vallès 08193, Spain

E-mail: Josep.Pinol@uab.es

Running title: Primers for quantitative metabarcoding

Abstract

DNA metabarcoding is a technique used to survey biodiversity in many ecological settings, but there are doubts about whether it can provide quantitative results, *i.e.* the proportions of each species in the mixture as opposed to a species list. While there are several experimental studies that report quantitative metabarcoding results, there are a similar number that fail to do so. Here we provide the rationale to understand under what circumstances the technique can be quantitative.

Basically, we simulate a mixture of DNA of *S* species with a defined initial abundance distribution. In the simulated PCR, each species increases its concentration following a certain amplification efficiency. The final DNA concentration will reflect the initial one when the efficiency is similar for all species; otherwise, the initial and final DNA concentrations would be poorly related. Although there are many known factors that modulate amplification efficiency, we focused on the number of primer-template mismatches, arguably the most important one. We used 15 common primers pairs targeting the mitochondrial COI region and the mitogenomes of *ca*. 1200 insect species.

The results showed that some primers pairs produced quantitative results under most circumstances, whereas some other primers failed to do so. Many species, and a high diversity within the mixture, helped the metabarcoding to be quantitative. In conclusion, depending on the primer pair used in the PCR amplification and on the characteristics of the mixture analysed (i.e., high species richness, low evenness), DNA metabarcoding can provide a quantitative estimate of the relative abundances of different species.

Introduction

Ideally, metabarcoding should be a technique used to quantify species abundance in natural communities (C_i in Figure 1) using high-throughput DNA sequencing (HTS). This is normally accomplished by sampling the organisms in the community using a particular sampling method (S_i; Morinière et al., 2016) or by collecting fragments of DNA shed from organisms (environmental DNA or eDNA, E_i; Taberlet, Coissac, Hajibababei, & Rieseberg, 2012). The target can also be the subset of the community consumed by a predator or an herbivore in what is termed diet analysis (D_i; Pompanon et al., 2012); the diet is estimated from the DNA remains in faecal samples or in the gut contents of the consumer (G_i). In all cases, the DNA is extracted into a solution with DNA of many species at relative abundances O_i. Then, the extracted DNA can be directly sequenced (shotgun metagenomics) or sequenced following amplification via PCR of one or more genomic regions (amplicon metabarcoding). Finally, the obtained DNA reads R_i are assigned to species or OTUs (F_i). Every process described in Figure 1 introduces its own biases (Leray & Knowlton, 2017; Pompanon et al., 2012), and so the estimation of the community composition C_i (or D_i in diet analysis) from the final read abundance (F_i) is a daunting task that we are now just beginning to grasp (Barnes & Turner, 2016; Elbrecht, Vamos, Meissner, Aroviita, & Leese, 2017). Only when all biases are avoided or corrected, will it be possible to perform quantitative metabarcoding.

The processes involved in the transformation of species counts or biomass (C_i or D_i) to the DNA solution (O_i) are complex. For instance, in the diet analysis, not all the DNA of the consumed species (or even of different tissues of the same species) is digested with the same efficiency. The case of environmental DNA is even worse, as there are many factors that affect the production and stability of eDNA (origin, state, decay, transport, persistence; Barnes & Turner, 2016). The extraction of DNA from samples (G_i, E_i or S_i) to the solution (O_i) would apparently be straightforward, but this is far from true for some organisms: Pornon et al. (2016) report a difference of *ca* 300 times in the extracted DNA yield (before amplification) from the same number of pollen grains of three plant species. These authors attribute this variability to interspecific differences in pollen wall structure, pollen size, genome size, the number of marker copies and DNA extraction efficiency.

The processes leading from the extracted DNA (O_i) to the relative species abundance (R_i) are no better. Amplicon metabarcoding (shaded region in Figure 1) is mostly affected by the PCR amplification step using 'universal' primers targeting a certain region of the genome. Universal primers do not perfectly match the DNA of all species, and so there is a variable number of template-primer mismatches across species. Consequently, some species are better amplified than others and the proportions in the final mixture do not reflect the original proportion of each species (Elbrecht and Leese, 2015; Leray et al., 2013; Bista et al., 2018). There are other complications in the PCR step that produce more biases; for instance, the use of indexed PCR primers (used to minimize the per sample cost of sequencing by allowing the sequencing of many samples in a single run) might induce further biases (Leray & Knowlton, 2017; O'Donnell, Kelly, Lowell, & Port, 2016). The avoidance of the PCR step (shotgun metagenomics) would in theory render a faithful list of R_i (Bista et al., 2018), and this is what is mostly used in microbial metabarcoding nowadays (Jovel et al., 2016). However, in eukaryotes, the scarcity of assembled genomes and the vast amount of sequencing depth needed, makes shotgun metabarcoding still unsuitable in most circumstances (Gómez-Rodríguez, Crampton-Platt, Timmermans, Baselga, & Vogler, 2015; Zhou et al., 2013).

Whether the metabarcoding provides quantitative results has been usually evaluated using mock communities of known composition that are amplified and sequenced, or using a classical quantification method alongside the DNA metabarcoding. There is a growing number of these studies and the results are contradictory (Table 1). Whilst many studies report a significant quantification, albeit with a variable explanatory power, many others do not. According to these results, the right question to ask is not whether, but in which circumstances, is DNA metabarcoding quantitative.

Here we do not attempt to tackle all the problems in quantitative metabarcoding depicted in Figure 1, but just a subset of them. We concentrate on the template-primer bias that complicates the quantification of the initial DNA concentration in a heterogeneous solution (O_i)

from the reads obtained after PCR amplification and HTS sequencing (R_i). We focus on this for two reasons. First, the process $O_i \rightarrow R_i$ is an obligatory step for diet (G_i), eDNA (E_i), and fresh or well conserved sample analyses (S_i), and so our contribution can potentially benefit people in several fields. Second, whereas there are several causes that influence the number of reads R_i (i.e. genome size, mitochondrial copy number, …), the number of template-primer mismatches is probably the most important one (Elbrecht & Leese, 2015; Mao, Zhou, Chen, & Quan, 2012; Pinto & Raskin, 2012; Piñol, Mir, Gomez-Polo, & Agustí, 2015). We address the problem using a simple model that simulates the process of PCR amplification in heterogeneous mixtures. We test the model using the mitogenomes of *ca.* 1200 species of insects available in RefSeq and 15 primer pairs targeting the COI region (Elbrecht & Leese, 2017a). Maybe the COI region is not the best suited for designing metabarcoding primers (Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014; Elbrecht et al., 2016), but it remains the region with most extensive information in genomic databases. The objectives are to evaluate *in silico* which primer pairs and which characteristics of species mixtures provide a quantitative relationship between the pre- and post-PCR marker abundance of the species in the mixture.

Material and Methods

Rationale of the model

Let's consider a mixture of DNA of S species each with original DNA concentration O_i that is PCR-amplified with a universal primer pair. Each species increases its concentration to F_i according to a certain efficiency Λ_i (here we are assuming that R_i in Figure 1 equals F_i, and so the biases in the bioinformatic pipeline from read number to species abundance are assumed to be negligible in the present application),

$$F_i = \Lambda_i \cdot O_i \qquad \text{Eq. 1}$$

Λ_i varies between 1 (no amplification) and 2^c (maximum amplification, with c the number of PCR cycles). F_i will be proportional to O_i when Λ_i is the same for all species; on the other hand, when Λ_i is very different among species, F_i would poorly reflect O_i. This model is equivalent to the basic one of Suzuki and Giovannoni (1996).

In DNA metabarcoding, only the relative proportions of each species in the mixture are of interest. Let's call o_i and f_i the original and final relative concentration of DNA of species *i* in the mixture. These two magnitudes are related by an equation like the previous one

$$f_i = \lambda_i \cdot o_i / a \qquad \text{Eq. 2}$$

where $a = \sum_{i=1}^{S} \lambda_i \cdot o_i$ is a scaling constant to assure that $1 = \sum_{i=1}^{S} f_i$. Here λ_i is the relative amplification efficiency of species *i* and belongs to the interval (0, 1].

