Received: 23 September 2020 | Revised: 21 June 2021 | Accepted: 7 July 2021

DOI: 10.1111/1755-0998.13464

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Relative species abundance estimation in artificial mixtures of insects using mito-metagenomics and a correction factor for the mitochondrial DNA copy number

Lidia Garrido-Sanz[1] | Miquel Àngel Senar[1] | Josep Piñol[1,2]

[1]Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain

[2]CREAF, Cerdanyola del Vallès, Spain

Correspondence
Lidia Garrido-Sanz, Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain.
Email: Lidia.Garrido@uab.cat

Funding information
Spanish Government grant, Grant/Award Number: TIN2017-84553-C2-1-R; Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Grant/Award Number: 2017-SGR-1001

Abstract

Mito-metagenomics (MMG) is becoming an alternative to amplicon metabarcoding for the assessment of biodiversity in complex biological samples using high-throughput sequencing. Whereas MMG overcomes the biases introduced by the PCR step in the generation of amplicons, it is not yet a technique free of shortcomings. First, as the reads are obtained from shotgun sequencing, a very low proportion of reads map into the mitogenomes, so a high sequencing effort is needed. Second, as the number of mitogenomes per cell can vary among species, the relative species abundance (RSA) in a mixture could be wrongly estimated. Here, we challenge the MMG method to estimate the RSA using artificial libraries of 17 insect species whose complete genomes are available on public repositories. With fresh specimens of these species, we created single-species libraries to calibrate the bioinformatic pipeline and mixed-species libraries to estimate the RSA. Our results showed that the MMG approach confidently recovers the species list of the mixtures, even when they contain congeneric species. The method was also able to estimate the abundance of a species across different samples (*within*-species estimation) but failed to estimate the RSA within a single sample (*across*-species estimation) unless a correction factor accounting for the variable number of mitogenomes per cell was used. To estimate this correction factor, we used the proportion of reads mapping into mitogenomes in the single-species libraries and the lengths of the whole genomes and mitogenomes.

KEYWORDS
Metazoa, mitochondrial genomes, mitogenome skimming, mock sample, next-generation sequencing, PCR-free

1 | INTRODUCTION

Mitochondrial metagenomics or mito-metagenomics (hereafter, MMG) is becoming an alternative to the classical amplicon metabarcoding for the large-scale assessment of biodiversity of Metazoa (Crampton-Platt et al., 2015; Tang et al., 2014; Zhou et al., 2013). MMG consists in the shotgun sequencing of a DNA sample followed by the mapping of the reads to mitochondrial genomes (hereafter, mitogenomes) obtained from online repositories or *ad hoc* assemblages (Crampton-Platt et al., 2016). On the positive side, MMG avoids the amplification biases caused by the PCR step (Elbrecht & Leese, 2015; Piñol et al., 2015; Taberlet et al., 2012). On the negative

side, the tiny size of the mitogenome compared to the nuclear genome produces a high number of non-informative reads (Tang et al., 2014), so a great sequencing depth is needed.

It is generally assumed that MMG quantifies satisfactorily the relative species abundance (hereafter, RSA) of complex mixtures (Gómez-Rodríguez et al., 2015; Zhou et al., 2013). However, as far as we know, there are only five studies that tested the MMG method (plus one using chloroplast metagenomics in plants) using shotgun samples of known composition (Bista et al., 2018; Gómez-Rodríguez et al., 2015; Gueuning et al., 2019; Ji et al., 2020; Lang et al., 2019; Tang et al., 2015). In general, the relationship between the expected and estimated RSA is statistically significant, but with high variability in the goodness of fit.

How the RSA of complex mixtures is presented in the literature requires some clarification. First, the RSA can be expressed as a proportion of the species biomass (e.g., Gueuning et al., 2019) or individual counts (e.g., Lang et al., 2019); alternatively, the RSA can refer to the proportion of the DNA amount of each species in the mixture. Whilst the former approach is more meaningful for most ecological studies, we adopt the latter approach here because it allows the independent evaluation of different sources of bias on the RSA estimation. Second, some studies provide the relative abundance of one species in different samples (e.g., Bista et al., 2018), whereas others report the abundance of several species in a single sample (e.g., Saitoh et al., 2016). Ji et al. (2020) named *within*-species estimation the former (is species *i* more abundant in sample *s* than in sample *r*?) and *across*-species the latter (is species *i* more abundant than species *j* in sample *s*?). This distinction is important because there are species-specific characteristics that influence the *across*-species estimation but not the *within*-species estimation. The most important of these characteristics is the variable number of mitogenomes per nuclear genome (mitochondrial DNA copy number). Thus, a species *i* with twice the number of mitogenomes per nuclear genome than another species *j* will produce twice many mitochondrial reads as well; without a proper correcting factor, species *i* would, apparently, be twice more abundant in the mixture than species *j*. This fact is known (Bista et al., 2018; Piñol et al., 2015; Tang et al., 2014), but there are not reliable solutions to the problem because little is known about the causes of the variation of the mitochondrial copy number *across*-species (but see Liu et al., 2018, that reported a higher mitochondrial copy number in organs with a high metabolic rate and in species living at low altitude than in their counterparts at high altitude in the Tibetan Plateau).

The size of the nuclear genome of the species also affects the *across*-species RSA estimation. Being all other things equal, a species *r* with a nuclear genome half as big as that of another species *s* will produce twice many mitochondrial reads because the mitochondrial DNA is diluted in a smaller amount of nuclear DNA. Therefore, without a proper correcting factor, species *r* would, apparently, be twice more abundant in the mixture than species *s*. The effect of genome size on RSA estimation is also known (Crampton-Platt et al., 2016; Krehenwinkel et al., 2017; Tang et al., 2014), but it is difficult to consider it because measuring the genome size is not an easy task (there is a database of genomes sizes with 1344 insect species on it; Gregory (2020), accessed on 25 March 2020). Both the variation *across*-species of mitochondrial copy number and genome size affect MMG, but also any amplicon metabarcoding method that targets genomic regions with a variable copy number, such as COI in animals (Hebert et al., 2003), ITS in fungi (Schoch et al., 2012), or rbcL + matK in plants (CBoL Plant Working Group, 2009).