The application of the model is straightforward. First, it requires a pool of species and the primer pairs of interest. Second, a method to generate random mixtures of species with a certain initial abundance distribution (o_i). Third, an estimation of the amplification efficiency λ_i of each species in the mixture. Finally, the computation of f_i using equation 2. Figure 2 summarizes the computational pipeline that implements the model above. On its left-hand side, there are the procedures that calculate the template-primer mismatches for each combination of the species and primer pairs in the pool. In the right-hand side, there is the algorithm that performs many simulations for each primer pair and compares o_i with f_i.

Primer pairs and species

We only considered the 15 COI primer pairs targeting the mitochondrial Folmer region (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994) analysed by Elbrecht and Leese (2017a) (Tables 2 and 3). The selection includes the most common universal primer pairs currently used for DNA metabarcoding of insects. We compiled a pool *P* of 1204 species of insects with an assembled mitochondrial genome at RefSeq (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/ visited at 25-May-2017) (see the distribution of the species among orders in Figure S1).

We calculated the number of primer-template mismatches for each primer and each mitochondrial genome using the function matchPattern of the R package Biostrings (Pagès, Aboyoun, Gentleman, & DebRoy, 2017). We limited the maximum number of mismatches per primer to five. When the primer mapped in more than one region, we choose the region with the lowest number of mismatches.

Next, we combined the pairs of primers in table 3 and computed the total number of primer-mismatches in the two primers and the amplicon length. We only retained the primer pair/genome combinations that produced an amplicon of length equal to the expected amplicon length for each primer pair (table 3). As primer pair #15 produced a very low number of useful species to analyse (table 3), it was not considered further here.

Generation of random communities

A subset of *S* species was randomly sampled from the pool of species *P*. The relative proportions of each species in the mixture was established following a geometric model (Magurran, 2004)

$$o_i = C_k \cdot k \cdot (1-k)^{i-1} \qquad \text{Eq. 3}$$

Where o_i is the proportion of the species *i* in the mixture, *k* is the parameter of the model and $C_k = [1-(1-k)^S]^{-1}$ is a constant that makes $\sum_{i=1}^{S} o_i = 1$. The parameter *k* belongs to the interval (0, 1). Small values of *k* produce communities in which the species have similar abundances, whereas high values of *k* produce communities dominated by a few abundant species.

Estimation of the PCR amplification efficiency

It is acknowledged that the number of template-primer mismatches influences the efficiency of the PCR reaction. However, much less is known about the nature of the relationship between number of mismatches *vs*. efficiency. Here we use a basic model in which each new mismatch reduces the efficiency in a certain proportion β:

$$\lambda_i = \beta^{-m_i} \qquad \text{Eq. 4}$$

Where *m* is the total number of primer-template mismatches in both the forward and the reverse primers. According to this model, the number of mismatches has a multiplicative effect on the amplification efficiency, and so when $m_i = 0$, $\lambda_i = 1$ (perfect match and maximum amplification efficiency); when $m_i = 1$, $\lambda_i = 1/\beta$; when $m_i = 2$, $\lambda_i = 1/\beta^2$; and so on.

In this application, we used two different formulations of the model above. In the first one (model 1), *m* is the number of template-primer mismatches in the entire length of both the forward and reverse primers; in the second one (model 2), *m* is the number of template-primer mismatches that occur only in the five 3'-end positions of both the forward and reverse primers. We have done so because mismatches closer to the 3' end of the primer has a greater effect on the PCR efficiency than mismatches occurring further away from the 3' end of the primer (Stadhouders et al., 2010).

Parameters of the simulation

For this model to be useful, the parameters of the model *S*, *k*, and β must have realistic values, *i.e.* they should correspond to values normally found in natural communities that could eventually be subjected to DNA metabarcoding. The number of species in a sample can be very different depending on the studied community. Here we are pretending to simulate communities of insects in temperate areas, and so a reasonable range for *S* in samples of temperate communities would be 5-100.

To find reasonable values of k, we took advantage of a dataset of insects in tree canopies of a citrus grove (Piñol, Espadaler, & Cañellas, 2012). We used biomass data of species of Dermaptera, Coleoptera, Hemiptera, Neuroptera, Psocoptera, and Hymenoptera from 133 sampling events with 5 or more species. Then we fitted a geometric model (equation 3) to the species biomass for every sampling event. The fitted k values varied between 0.2 and 0.95 (Figure S2a). The goodness of fit of the geometric model to the data was very high (Figure S2b), indicating that the use of a geometric model to describe the relative proportions of the species is sound.

The β value that relates the number of template-primer mismatches to the amplification efficiency was estimated from Piñol et al. (2015), who report a significant negative relationship between the logarithm of the amplification efficiency and the number of template-primer mismatches:

$$log_{10}y = -0.25 + 0.61 \cdot (-m_T), r^2 = 0.73, F_{1,9} = 24.35, P = 0.0008$$

$$log_{10}y = -1.09 + 1.73 \cdot (-m_5), r^2 = 0.81, F_{1,9} = 38.24, P = 0.0002$$

where m_T is the total number of template-primer mismatches in the entire length of the both primers, and m_5 is the total number of mismatches in the five 3'-end nucleotides of the forward and reverse primers. We are not aware of any other study that explicitly states the relationship between the number of template-primer mismatches and the amplification efficiency.

The above empirical relationships are equivalent to equation 4, where β can be estimated as 10^{slope}. Thus, for model 1, $\beta = 10^{0.61} = 4.07$ and for model 2, $\beta = 10^{1.73} = 53.70$. The error in the estimated values of the slope translates to the estimate of β: the 95% interval of confidence for β in model 1 is (2.1 to 7.8) and for model 2 is (12.6 to 234).

Simulations

For each primer pair considered, we ran 10000 simulations using an R script that performed the following steps (summarized in Figure 2):

1. We generated a random number of species *S* in the community between 5 and 100. (The random number and all those below followed a uniform distribution).
2. We randomly chose *S* species from the pool *P* of species.
3. We generated a random value of *k* between 0.2 and 0.95.
4. We used this *k* value to estimate the initial relative DNA concentration o_i of each of the *S* species in the community using equation 3.
5. We generated a random value of β in the interval (2, 8) for model 1 and in the interval (12, 240) for model 2.
6. Using equation 4, we calculated the relative amplification efficiency λ_i of each species in the community using the above β value and the previously calculated number of mismatches between this primer pair and each species. For model 1 we used the number of template-primer mismatches in the entire length of the primers; for model 2 we used only the mismatches occurring in the five 3'-end positions of both primers.
7. Finally, we calculated the relative DNA concentration of each species in the mixture f_i from Equation 2 using the above estimates of λ_i and o_i.
8. Each simulation was summarized by the linear correlation coefficient r_i between f_i and o_i. We also calculated whether the most abundant species at the beginning (o_i) was also the most abundant at the end of the PCR reaction (f_i).

Analysis of the results of the simulations

For each primer pair, we set the following statistical test:

H_0: $r = 0$ (there is not a linear relationship between o_i and f_i)
H_1: $r \neq 0$ (there is a linear relationship between o_i and f_i)

To decide between H_0 and H_1 we considered the empirical 95% confidence interval (CI) of the r_i distribution: when $0 \in$ CI we accepted H_0 and when $0 \notin$ CI we accepted H_1. The probability of error when accepting H_1 is 0.05.

For each set of simulations of each primer pair we also calculated the proportion in which the same species was the most abundant before and after the simulated PCR reaction. If this value was above 0.95, then it would be safe to consider that the observed most abundant species was correctly guessed with a probability of 0.95.

Finally, with all the simulations of all the primer pairs we calculated using a linear model the proportion of the variance of r (after the Fisher z-transformation) associated with the factors: primer pair; S; k; β; and all 2-way interactions.

All the calculations were conducted with R in-house scripts (R Core Team, 2016) also using the database manager SQLite (Müller, Wickham, James, & Falcon, 2017).

Results

Number of mismatches

Considering the entire length of both the forward and the reverse primers, the median of the number of template-primer mismatches was 0 for primer pairs #10 to #14, 1 for primer pair #7, and 3 or higher for the rest of primer pairs (Figure 3A). When only the five 3'-end positions of each primer were considered, the median of the number of template-primer mismatches was 2 for primer pair #4, 1 for primer pairs #2 and #5, and 0 for the remainder (Figure 3B). Primer pairs #10 and #14 were especially good, as more than 99% of the tested species (~1150) had no template-primer mismatches in the five 3'-terminal positions.

Relationship o_i-f_i

Considering the entire length of both the forward and the reverse primers, the simulations of primer pairs #1, #2, #3, and #5 generated an empirical 95% confidence interval (CI) for the linear correlation coefficient that included the 0 value, indicating that it is not justified to assume a significant linear relationship between o_i and f_i (Figure 4A). The opposite was true for the rest of primer pairs. The relationship o_i-f_i was especially good for primers pairs #10 to #14, and to a lesser extent to primer pair #7; for all these primer pairs, it is safe to assume that the final concentration of DNA after the PCR reaction (f_i) quantitatively reflects what was there initially (o_i). However, for none of the primer pairs analysed it is safe to assume that the most abundant species after the PCR reaction was the most abundant initially (Table 4).

The overall picture was slightly better when only the five 3'-terminal bases of both primers were considered (Model 2). In this case, only primer pairs #4 and #5 generated a CI for r that included the 0 value, while the opposite was true for the rest of them (Figure 4B). Primer pairs #10, #11, #12, and #14 where again especially good, generating CI that were always above the value of r=0.9. In addition, for primers pairs #10 to #14, it is safe to assume that the most abundant species was correctly attributed (Table 4).

Effect of the characteristics of the mixture of species on the correlation o_i-f_i

The mixture of the species is characterised in the model by S and k. When the number of species S in the random sample was low (5-15) the relationship o_i-f_i was not significant for all the primers pairs except #10 to #14 (Model 1; Figure S3-A) and for #10 to #14, #3, #6 and #8 (Model 2; Figure S3-B). When S was high (51-100) all primer pairs produced a significant linear correlation between o_i-f_i for both models (Figure S3-CD).

Low values of the parameter k of the geometric distribution (i.e., species with not very different abundances; $k < 0.45$) produced worst results, especially for Model 1, that when k was higher ($k > 0.70$), where only primer pair #5 (Model 1) and #4 and #5 (Model 2) had a 95% CI that included the 0 value (Figure S4).

Relative importance of each factor on the magnitude of the correlation o_i-f_i

We decomposed the variance of the correlation coefficient r in the 140 000 runs (14 primers pairs x 10000 runs each) according to the factors considered in Model 1 and 2 (Table 5). For both models 1 and 2, the main factor was the choice of primer pair that accounts for more 20% of the total variance; k (models 1 and 2) and β (model 1) also had some importance in the decomposition, but not S. However, in both models most of the variance was unexplained by the considered factors. This implies that there are more important reasons on top of those considered above that affect the correlation coefficient r. In the model, the main reason is the idiosyncratic species composition of each simulated mixture; this means that two simulations with identical S, k and β values, but with a different choice of species will likely produce a very different value of r.

The primer pairs that do better (Figure 4) are those with fewer template-primer mismatches (Figure 5-AB). Indeed, the mean number of mismatches per primer is linearly correlated with the mean r of the simulations for both models. However, and following the rationale of the model, the mean r was even better correlated with the standard deviation of the number of mismatches per primer (Figure 5-CD). Consequently, a proxy for the potential of a certain primer pair for conducting quantitative metabarcoding would be the mean, or even better, the standard deviation, of the template-primer mismatches of that primer within the pool of the genomes of interest.

Discussion

The model is intended to establish whether the results of a part of the metabarcoding analysis, but not of the entire metabarcoding pipeline (Figure 1), are *likely* to be quantitative. By quantitative we more precisely mean that there exists a significant linear correlation (at a certain significance level) between the relative DNA concentration before and after the PCR reaction (O_i and F_i in Figure 1) using a particular primer set and a group of organisms. What the model does not provide is the *certainty* of a significant relationship for a given analysis. This approach may have greater utility for the analysis of eDNA samples or community DNA than for gut content analyses, given the additional sources of error associated with digestion. Also, the number of different species in the mix is likely to be much lower in predator/herbivore gut samples than in eDNA samples.

It is also important to realize that the model only considers one of the many factors that affect the PCR amplification efficiency, i.e. the number of template-primer mismatches. Among the non-considered factors there are variable mtDNA copy number, the genome size, the position and type of the mismatches (Stadhouders et al., 2010), and the G+C content of the amplicon (Wintzingerode, Göbel, & Stackebrandt, 1997). However, we are aware of only one study that explicitly correlated the number of mismatches with the amplification efficiency. In that study, the variance explained by the number of mismatches was ~0.75 of the total (Piñol et al., 2015), so there is only a mere ~0.25 of the total variance in the amplification efficiency left to be explained by the rest of the unaccounted factors mentioned above. Thus, this model makes the strong assumption that the PCR amplification efficiency mainly depends on the number of template-primer mismatches, and considering the empirical information available so far, it is a reasonable assumption.

When summarizing which factors most affect the correlation between the pre- and post-PCR DNA concentrations (o_i-f_i) it stands out that the most important of them all was not included in the model (i.e., the unexplained variance in Table 5). This unexplained variance is the idiosyncratic species composition of the mixture and their relative abundances. The following example shows that there can be huge effects even when all the model parameters are the same. In the mixture of table 6a (S = 10, k = 0.4, β = 4) the linear correlation o_i-f_i is highly significant (r = 0.996, P < 0.001). In the mixture in table 6b, the species (and the parameters of the model) are the same as in 6a, but now the most abundant species is #8 instead of #1;

in this case the relationship o_i-f_i becomes non-significant (r = 0.46, P > 0.05). This example shows that it is impossible to be sure that a particular metabarcoding analysis will produce a significant o_i-f_i correlation, unless we know in advance the exact composition of the species in the mixture. For this reason, we highlighted above that our analysis can only provide the *likelihood* of the PCR step of a certain metabarcoding experiment being quantitative, but never its *certainty*.

Table 5 also shows that the factors considered in the model are also important. Below we discuss the importance of the selected primer sets and the macroscopic characteristics of the mixture, i.e. the species richness S and the slope (k) of their relative abundances.

Choice of primer pairs

The choice of primer pairs is the most important decision to make for DNA metabarcoding. The model suggests that some of the primers tested in this study are better suited than others for quantitative DNA metabarcoding. Among the primer pairs tested here, the best choice seems to be the primer pair #10 (Gibson et al., 2014) and the primer pairs #11 to #14 (Elbrecht & Leese, 2017a). All of them guarantee (with a probability of 0.95) a significant linear relationship o_i-f_i; moreover, the linear correlation coefficient between pre- and post-PCR DNA concentrations is likely to be high (in the range 0.4 – 1) (Figure 4). The rest of the primer pairs, except #4 and #5, also provide significant results, but with lower r values. Primers sets #10 to #14 are highly degenerated, so, it is justified that they amplify better *in silico* than other sets with a much lower degeneracy (e.g., #4 and #5).

The primer pair #10 (Gibson et al., 2014) was developed to amplify a 310 bp region of many families of arthropods. This primer contains the universal base inosine (I); in our calculations, we considered that inosine could pair any base, but, in reality, its capacity to pair with the four bases is variable (Martin, Castro, Aboul-ela, & Tinoco, 1985); besides, the use of inosine increases the price of the primers. The four primer pairs of Elbrecht and Leese (2017a) are all possible combinations of two forward and two reverse primers that produce amplicons of different length. As these primers pairs were developed recently, they have been hardly used by other researchers for DNA metabarcoding (Krehenwinkel et al., 2017); considering our results, we would recommend their use. In any case, *in vivo* validations of primers sets with mock samples of the species of interest is still advisable before embarking on a metabarcoding study.

The above recommendation does not imply that the rest of the primers are of no use in metabarcoding. Some of them can have great coverage of some groups, like primers #5 that are particularly good for Lepidoptera (Zeale, Butlin, Barker, Lees, & Jones, 2011) and have been used with profit for the characterisation of the diet of bats (Clare, Symondson, & Fenton, 2014).