Here we explore the quantitative capabilities of MMG for the estimation of RSA of heterogeneous mixtures of insects. For this purpose, we prepared single-species and artificial mixed-species libraries with several species of insects whose entire genome has already been sequenced. We included four species of *Drosophila* to assess the ability of the method to set apart closely related species. The single-species libraries allowed the calculation of a reliable mitochondrial DNA copy number (N_M) for each species that was further used as a correction factor for the *across*-species estimation of RSA of the mixed-species libraries. In particular, we addressed the following questions: (1) Is the MMG method able to identify species in complex mixtures, even when they are of the same genus? As an approach to real samples, we investigated the robustness of the method in the absence of the mitogenome of the focal species. (2) Can MMG estimate the RSA of complex samples? Is it necessary the use of the N_M correction factor for the *across*-species estimation of RSA? (3) Finally, can the number of sequenced reads be reduced and still recover all species in a complex sample of insects?

2 | MATERIALS AND METHODS

2.1 | Selection of species, preparation of the DNA libraries, sequencing, and quality control

We selected 17 species of insects whose complete genome is already sequenced and available on the RefSeq repository (Table 1). Individuals of these species were captured alive or in fly traps from various locations and DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen) from c. 20 mg of fresh material. With the DNA extracts, we prepared two kinds of libraries: 21 single-species libraries and six mixed-species libraries. Four species were sequenced twice in different runs using different extractions to test the repeatability of the method (Table 1). The same extracts used for the first run of single-species libraries were also used to create six artificial mixed-species libraries of 7–8 species at known relative DNA concentrations to test the ability of the method to estimate the RSA (Table S1). In libraries no. 1 and no. 2, the RSA was highly variable (from ~50%, the most abundant, to ~0.4%, the least abundant); in libraries no. 3 and no. 4, it was intermediate (from ~35% to ~3%); and in libraries no. 5 and no. 6, the variability among species was much lower (from ~24% to ~8.5%). As stated in the Introduction, here the RSA is the relative DNA concentration of the species in the mixture, not their relative biomass. Consequently, all sources of variation between the fresh biological material and the extracted DNA

TABLE 1 Summary information of single-species libraries

Run	Library	Species	Order	Family	Cultured/Wild	Origin (country)	Number of raw reads (paired-end)	Number of reads after quality control step (paired-end)	Number of candidate mito-reads (paired-end)
1	1	*Papilio machaon*	Lepidoptera	Papilionidae	Wild	Spain	434,520	432,712	14,774
1	2	*Drosophila virilis*	Diptera	Drosophilidae	Cultured	Spain	4,714,902	4,652,442	152,568
1	3	*Drosophila melanogaster*	Diptera	Drosophilidae	Cultured	Spain	2,283,768	2,275,386	50,944
1	4	*Drosophila mojavensis*	Diptera	Drosophilidae	Cultured	Spain	1,668,424	1,646,524	109,458
1	5	*Bactrocera oleae*	Diptera	Tephritidae	Wild	Spain	580,996	577,720	23,204
1	6	*Linepithema humile*	Hymenoptera	Formicidae	Wild	Spain	1,422,342	1,402,800	73,580
1	7	*Bombus terrestris*	Hymenoptera	Apidae	Wild	Spain	1,994,938	1,987,920	43,872
1	8	*Apis mellifera*	Hymenoptera	Apidae	Wild	Spain	1,262,388	1,236,630	223,202
1	9	*Acyrthosiphon pisum*	Hemiptera	Aphididae	Cultured	USA	684,688	596,972	104,344
2	1	*Atta colombica*	Hymenoptera	Formicidae	Cultured	Denmark	3,272,710	3,267,688	333,064
2	2	*Bemisia tabaci*	Hemiptera	Aleyrodidae	Cultured	Spain	2,513,212	2,501,586	23,096
2	3	*Cimex lectularius*	Hemiptera	Cimicidae	Wild	Spain	3,506,722	3,474,010	122,826
2	4	*Drosophila melanogaster*	Diptera	Drosophilidae	Cultured	Spain	2,909,608	2,903,546	74,984
2	5	*Drosophila mojavensis*	Diptera	Drosophilidae	Cultured	Spain	1,797,470	1,786,926	82,794
2	6	*Drosophila virilis*	Diptera	Drosophilidae	Cultured	Spain	1,336,884	1,332,888	49,146
2	7	*Drosophila suzukii*	Diptera	Drosophilidae	Cultured	Spain	2,510,384	2,501,862	79,498
2	8	*Linepithema humile*	Hymenoptera	Formicidae	Wild	Spain	2,164,404	2,157,154	95,936
2	9	*Plutella xylostella*	Lepidoptera	Plutellidae	Wild	Spain	4,250,124	4,244,328	85,882
2	10	*Solenopsis invicta*	Hymenoptera	Formicidae	Wild	Argentina	3,661,374	3,648,438	146,812
2	11	*Vollenhovia emeryi*	Hymenoptera	Formicidae	Wild	Japan	3,487,834	3,480,802	66,664
2	12	*Wasmannia auropunctata*	Hymenoptera	Formicidae	Wild	Spain	3,335,212	3,293,182	179,000

(e.g., DNA-to-biomass ratio) are ignored (Matesanz et al., 2019; Tang et al., 2015).