Characteristics of the mixture of species

The number of species in the mixture and their relative abundance were also important in the quantification of the species. The number of species S does not explain the magnitude of r (Table 5) but affect the width of the CI of r (Figure S3). When S is low, most primer pairs (#1 to #9) give a CI of r that includes the value r=0; on the contrary, when S is high all tested primer pairs guarantee (at the 95% level) a quantitative metabarcoding. This result is a consequence of the higher effect that an outlier (e.g., one species with one of more mismatches, but very abundant initially, in an assemblage where most species have no mismatches) has on r when S is low that when S is high. Thus, as a rule, the higher the number of species in the mixture, the higher the likelihood of the results reflecting the original relative abundance of the species. These results have a relevant corollary for diet analysis: DNA metabarcoding is more likely to provide a quantitative diet for polyphagous than for stenophagous predators. This implies that it would be more quantitative when analysing polyphagous species in diverse tropical ecosystems than in less diverse temperate ecosystems. Thus, good dietary quantification would be expected, for example, for larger predators eating many small prey (e.g. an insectivorous bird or bat) than for a small predator (e.g. an insect) that may be polyphagous but have few prey in its guts at any moment in time.

The relative abundance of the species in the mixture also affects the quantification of species by metabarcoding. When the relative abundance of the species is similar among them (k low) the method was less reliable that when a few species were very abundant and the rest were not (k high) (Figure S4). This behaviour is easy to understand by observing equation 2 that describes the PCR reaction. The linear correlation o_i-f_i is going to be higher when the variance of o_i (in relation to λ_i) is also high.

Relationship between the number of mismatches and amplification efficiency

Here we considered that all mismatches have the same importance and that each new mismatch reduces the amplification efficiency in the same factor β (equation 4). However, there are other characteristics of the mismatches, besides their total number, that affect the amplification efficiency.

It is known that mismatches near the 3'-end of the primer have a higher effect than in other positions of the primer (Bru, Martin-Laurent, & Philippot, 2008; Stadhouders et al., 2010). We partially took into account this effect by using two versions of the model, one considering all mismatches in both primers (model 1) and one considering only the mismatches in the five 3'-terminal positions of both primers (model 2). The results produced similar conclusions with both versions of the model regarding which primers produced better quantitative results.

It is also known that some types of mismatch reduce more than others the amplification efficiency (Kwok et al., 1990; Stadhouders et al., 2010; Wright et al., 2014). In general, it has been reported a general purine-purine > pyrimidine-pyrimidine > purine-pyrimidine hierarchy of mismatch impact (Stadhouders et al., 2010), but there are some discrepancies. In addition, most of the studies refer only to the 4-5 bases in the 3'-end of the primers, and very little is known about mismatches in the rest of the primer positions (Sipos et al., 2007).

Considering that there is not enough quantitative information about the effect of the mismatch position and type throughout the entire length of the primer, we preferred to keep the model simple. More experimental work in this respect would be needed to parametrise with confidence more realistic models of amplification efficiency. It is worth mentioning that other models already consider the position, adjacency and type of the mismatches (Elbrecht & Leese, 2017b), but their parametrisation is limited as it is based on the scarce empirical information available.

If proved robust, the assumption that the amplification efficiency depends basically on the number of template-mismatches suggests a possible avenue for quantifying mixtures amplified with any primer set. Once (or when) the species composition of the mixture is known, and given the number of primer-template mismatches, it would be possible to estimate the initial abundance of each species (o_i) using equation 2 in reverse. Thus, it should be possible, at least in theory, to quantify the relative composition of any mixture in two steps: the first one would provide the list of species and the second one the relative abundance of each one.

Limitations of the model

The model was applied to approximately 1200 species of insects with a sequenced mitogenome in RefSeq. The model says nothing about other genomic regions, groups of organisms, or sets of primers. For instance, it could be perfectly possible that some of the primers that did not perform well in our analysis, behave much better for a subset of insect orders. However, it is fair to suppose that the same kind of conclusion would be obtained elsewhere: some primer pairs would do better than others, mixtures with more species would do better than mixtures with fewer species, and mixtures with less evenness would also do better than mixtures with a higher evenness. So, it would be worthwhile to conduct similar studies to the present one using different primers and relevant groups of organisms before embarking on metabarcoding experiments.

The model has implicitly assumed that all species in the mixture are amplified to some extent in the simulated PCR. This assumption is at odds with the fact that all primers fail to amplify some species (Brandon-Mog et al., 2015; Mao et al., 2012). This is of little importance in our approach. If some species fail to amplify its final concentration would be $f_i = 0$; in our model f_i would be a very small number, but never 0. However, as we calculate the linear correlation o_i-f_i without any transformation of the raw data, the fact that f_i is 0 or a very small number like, let's say 0.00001, is of minor importance.

It is also important to mention that, implicitly, we considered that the initial DNA concentration was proportional to some measure of abundance, like biomass or individual number, but this is not necessarily the case, especially when multiple-copy markers are used. In plants, there is the added problem of ploidy. Unfortunately, interspecific comparisons using single copy nuclear markers are not usually viable in dietary analysis, as multi-copy targets are needed to amplify the degraded DNA associated with herbivory and predation. In addition, whilst there is some information about gene copy number across taxa in prokaryotes (i.e., 16S rRNA gene; Farrelly, Rainey, & Stackebrandt, 1995), and even ways to use this information for *a posteriori* correction of read numbers (Angly et al., 2014), we are not aware of any reliable data on mtDNA copy number across arthropod species.

Despite the overwhelming complexity of the entire metabarcoding process (Figure 1), the model presented here offers some hope for making the process more quantitative. By simply choosing a primer set with a low variance in the number of mismatches it is possible to obtain greater quantitative accuracy. It is true that other sources of bias remain unchanged, like different digestion rates for DNA from different species, but the results presented here would help to reduce the overall bias.

Acknowledgements

We thank Llorenç Badiella for giving statistical advice and to Simon Creer, Vasco Elbrecht, and two anonymous reviewers for their comments in previous versions of the manuscript. Financial help was provided by the Spanish Government grants TIN2014-53234-C2-1-R and TIN2017-84553-C2-1-R.

Literature

Albaina, A., Aguirre, M., Abad, D., Santos, M., & Estonba, A. (2016). 18S rRNA V9 metabarcoding for diet characterization: a critical evaluation with two sympatric zooplanktivorous fish species. *Ecology and Evolution*, *6*, 1809–1824.

Angly, F.E, Dennis, P.G., Skarshewski, A., Vanwonterghem, I.,Hugenholtz, P., & Tyson, G.W. (2014). CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, *2*, 11.

Barnes, M.A., & Turner, C.R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, *17*, 1-17.

Bista, I., Carvalho, G.R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., ... Creer, S. (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources*, DOI:10.1111/1755-0998.12888.

Blanckenhorn, W.U., Rohner, P.T., Bernasconi, M.V., Haugstetter, J., & Buser, A. (2016). Is qualitative and quantitative metabarcoding of dung fauna biodiversity feasible? *Environmental Toxicology and Chemistry*, *35*, 1970–1977.

Brandon-Mong, G.J., Gan, H.M., Sing, K.W., Lee, P.S., Lim, P.E., & Wilson, J.J. (2015). DNA metabarcoding of insects and allies: an evaluation of primers and pipelines. *Bulletin of Entomological Research*, *105*, 717-727.

Bru, D., Martin-Laurent, F., & Philippot, L. (2008). Quantification of the Detrimental Effect of a Single Primer-Template Mismatch by Real-Time PCR Using the 16S rRNA Gene as an Example. *Applied and Environmental Microbiology*, *74*, 1660–1663.

Clare, E.L., Symondson, W.O., & Fenton, M.B. (2014). An inordinate fondness for beetles? Variation in seasonal dietary preferences of night-roosting big brown bats (*Eptesicus fuscus*). *Molecular Ecology*, *23*, 3633-3647.

Clarke, L.J., Beard, J.M., Swadling, K.M., & Deagle, B.E. (2017). Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution*, *7*, 873–883.

Deagle, B.E., Thomas, A.C., Shaffer, A.K., Trites, A.W., & Jarman, S.N. (2013). Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular Ecology Resources, 13*, 620-633.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, *10*, 20140562.