All libraries were prepared with the TruSeq DNA PCR-Free LT Kit of Illumina following the manufacturer's instructions (Ref. 15037063) and sequenced using Illumina MiSeq with the 2 × 150 bp chemistry in three different runs. Both the libraries and the sequencing runs were also used in a previous study (Garrido-Sanz et al., 2020). In the parent study, the obtained reads were mapped against a reference database of whole-genomes instead of a reference database of mitogenomes as we do here.

We pre-processed the raw reads through a quality control step using FASTQC v0.11.7 (Andrews, 2015) and TRIMMOMATIC v0.36 (Bolger et al., 2014) to trim the reads to a 150 bp length and to remove those shorter than 140 bp. Only paired reads were kept.

2.2 | Reference genomes

We downloaded all mitogenomes of insects available at the NCBI RefSeq database on 1 August 2019, plus the complete genomes of the 17 species selected for the study from the same database. Ten species had the complete genomes but not the mitogenomes on RefSeq, and we downloaded their mitogenomes from GenBank (accessed on 2 August 2019). We used high-quality mitogenomes of as many species as possible to test the ability of the method to find the selected species among the many others in the reference database. Species with several mitogenomes were deduplicated, and we obtained the mitogenomes of 1794 species of insects, comprising 1174 genera, 331 families, and 27 orders (hereafter, Mito1794; Table S2).

2.3 | Mapping of reads into references

In MMG studies, a small proportion of shotgun reads map into the mitogenome (e.g., Tang et al., 2014), hence most reads are not useful and slow down the mapping process. Thus, it seems reasonable to eliminate the reads that are not mitochondrial before the mapping step (Crampton-Platt et al., 2015, 2016; Zhou et al., 2013). For this purpose, we created a reference database with one mitogenome per family (hereafter, Mito331) where the representative species per family was chosen randomly. We then mapped the raw reads against the Mito331 reference data set using a permissive criterion and kept the putative mitochondrial reads (hereafter, candidate mito-reads). The mapping was done using BWA 0.7.15-r1140 (Li, 2013) with mem algorithm and an alignment score of zero (-T0). SAMTOOLS 1.10 (Li et al., 2009) was subsequently used to filter the paired-end reads with no mapping reads (view -F2316 -b) and recovered the mito-reads in FASTQ format (bam2fq).

As low-complexity regions are prone to misclassify the reads (Lu & Salzberg, 2018; Pearman et al., 2019), we prepared a new set of filtered mitogenome references by removing low-complexity regions from the Mito1794 reference (hereafter, FilteredMito1794). Low-complexity regions were identified using dustmasker (-level 45) (Morgulis et al., 2006) and replaced with Ns using an in-house python script.

To avoid confusion, we recapitulate below the name and meaning of the three different databases of mitochondrial genomes that we used for the mapping of reads:

- Mito1794: the original mitogenomes of 1794 species.
- FilteredMito1794: as Mito1794, but with the low complexity regions removed.
- Mito331: a subset of Mito1794 with only one mitogenome per family; this reference was only used to obtain the candidate mito-reads from the total of reads of each sample.

As we did not know to which extent the filtering of raw reads and mitogenomes was useful, we conducted four different kinds of mapping of reads to reference mitogenomes in the single-species libraries:

- Raw reads against Mito1794 reference database
- Raw reads against FilteredMito1794 reference database
- Candidate mito-reads against Mito1794 reference database
- Candidate mito-reads against FilteredMito1794 reference database

In all cases, the mapping was conducted using BWA with mem algorithm and default options. Because the mapping of a sample was done against each mitogenome individually, we obtained the same number of SAM files as references used. We subsequently used SAMtools to remove reads that did not map to any reference (view -F2308).

2.4 | Assignment of mapped reads to species

In general, a read mapped to several reference mitogenomes (e.g., homologous sequences in several species), so an algorithm was needed to assign reads to species. For this purpose, we used the γ–δ algorithm described in Garrido-Sanz et al. (2020). Briefly, what the γ–δ algorithm does is to quantify the similarity between the query read and the reference (i.e., the mapping ratio) and then decide whether a read is informative or non-informative. It is informative when it is very similar (mapping ratio above γ) to species *i* and not very similar (mapping ratio below δ) to the rest of species; in this case, the read is assigned to species *i*. It can be non-informative for two reasons, either because the read is not similar enough to any species (mapping ratio below γ for all species) or because it is too similar (mapping ratio above δ) to two or more species; in this case, the read is discarded. In all cases $\gamma > \delta$.

The γ–δ algorithm has never been applied before to MMG data, hence the appropriate values of γ and δ are unknown, so the

single-species libraries were used to find the best combination of the parameters γ and δ. The tested values were all the combinations of γ = {0.99, 0.98, 0.97} and δ = {0.98, 0.97, 0.96} provided that γ > δ. To find the best values of γ and δ we relied on the criterion that the number of recovered species had to be one in the single-species libraries.

The reads in single-species libraries were divided into a training set with 75% of the reads randomly selected, and a test set with the remaining 25% of the reads. The training set was used for the calibration of the procedure; the test set was used to assess the goodness of fit of the model and to calculate the summary statistics.

A situation that can arise in real samples, as opposed to the artificial samples used here, is that the mitogenomes of some species in the sample are not in the reference database. We explored this situation by running again the complete pipeline with all the single-species libraries using the best set of input data and parameters. Ideally, no read should be assigned to any species because the mitogenome of the only species in the library is not in the database. However, the reads might eventually be wrongly assigned to other species in the database and, thus, generate false positives. The outcome of this experiment should reveal the robustness of the γ–δ algorithm in the assignment of reads to species.