Dell'Anno, A., Carugati, L., Corinaldesi, C., Riccioni, G., & Danovaro, R. (2015). Unveiling the biodiversity of deep-sea nematodes through metabarcoding: are we ready to bypass the classical taxonomy? *PLoS ONE*, *10*, e0144928.

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*, *10*, e0130324.

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.N., … Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *Peer Journal*, *4*, e1966.

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

Elbrecht, V., & Leese, F. (2017b). PrimerMiner: An R package for development and in silico validation of DNA metabarcoding primers. *Methods in Ecology and Evolution*, *8*, 622-626.

Elbrecht, V., Vamos, E.E., Meissner, K., Aroviita, J., & Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, *8*, 1265-1275.

Evans, N.T., Olds, B.P., Renshaw, M.A., Turner, C.R., Li, Y., Jerde, C.L., … Lodge, D.M. (2016). Quantification of mesocosm fish and amphibian species diversity via DNA metabarcoding. *Molecular Ecology Resources*, *16*, 29-41.

Farrelly, V., Rainey, F.A., & Stackebrandt, E. (1995). Effect of Genome Size and rrn Gene Copy Number on PCR Amplification of 16S rRNA Genes from a Mixture of Bacterial Species. *Applied and Environmental Microbiology*, *61*, 2978-2801.

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*, 294-299.

Geisen, S., Laros, I., Vizcaíno, A., Bonkowski, M., & de Groot, G.A. (2015). Not all are free-living: high-throughput DNA metabarcoding reveals a diverse community of protists parasitizing soil metazoan. *Molecular Ecology*, *24*, 4556-4569.

Geller, J., Meyer, C., Parker, M., & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, *13*, 851–861.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., … Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences*, *111*, 8007–8012.

Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M.J., Baselga, A., & Vogler, A.P. (2015). Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, *6*, 883-894.

Hajibabaei, M., Spall, J.L., Shokralla, S., & van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, *12*, 12-28.

Hänfling, B., Handley, L.L., Read, D.S., Hahn, C., Li, J., Nicholls, P., … Winfield, I.J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, *25*, 3101–3119.

Hawkins, J., de Vere, N., Griffith, A., Ford, C.R., Allainguillaume, J., Hegarty, M.J., … Adams-Groom, B. (2015). Using DNA metabarcoding to identify the floral composition of honey: a new tool for investigating honey bee foraging preferences. *PLoS ONE*, *10*, e0134735.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator. Proceedings of the National Academy of Sciences*, *101*, 14812–534.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., … Wong, G.K.S. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*, *7*, 459.

Klymus, K.E., Marshall, N.T., & Stepien, C.A. (2017). Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. *PLoS ONE*, *12*, e0177643.

Kraaijeveld, K., de Weger, L.A., Ventayol-García, M., Buermans, H., Frank, J., Hiemstra, P.S., & den Dunnen, J.T. (2015). Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Molecular Ecology Resources*, *15*, 8-16.

Krehenwinkel, H., Wolf, M., Lim, J.Y., Rominger, A.J., Simison, W.B., & Gillespie, R.G. (2017) Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, *7*, 17668.

Kwok, S., Kellogg, D.E., McKinney, N., Spasic, D., Godal, L., Levenson, C., & Sninsky, J.J. (1990). Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nucleic Acids Research*, *18*, 999-1005.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., … Machida, R.J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity; application for characterizing coral reef fish gut contents. *Frontiers of Zoology*, *10*, 34.

Leray, M., & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *Peer Journal*, *5*, e3006.

Magurran, A.E. (2004). *Measuring Biological Diversity*. Malden: Blackwell Publishing.

Mao, D.P., Zhou, Q., Chen, C.Y., & Quan, Z.X. (2012). Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC microbiology*, *12*, 66.

Martin, F.H., Castro, M.M., Aboul-ela, F., & Tinoco Jr. I., (1985). Base pairing involving deoxyinosine: implications for probe design. *Nucleic Acids Research*, *13*, 8927-8938.

Meusnier, I., Singer, G.A.C., Landry, J.F., Hickey, D.A., Hebert, P.D.N., & Hajibabaei, M. (2008). A universal mini-barcode for biodiversity analysis. *BMC Genomics*, *9*, 214.

Morinière, J., de Araujo, B.C., Lam, A.W., Hausmann, A., Balke, M., Schmidt, S., … Haszprunar, G. (2016). Species identification in malaise trap samples by DNA barcoding based on NGS technologies and a scoring matrix. *PLoS ONE*, *11*, e0155497.

Müller, K., Wickham, H., James, D.A., & Falcon, S. (2017). RSQLite: 'SQLite' Interface for R. R package version 1.1-2. https://CRAN.R-project.org/package=RSQLite.

Nichols, R.V., Åkesson, M., & Kjellander, P. (2016). Diet Assessment Based on Rumen Contents: A Comparison between DNA Metabarcoding and Macroscopy. *PLoS ONE*, *11*, e0157977.

O'Donnell, J.L., Kelly, R.P., Lowell, N.C., & Port, J.A. (2016). Indexed PCR primers induce template-specific bias in large-scale DNA sequencing studies. *PLoS ONE*, *11*, e0148698.

Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. (2017). Biostrings: Efficient manipulation of biological strings. R package version 2.46.0.

Pinto, A.J., & Raskin, L. (2012). PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. *PLoS ONE*, *7*, e43093.

Piñol, J., Espadaler, X., & Cañellas, N. (2012). Eight years of ant-exclusion from citrus canopies: effects on the arthropod assemblage and on fruit yield. *Agricultural and Forest Entomology*, *14*, 49-57.

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, *15*, 819–830.

Pompanon, F., Deagle, B.E., Symondson, W.O.C., Brown, D.S., Jarman, S.N., & Taberlet, P. (2012). Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, *21*, 1931-1950.

Porazinska, D.L., Giblin-Davis, R.M., Faller, L., Farmerie, W., Kanzaki, N., Morris, K., … Thomas, W.K. (2009). Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Molecular Ecology Resources*, *9*, 1439-1450.

Pornon, A., Escaravage, N., Burrus, M., Holota, H., Khimoun, A., Mariette, J., … Andalo, C. (2016). Using metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific Reports*, *6*, 27282.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Richardson, R.T., Lin, C.H., Sponsler, D.B., Quijia, J.O., Goodell, K., & Johnson, R.M. (2015). Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. *Applications in Plant Sciences*, *3*, 1400066.

Saitoh, S., Aoyama, H., Fujii, S., Sunagawa, H., Nagahama, H., Akutsu, M., … Nakamori, T. (2016). A quantitative protocol for DNA metabarcoding of springtails (Collembola). *Genome*, *59*, 705-723.

Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., … Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, *5*, 9687.

Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K., & Nikolausz, M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on16S rRNA gene-targeting bacterial community analysis. *FEMS Microbiol. Ecol.*, *60*, 341–350.

Stadhouders, R., Pas, S.D., Anber, J., Voermans, J., Mes, T.H., & Schutten, M. (2010). The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5′ nuclease assay. *The Journal of Molecular Diagnostics*, *12*, 109-117.

Suzuki, M.T., & Giovannoni, S.J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, *62*, 625-630.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L.H. (2012). Environmental DNA. *Molecular Ecology*, *21*, 1789-1793.

Thomsen, P.F., Møller, P.R., Sigsgaard, E.E., Knudsen, S.W., Jørgensen, O.A., & Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS ONE*, *11*, e0165252.

Van Houdt, J.K.J., Breman, F.C., Virgilio, M., & De Meyer, M. (2010). Recovering full DNA barcodes from natural history collections of Tephritid fruitflies (Tephritidae, Diptera) using mini barcodes. *Molecular Ecology Resources*, *10*, 459-465.

Wintzingerode, F.V., Göbel, U.B., & Stackebrandt, E. (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews*, *21*, 213-229.

Wright, E.S., Yilmaz, L.S., Ram, S., Gasser, J.M., Harrington, G.W. & Noguera, D.R. (2014). Exploiting extension bias in polymerase chain reaction to improve primer specificity in ensembles of nearly identical DNA templates. *Environmental Microbiology*, *16*, 1354-1365.

Zeale, M.R.K., Butlin, R.K., Barker, G.L.A., Lees, D.C., & Jones, G. (2011). Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources*, *11*, 236–244.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., … Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, *2*, 4.