2.5 | Quantification of the RSA in mixed-species libraries and the need for a species-specific correction factor

In the literature, the relative abundance of one species is sometimes compared among different samples and on other occasions, the relative abundance of several species is compared within one sample (*within*-species and *across*-species RSA, respectively, following Ji et al., 2020). Here, we present the comparison of actual *versus* estimated RSA in the mixed-species libraries using both approaches. As we observed in the Introduction, from a conceptual point of view, the quantitative estimation of *within*-species RSA in MMG is easier than the *across*-species RSA, because in the latter the mitochondrial DNA copy number can vary widely between species.

With the single-species libraries we estimated the mitochondrial DNA copy number (N_M) of each species in the following way:

1. Let x_i be the ratio of the genomic mitochondrial information divided by the total (haploid) genomic information for species *i*. The mitochondrial information is the mitogenome length (M_i) times N_{Mi}; the total genomic information is the sum of the nuclear genome length (G_i) and the mitochondrial information.

$$x_i = \frac{M_i \cdot N_{Mi}}{G_i + (M_i \cdot N_{Mi})} \quad (1)$$

2. The re-arrangement of Equation (1) allows the estimation of N_{Mi}.

$$N_{Mi} = \frac{x_i \cdot G_i}{M_i \cdot (1 - x_i)} \quad (2)$$

3. G_i and M_i are known for species with sequenced genomes, but x_i is not. In our experimental setting, x_i can be estimated in the single-species libraries as the ratio between the number of reads that map into the mitogenome (R_{Mi}) divided by the total number of reads of species *i* (R_{Gi}).

$$x_i = \frac{R_{Mi}}{R_{Gi}} \quad (3)$$

We obtained R_{Mi} by mapping the reads of species *i* to its mitogenome when this mitogenome was the only one used as the reference in the mapping. Regarding R_{Gi}, we assumed that all reads of the single-species library of species *i* belong to species *i*.

In the comparison of actual *versus* estimated RSA using the mixed-species libraries, we multiplied the actual relative abundance of species *i* (Table S1) by N_{Mi}, and then renormalized the values to sum 1.

Finally, we estimated the importance of knowing or ignoring the individual values of G_i and M_i of each species in the mixture in the estimated RSA by comparing the results obtained with the correction factor of Equation (2) (N_{Mi}) with another factor that uses the mean value of *G* ($\overline{G}$ = 338 Mbp) and *M* ($\overline{M}$ = 16.3 kbp) for all the species in the mixture ($\overline{N}_{Mi}$):

$$\overline{N}_{Mi} = \frac{x_i \cdot \overline{G}}{\overline{M} \cdot (1 - x_i)} \quad (4)$$

2.6 | Rarefaction of the input samples

We only multiplexed six mixed-species libraries in a single Illumina MiSeq run (Table S1), with the consequence of a high economic cost per library. However, from a practical point of view, it would be interesting to use fewer reads per library and still have a good quantitative estimation of RSA. To test this possibility, we randomly rarefacted the mixed-species samples at various proportions of the original number of reads {0.5, 0.1, 0.05, 0.01} and run the new data sets through the entire pipeline. We repeated each simulation 100 times using different subsets and from every simulation we recorded the number of recovered species.

2.7 | Statistical analyses and hardware

All statistical analyses were performed with R 3.4.2 (R Core Team, 2016) in RSTUDIO 1.0.143 (RStudio Team, 2015). Plots were

created using the R packages ggplot2 (Wickham, 2016) and ggpubr (Kassambara, 2018).

We run the complete pipeline on a server with two Intel Xeon E5-2620 v3 processors with six cores each and hyperthreading technology, so a maximum of 24 threads were available.

3 | RESULTS

3.1 | Species identification

The 21 single-species libraries (Table 1) generated 2,371,091 ± 1,210,091 (mean ± SD) paired-end reads. A proportion of 0.012 ± 0.027 reads were eliminated in the trimming step, remaining a proportion of 0.987 ± 0.027 reads available for further analysis.

The results that follow correspond to the application of the γ–δ algorithm for the assignation of reads to species on the training set (i.e., 75% of the sequenced data). We also eliminated from the following results the reads assigned to species that could legitimately be attributed to physical contamination in the laboratory or the sequencing. These contaminants were species sequenced in different libraries of the same Illumina run and the fly *Ceratitis capitata* that contaminated the library of *Bactrocera oleae* (see the discussion for the reason behind this contamination).

Because the MMG method must recover only one species in single-species libraries, we fixed the values of γ = 0.99 and δ = 0.96 in the γ–δ algorithm as this combination was the only one to provide the expected result (Table S3). All the other tested combinations reported false positives, such as *Bactrocera biguttula* in libraries of *B. oleae*, *Drosophila formosana* in libraries of *D. melanogaster* and *Solenopsis richteri* in libraries of *S. invicta*. Results of all tested γ, δ and input data combinations (raw reads *versus* candidate mito-reads and Mito1794 *versus* FilteredMito1794 references) are provided as Supporting Information (Tables S4–S7).

Filtering out the repetitive regions of the mitogenomes (i.e., FilteredMito1794) had a dramatic effect on the number of identified species. With the FilteredMito1794 database, we only detected the focal species in all the libraries, whereas with the Mito1794 database there appeared several false positives in many libraries (2.5 ± 1.2 species per library using all raw reads or 1.7 ± 0.9 species using only candidate mito-reads; Table 2a). The masked regions mostly belonged to non-coding regions of the mitogenome, including the control region (Table S8). The use of only candidate mito-reads instead of all reads produced a loss of c. 6% of informative reads (Table 2b) but reduced ~18 times the execution time (Table 2c). In summary, the elimination of the repetitive regions from the genomes removed all the false positives and the mining of candidate mito-reads reduced 18-fold the execution time of the pipeline with a moderate loss of informative reads. Therefore, in the subsequent steps, we used both the FilteredMito1794 database and only the candidate mito-reads (Figure 1).

We evaluated the goodness of fit of the model with the test set (i.e., the remaining 25% of reads not used in the previous calibration) using the best set of input data and parameters (i.e., the FilteredMito1794 database, the candidate mito-reads and the parameters γ = 0.99 and δ = 0.96). The number of identified species per library was 1 in all cases (Table S9) and the proportion of informative reads was 0.0046 ± 0.0056.

The absence of the mitogenome of the focal species in the reference database did not produce many false positives in the single-species libraries (Table 3). When there were no congeneric species of the focal species in the reference database (six out of 17 species), no read was assigned to any species; when there were congeneric species in the database, in six cases no reads were assigned to any species and in four cases some reads were assigned to another species of the same genus (*Bactrocera*, *Drosophila*, *Plutella*, and *Solenopsis*); only in one species (*Drosophila melanogaster*) appeared some reads belonging to species of a different genus (*Exorista sorbillans*, Diptera:Tachinidae).

The six mixed-species libraries (Table S1) generated 3,376,087 ± 424,238 paired-end reads. A proportion of 0.003 ± 0.001 reads were eliminated in the trimming step and a proportion of 0.925 ± 0.006 in the mito-reads mining step. Therefore, only a proportion of 0.075 ± 0.006 of the raw reads were candidate mito-reads retained for further analysis.

TABLE 2 Summary of the results per library (mean ± SD) on the training data set of single-species libraries for the four combinations of input data assessed in this study (raw reads and candidate mito-reads mapped to Mito1794 and FilteredMito1794 databases) and using γ = 0.99 and δ = 0.96. (a) Number of recovered species per library, (b) Relative proportion of informative reads per library, and (c) processing time per library (format h:mm:ss; the time necessary to find the candidate mito-reads is included in the processing time). Reads from contaminant species have not been considered

		Reference database	
Metric	**Input data**	**Mito1794**	**FilteredMito1794**
(a) Number of identified species	Raw reads	2.52 ± 1.21	1 ± 0
	Mito-reads	1.71 ± 0.90	1 ± 0
(b) Relative proportion of informative reads	Raw reads	0.0049 ± 0.0057	0.0047 ± 0.0056
	Mito-reads	0.0049 ± 0.0057	0.0046 ± 0.0057
(c) Processing time	Raw reads	7:19:43 ± 3:46:53	7:21:10 ± 3:53:34
	Mito-reads	0:25:13 ± 0:17:07	0:24:37 ± 0:16:37

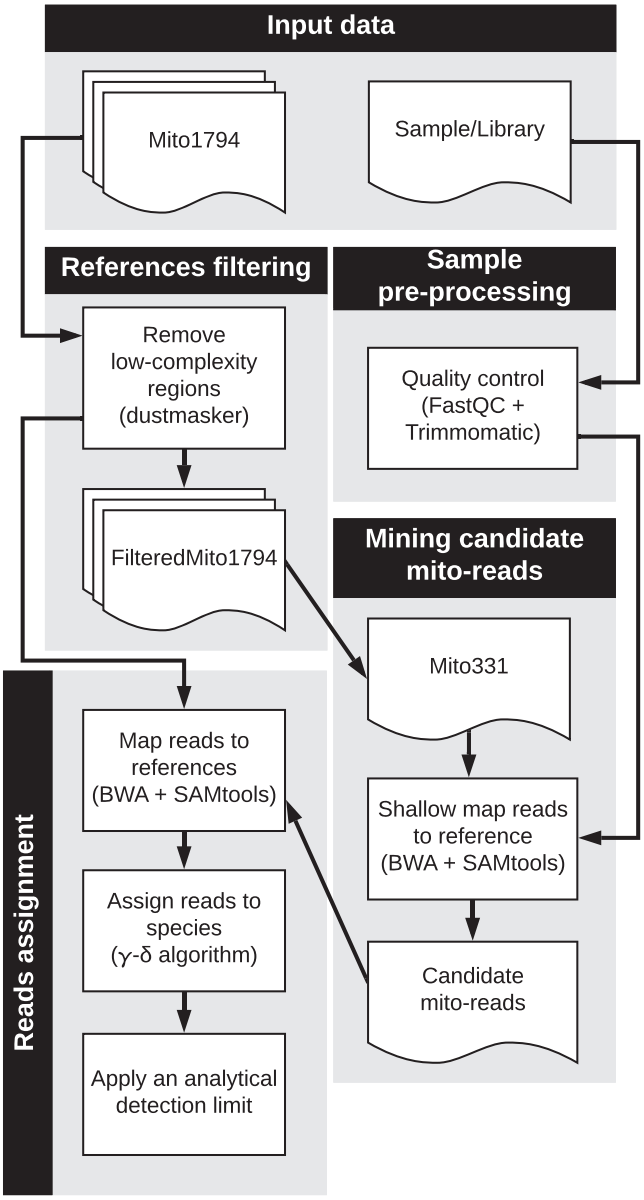

FIGURE 1 MMG pipeline applied in this study. In brackets, the tools used in each step

In the mixed-species libraries, we recovered all species included in the libraries except *Papilio machaon* in library no. 2 (Table S10). As in the single-species libraries, in the mixed-species libraries, we found reads of *Ceratitis capitata* and discarded them as laboratory contamination. In addition, in libraries no. 2 and no. 3, one single read was attributed to *Bactrocera biguttula* (Table S10), a species not handled in the laboratory; therefore, an analytical detection limit of $\varepsilon = 0.0001$ would be useful for the elimination of all false positives.