Authors contribution

JP, MAS, and WOCS designed the study. JP and MAS wrote the code and performed the statistical analyses. All authors played a role in editing the final version of the paper.

Data Accessibility

The mitochondrial genomes and the R scripts used to generate the results are archived in Dryad (doi:10.5061/dryad.q2r3b1f).

Supporting information

Additional supporting information may be found in the online version of this article.

Figure S1. Distribution in orders of the species of insects used in the study.

Figure S2. (A) Histogram of the best fitting *k* values of a geometric distribution.

Figure S3. Effect of the species richness (*S*) on the 95% confidence interval (95% CI) for the Pearson correlation coefficient r.

Figure S4. Effect of the parameter *k* of the geometric distribution on the 95% confidence interval (95% CI) for the Pearson correlation coefficient r.

Table 1. A compilation of experiments attempting to establish whether DNA metabarcoding can be said to be quantitative. The goodness of fit was usually estimated as the Pearson or Spearman squared correlation coefficient; its significance is given as NS (P > 0.05), * (P< 0.05), ** (P<0.01), and *** (P<0.001).

Organisms	Marker	Goodness of fit	Significance	Reference
Eight fish and one amphibian species in mesocosms compared with DNA metabarcoding	cit b, 12S, 16S	0.49-0.88	** to ***	Evans et al. (2016)
Calanoid copepods measured as biomass and by DNA metabarcoding	COI, 16S, 18S	0.26-0.83	NS to ***	Clarke, Beard, Swadling, & Deagle (2017)
Mock community of 41 species of nematodes at variable abundances	LSU, SSU rRNA	Not given	NS	Porazinska et al. (2009)
Analysis of faeces of seals fed with 3 species of fish in known proportions	16S	Not given	NS	Deagle, Thomas, Shaffer, Trites, & Jarman (2013)
Three samples of airborne pollen measured by classical methods and by DNA metabarcoding	trnL	0.23-0.45	***	Kraaijeveld et al. (2015)
Nine samples of pollen assemblages measured by classical methods and by DNA metabarcoding	rbcL	Negative to 0.55	NS to **	Hawkins et al. (2015)
Six samples of pollen assemblages measured by classical methods and by DNA metabarcoding	ITS2, matK, rbcL	Negative to 0.88	NS to *	Richardson et al. (2015)
Marine nematodes identified morphologically and by DNA metabarcoding	18S	Not given	NS	Dell'Anno, Corinaldesi, Riccioni, & Danovaro (2015)
Lake fish assemblages of 16 fish species measured as eDNA and compared with estimates from surveys	12S, cit b	0.05-0.70	NS to ***	Hänfling et al. (2016)
Mock communities of 4 to 9 insect species common in dung fauna in variable proportions	COI	0.01 0.86	NS to *	Blanckenhorn, Rohner, Bernasconi, Haugstetter, & Buser (2016)
Plants in rumen contents measured by DNA metabarcoding and by macroscopic identification	trnL	0.15 – 0.27	**	Nichols, Akesson, & Kjellander (2016)
Natural marine fish assemblages measured as eDNA and as trawl catches. In addition, a mock community of 5 fish species at variable abundances	12S	Natural: 0.10– 0.14 Mock: 0.81	* to ***	Thomsen et al. (2016)
Mock community of an equimolar mix of 12 species of insects and spiders	COI	Not given	NS	Piñol et al. (2015)
Mock community of 8 species of soil protist of 4 different phyla at variable abundances	18S	Not given	NS	Geisen, Laros, Vizcaíno, Bonkowski, & de Groot (2015)
Mock community of 6 species of zooplankton at variable abundances	18S	0.96	**	Albaina, Aguirre, Abad, Santos, and Estonba (2016)
Mock community of 6 species of Collembola at variable abundances	COI, 16S	0.83 – 0.98	***	Saitoh et al. (2016)
Mock community of an equimolar mix of 34 species of aquatic invertebrate belonging to 6 different phyla	COI	Not given	NS	Leray and Knowlton (2017)
Mock community of 10 species of freshwater bivalve and gastropod molluscs at variable abundance	16S	0.79 – 0.92	*	Klymus, Marshall, & Stepien (2017)

Table 2. Universal primers targeting the mitochondrial COI region used in this study.

Name	**Strand**	**Sequence (5 → 3')**	**Reference**
LCO1490	F	GGTCAACAAATCATAAAGATATTGG	Folmer et al. (1994)
HC02198	R	TAAACTTCAGGGTGACCAAAAAATCA	Folmer et al. (1994)
Uni-MinibarR1	R	GAAAATCATAATGAAGGCATGAGC	Meusnier et al. (2008)
Uni-MinibarF1	F	TCCACTAATCACAARGATATTGGTAC	Meusnier et al. (2008)
ZBJ-ArtF1c	F	AGATATTGGAACWTTATATTTTATTTTTGG	Zeale et al. (2010)
ZBJ-ArtR2c	R	WACTAATCAATTWCCAAATCCTCC	Zeale et al. (2010)
mlCOIintF	F	GGWACWGGWTGAACWGTWTAYCCYCC	Leray et al. (2013)
mlCOIintR	R	GGRGGRTASACSGTTCASCCSGTSCC	Leray et al. (2013)
LepF1	F	ATTCAACCAATCATAAAGATATTGG	Hebert, Penton, Burns, Janzen, & Hallwachs (2004)
EPT-long-univR	R	AARAAAATYATAAYAAAIGCGTGIAIIGT	Hajibabaei, Spall, Shokralla, & Konynenburg (2012)
MLepF1-Rev	R	CGTGGAAAWGCTATATCWGGTG	Brandon-Mong et al. (2015)
Ill_C_R	R	GGIGGRTAIACIGTTCAICC	Shokralla et al. (2015)
Ill_B_F	F	CCIGAYATRGCITTYCCICG	Shokralla et al. (2015)
BF1	F	ACWGGWTGRACWGTNTAYCC	Elbrecht and Leese (2017a)
BF2	F	GCHCCHGAYATRGCHTTYCC	Elbrecht and Leese (2017a)
BR1	R	ARYATDGTRATDGCHCCDGC	Elbrecht and Leese (2017a)
BR2	R	TCDGGRTGNCCRAARAAYCA	Elbrecht and Leese (2017a)
ArF5	F	GCICCIGAYATRKCITTYCCICG	Gibson et al. (2014)
ArR5	R	GTRATIGCICCIGCIARIACIGG	Gibson et al. (2014)
jgLCO1490	F	TITCIACIAAYCAYAARGAYATTGG	Geller, Meyer, Parker, & Hawk (2013)
jgHCO2198	R	TAIACYTCIGGRTGICCRAARAAYCA	Geller et al. (2013)
L499	F	ATTAATATACGATCAACAGGAAT	Van Houdt, Breman, Virgilio, & De Meyer (2010)
H2123d	R	TAWACTTCWGGRTGWCCAAARAATCA	Van Houdt et al. (2010)

Table 3. Primer pairs used in this study. Primer names as in table 2. Primer pair #15 was not further considered because it provided much fewer species with useful data than the other 14 primer pairs.

id	Forward Primer	Reverse Primer	Amplicon length (bp)	Number of species with useful data
#1	LCO1490	HC02198	658	1003
#2	LepF1	MLepF1-Rev	218	1035
#3	LepF1	EPT-long-univR	127	1048
#4	Uni-MinibarF1	Uni-MinibarR1	127	800
#5	ZBJ-ArtF1c	ZBJ-ArtR2c	157	937
#6	jgLCO1490	mlCOIintR	319	944
#7	mlCOIintF	jgHCO2198	313	1162
#8	LCO1490	Ill_C_R	325	1014
#9	Ill_B_F	HC02198	418	1143
#10	ArF5	ArR5	310	1157
#11	BF2	BR1	322	1146
#12	BF1	BR2	316	1155
#13	BF2	BR2	421	1157
#14	BF1	BR1	217	1143
#15	L499	H2123d	178	480

Table 4. Proportion of runs in which the same species is most abundant both before and after the simulated PCR reaction for Model 1 and Model 2 simulations. Proportion based on 10000 simulation runs per primer pair. It is indicated in bold face whether it is safe (at $\alpha = 0.95$) to conclude which species is the most abundant in the mixture.