3.2 | Estimation of the relative species abundance in mixed-species libraries

The *within*-species RSA was well estimated for all species ($r \geq .97$ and $p < .05$ for all species; Figure 2), but the *across*-species RSA estimation was very poor ($r \leq .67$ and $p > .05$ for all samples; Figure 3a). Thus, it seems clear the need for a species-specific correction factor that considers a variable ratio of mitochondrial to nuclear DNA (Table 4). When we modified the actual RSA with the N_{Mi} correction factor (Equation 2), the correlation between actual and estimated RSA *across*-species became significant in all samples ($r \geq .84$ and $p < .05$ for all libraries; Figure 3b). The use of the $\overline{N}_{Mi}$ correction factor (Equation 4) instead of N_{Mi} provided an even better quantitative estimation of RSA *across*-species ($r \geq .91$ and $p < .005$ for all libraries; Figure 3c).

The use of rarefacted samples showed that in the libraries with a more variable species abundance (libraries no. 1 and no. 2), the use of just half of the total available reads reduced the number of the identified species (Figure 4a) and promoted the presence of low-abundant false positives, like *Bactrocera biguttula* above the detection limit $\varepsilon = 0.0001$ in library no. 2. On the contrary, when the abundance of species was less variable (libraries no. 5 and no. 6), the expected number of species was obtained with half the reads (Figure 4f) or even with 10% of reads (Figure 4e).

3.3 | Computer use

The total consumed time by running the entire pipeline with the mixed-species libraries ranged between 66 and 82 min (Table S11). Most of the processing time was devoted to the mapping of the reads to the references (95%) and only 3% of the time was used by the γ−δ algorithm (Table S11).

4 | DISCUSSION

Mito-metagenomics (MMG) proved to be able to set apart and quantify the relative DNA abundance of insect species in artificial mixtures, even when the species were congeneric. The estimation was as good as the one obtained with whole genomes instead of mitogenomes by Garrido-Sanz et al. (2020), using the same DNA libraries and bioinformatic methods as here. However, to be able to quantify the RSA in a sample with several species (*across*-species RSA) it was necessary to correct the raw reads by the variable amount of mitochondrial to nuclear DNA (mitochondrial DNA copy number) among species.

4.1 | Species identification

The MMG approach used here recovered only the focal species from all the single-species libraries, with no false positives when low-complexity regions were filtered out from the mitogenomes (except for genuine contaminants, see below). Without this filtering step, some reads were attributed to non-focal species; these reads were sequences with biased composition, probably from repetitive regions (e.g., microsatellites) that mostly matched non-coding

TABLE 3 List of species detected on the single-species libraries when the mitogenome of the focal species is in the reference database (column A) and when it is not (column B). For each detected species we indicate its name and the number of assigned reads (in brackets). The number of congeneric species of the focal species included in the database is provided in column C. Libraries are divided into four groups: Group 1, species without congeneric species in the database and without false positive species; Group 2, species with congeneric species in the database and without false positive species; Group 3, species with congeneric species in the database but with false positive of the same genus; and group 4, species with congeneric species in the database but with false positive of a different genus

Group	Run	Library	Species used to prepare the library	(A) Focal species mitogenome present in database	(B) Focal species mitogenome not present in database	(C) Number of congeneric species in the database
1	1	9	*Acyrthosiphon pisum*	*Acyrthosiphon pisum* (154)	None	0
	2	1	*Atta colombica*	*Atta colombica* (19412)	None	0
	2	3	*Cimex lectularius*	*Cimex lectularius* (1082)	None	0
	1	6	*Linepithema humile*	*Linepithema humile* (218)	None	0
	2	8	*Linepithema humile*	*Linepithema humile* (913)	None	0
	2	11	*Vollenhovia emeryi*	*Vollenhovia emeryi* (1399)	None	0
	2	12	*Wasmannia auropunctata*	*Wasmannia auropunctata* (17)	None	0
2	1	8	*Apis mellifera*	*Apis mellifera* (3597)	None	7
	2	2	*Bemisia tabaci*	*Bemisia tabaci* (349)	None	1
	1	7	*Bombus terrestris*	*Bombus terrestris* (1877)	None	2
	1	4	*Drosophila mojavensis*	*Drosophila mojavensis* (4407)	None	18
	2	5	*Drosophila mojavensis*	*Drosophila mojavensis* (2646)	None	18
	2	7	*Drosophila suzukii*	*Drosophila suzukii* (4664)	None	18
	1	1	*Papilio machaon*	*Papilio machaon* (312)	None	13
3	1	5	*Bactrocera oleae*	*Bactrocera oleae* (425)	*Bactrocera biguttula* (137)	14
	1	2	*Drosophila virilis*	*Drosophila virilis* (8070)	*Drosophila littoralis* (20)	18
	2	6	*Drosophila virilis*	*Drosophila virilis* (2704)	*Drosophila littoralis* (6)	18
	2	9	*Plutella xylostella*	*Plutella xylostella* (2371)	*Plutella australiana* (127)	1
	2	10	*Solenopsis invicta*	*Solenopsis invicta* (131)	*Solenopsis richteri* (236)	2
4	1	3	*Drosophila melanogaster*	*Drosophila melanogaster* (979)	*Drosophila formosana* (312) *Exorista sorbillans* (26) *Drosophila mauritiana* (3)	18
	2	4	*Drosophila melanogaster*	*Drosophila melanogaster* (1384)	*Drosophila formosana* (478) *Exorista sorbillans* (53) *Drosophila mauritiana* (3)	18

regions on the reference mitogenomes (Table S8; Faber & Stepien, 1998; Wolff et al., 2012). Some popular tools do implicitly or explicitly filter out low complexity regions from the reference genomes: Kraken masks low complexity regions when adding references to the database (Wood & Salzberg, 2014); and BLAST filters both query sequences and references (Altschul et al., 1990, 1997; Camacho et al., 2009).

We also found contaminant species in all the sequenced libraries (Tables S9–S10). The origin of these reads of contaminants can be tag jumping during the sequencing reaction (Schnell et al., 2015) or actual contamination in the laboratory. The first reason is probably the cause of finding reads in single-species libraries belonging to other species sequenced in the same run but different libraries. The second reason is behind the presence of reads attributed to *Ceratitis capitata* in libraries where *Bactrocera oleae* was also present, because the two dipterans, which are agricultural pests, were captured together in fly traps. A more throughout discussion about this problem is provided in Garrido-Sanz et al. (2020). The removal of the genuine contaminant species in artificial libraries as we did here was possible because we knew the identity of the species in the mixture, but it is impossible in real samples.

4.2 | Quantification of the RSA and the need for a species-specific correction factor

With the mixed-species libraries, we estimated the *within*-species RSA with high statistical confidence (Figure 2). Similar good results have been reported in previous studies that used mock samples (Bista et al., 2018; Ji et al., 2020). On the contrary, the *across*-species RSA estimation within a sample was not statistically significant in any sample (Figure 3a). These results contrast with the study of

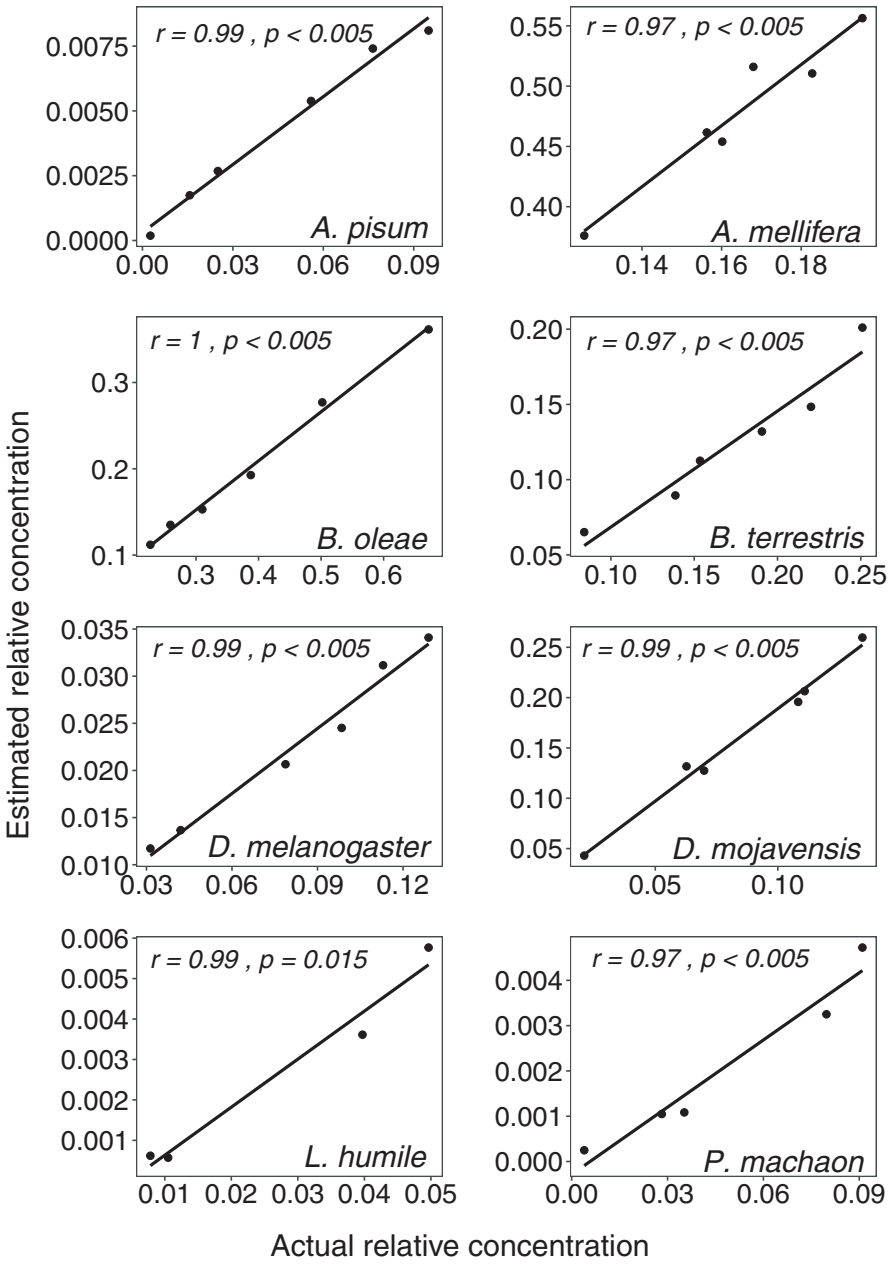

FIGURE 2 Scatter plot of the estimated *versus* the actual RSA for each species of the mixed-species libraries (i.e., *within*-species RSA). Each plot shows the Pearson correlation coefficient (*r*) and the corresponding *p*-value. The coordinate at the origin of all regression lines was not different to 0

Gueuning et al. (2019) that reported good quantitative estimations of *across*-species RSA in artificial mixtures of wild bees.

Using the mitochondrial DNA copy number correction factor, the RSA *across*-species correlated significantly with the real values in the six artificial samples analysed (Equation 2; Figure 3b), even when the mean genome and mitogenome sizes were used instead of the species-specific value (Equation 4; Figure 3c). Other studies reporting RSA estimation *across*-species also used some correction factor before comparing the expected and observed number of reads, but none included the genome size, as we did here. For instance, Gómez-Rodríguez et al. (2015) and Tang et al. (2015) considered the mitogenome size of the species and Tang et al. (2015) and Lang et al. (2019) the number of reads from the genome. Tang et al. (2015) is the only study that provides both the goodness of fit with and without the use of the correction factor, and the effect is very different from the one reported here, as the result was almost the same in both cases. One possible explanation is that Tang et al. (2015) dealt only with wild bees (a group of a few Hymenoptera families), so the interspecific differences might be low compared to our study which included species from four insect orders. Nevertheless, the effect of the mitochondrial DNA copy number on the estimation of RSA deserves more research effort if DNA-based techniques are to provide good quantitative results, both using mito-metagenomics and amplicon metabarcoding targeting variable copy number regions.