Primer pair	Model 1	Model 2
1	0.42	0.65
2	0.43	0.51
3	0.42	0.94
4	0.49	0.41
5	0.36	0.50
6	0.55	0.85
7	0.63	0.76
8	0.52	0.94
9	0.53	0.67
10	0.94	**1.00**
11	0.88	**0.98**
12	0.81	**0.98**
13	0.91	**0.97**
14	0.79	**1.00**

Table 5. Percentage of the variance in the linear correlation coefficient *r* explained by each parameter involved in the simulations using models 1 and 2.

Factor	Model 1	Model 2
Primer pair (pp)	23.0	20.9
S	0.1	0
k	11.8	2.2
β	2.0	0
pp:S	0	0
pp:k	0.9	0.9
pp:β	0.2	0
S:k	0	0
S:β	0	0
k:β	0	0
unexplained	62.0	76.0

Table 6. Two hypothetical mixtures of ten species, with the number of mismatches of each species, its original and final DNA concentration (o_i and f_i), and its amplification efficiency (β = 4). The mixture B is the same as the mixture A, but for the swap of o_i of species #1 and #8.

A				
#sp	m_{Ti}	o_i	e_i	f_i
1	0	0,403	1,000	0,437
2	0	0,241	1,000	0,261
3	0	0,145	1,000	0,157
4	0	0,087	1,000	0,094
5	1	0,052	0,250	0,014
6	1	0,031	0,250	0,008
7	0	0,019	1,000	0,021
8	2	0,011	0,063	0,001
9	1	0,007	0,250	0,002
10	0	0,004	1,000	0,004

B				
#sp	m_{Ti}	o_i	e_i	f_i
1	0	0,011	1,000	0,020
2	0	0,241	1,000	0,434
3	0	0,145	1,000	0,261
4	0	0,087	1,000	0,157
5	1	0,052	0,250	0,023
6	1	0,031	0,250	0,014
7	0	0,019	1,000	0,034
8	2	**0,403**	0,063	0,045
9	1	0,007	0,250	0,003
10	0	0,004	1,000	0,007

Figure 1. Conceptual diagram of the process of quantitative DNA metabarcoding, from the usual targets (relative abundance of diet components, D_i, or species in a community C_i) to the final assignment of abundances to species (F_i). The sub index *i* indicates the abundance of species *i* in the multispecies mixture. This study covers the process of amplicon metabarcoding ($O_i \rightarrow A_i \rightarrow R_i$).

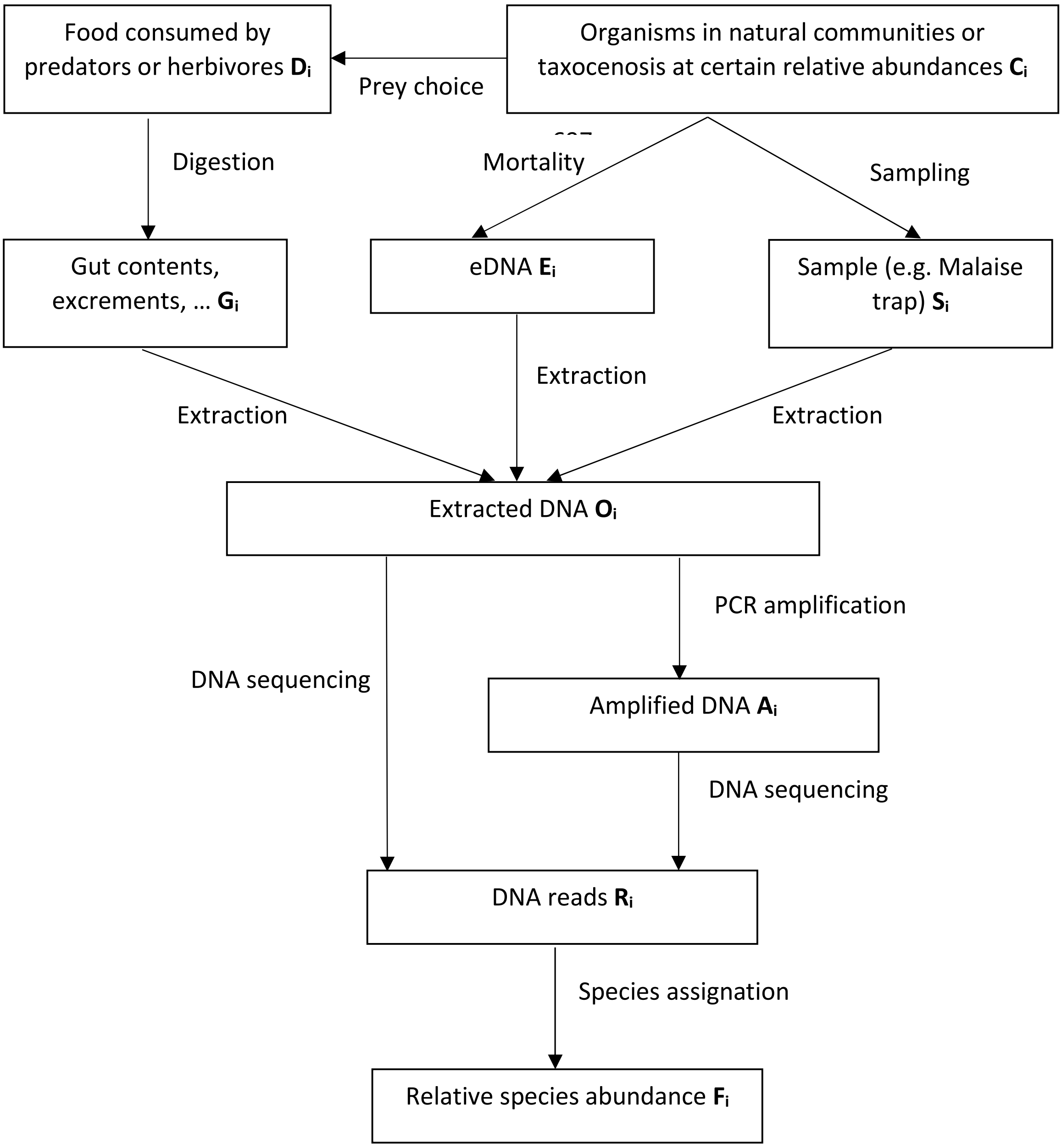

Figure 2. Flow diagram of the pipeline used in this study. On the left-hand side, calculations of template-primer mismatches for each primer pair and genome, both for all the nucleotides in both the forward and reverse primers (m_T; model 1) and only on five 3'-terminal nucleotides (m_5; model 2). On the right-hand side, the algorithm that generates random mixtures of species at random initial abundances (o_i), estimates an amplification efficiency for each species based on the number of template-primer mismatches, and simulates a PCR reaction to produce a final relative abundance of each species (f_i).

. . . . run once . . . . run many times . . . .

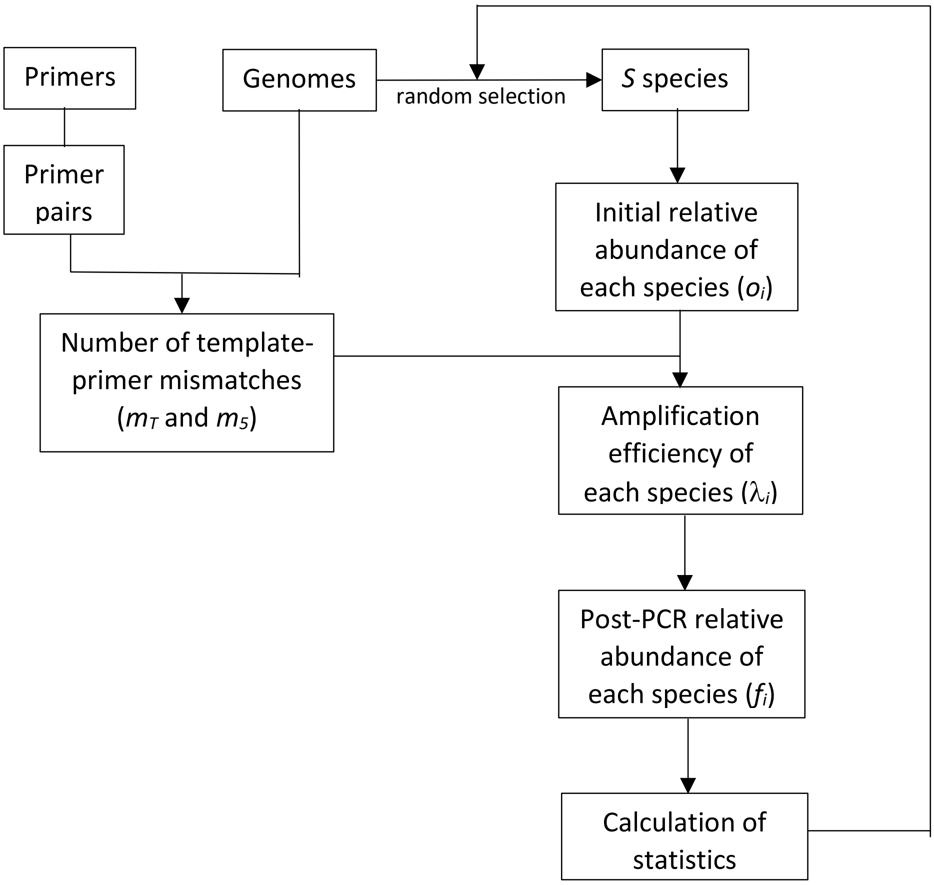

Figure 3. Boxplot of number of template-primer mismatches for each primer pair. (A) Model 1 considers the total number of mismatches in both the forward and reverse primers. (B) Model 2 considers only the mismatches in the five 3'-terminal positions of both primers. Primer pair numbering is the same as in Table 3.

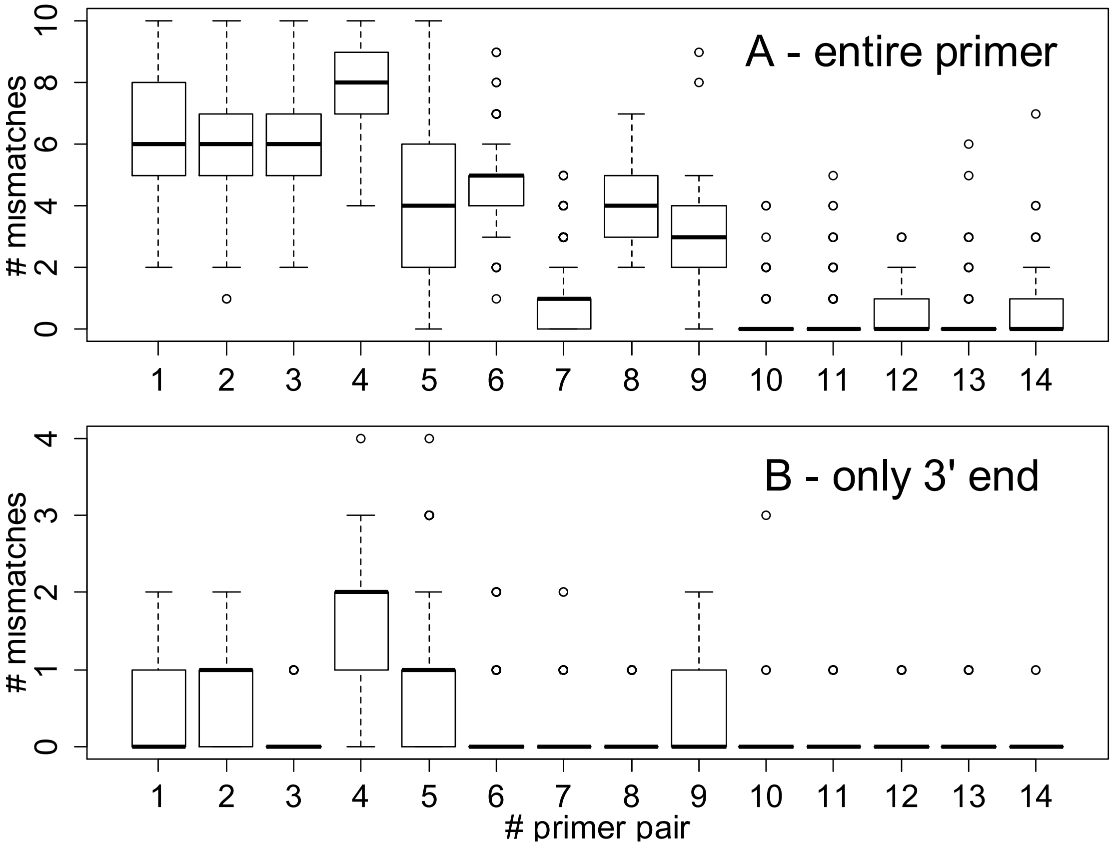

Figure 4. Ninety-five percent confidence interval (95% CI) for the Pearson correlation *r* for each primer pair analysed for Model 1 (A) and Model 2 (B). For primer pairs #10 and #14 in B, the CI is so small that the arrowheads could not be plotted. When the CI cuts the vertical line at *r* = 0 it indicates that it is not possible to consider that *r* > 0 (with a probability of 0.95). Primer pair numbering is the same as in Table 3.

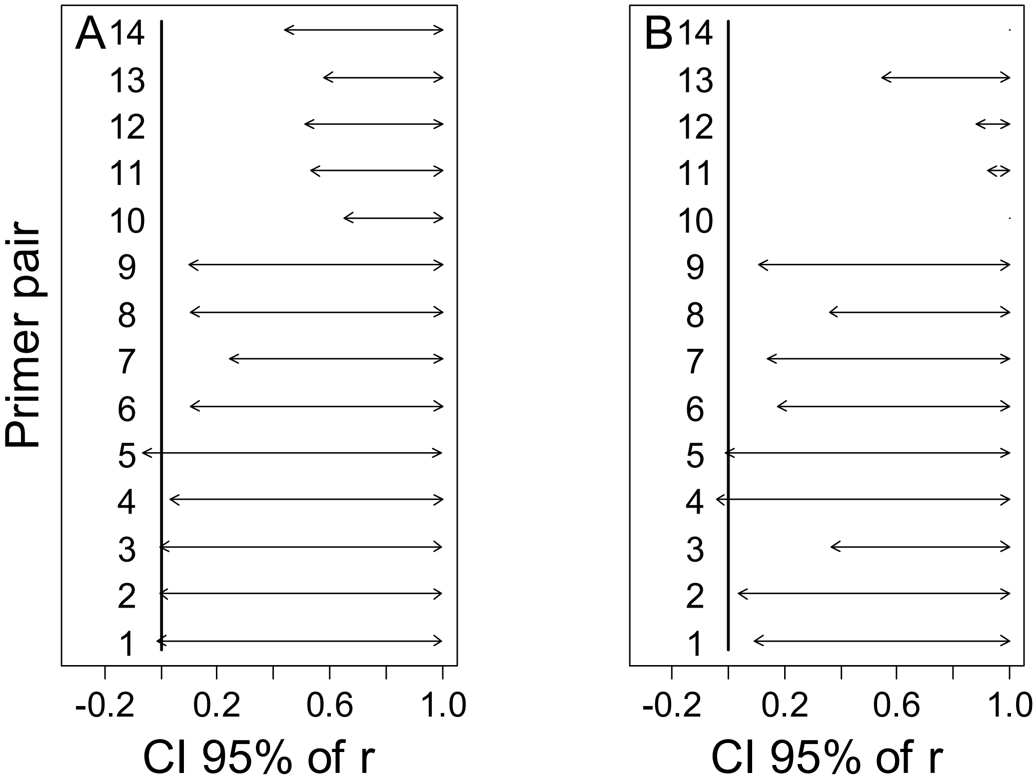

Figure 5. Relationship between the simulated mean of *r* and the mean number of template-primer mismatches (A, B) and the standard deviation of the number of mismatches (C, D) for model 1 (A, C) and model 2 (B, D); m_T = number of mismatches in the entire length of both primers; m_5 = number of mismatches in the five 3'-end positions of both primers.