The method used here to estimate the correction factor for the mitochondrial DNA copy number (i.e., preparation and sequencing of a single-species library to a depth of c. one million reads) has a cost that is not negligible. Ideally, there should be a method to independently estimate the mitochondrial DNA copy number of each species that did not involve sequencing. Such methods do exist

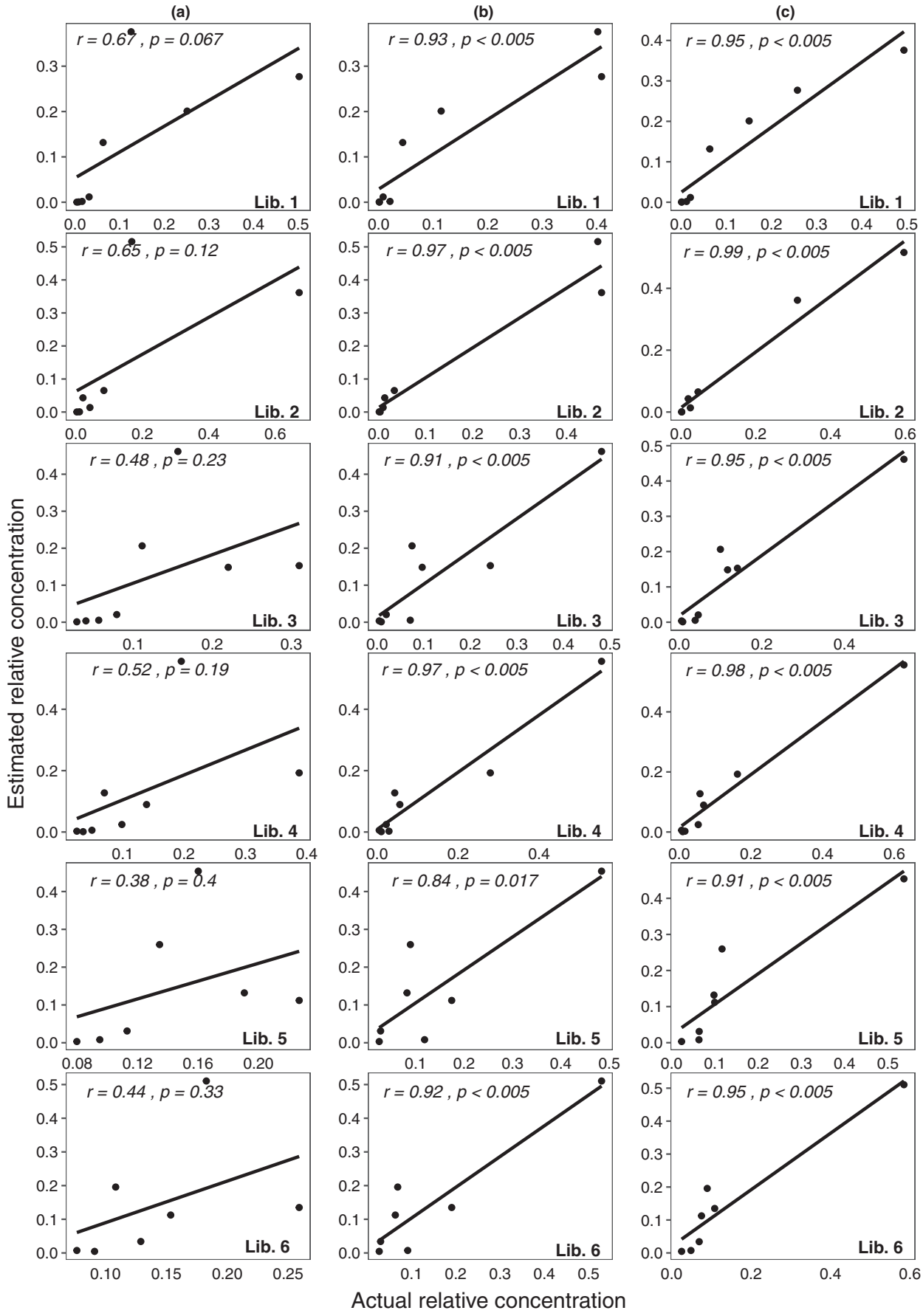

FIGURE 3 Scatter plot of the estimated *versus* the actual RSA in each mixed-species library (i.e., *across*-species RSA). At the top, it is indicated the way we conducted the actual RSA. (a) Original expected data. (b) Corrected expected data after applying the N_{Mi} correction factor. (c) Corrected expected data after applying the $\overline{N}_{Mi}$ correction factor. Rows from top to bottom correspond to mixed-species libraries from 1 to 6. Each plot shows the Pearson correlation coefficient (*r*) and the corresponding *p*-value

because there is an interest in medicine to measure the mitochondrial DNA copy number for its relationship with several diseases. In medicine, the mitochondrial DNA copy number is usually estimated using quantitative PCR (qPCR; Thyagarajan et al., 2012); this method requires two primer pairs, one for a mitochondrial marker and one for a single-copy nuclear marker. These primers are known for humans, but it would be costly to generate them for every species in an environmental sample, especially for those species whose genome has not yet been sequenced (but see Liu et al., 2018). The qPCR itself is cheap, but the preparative work for each species would be long.

It is important to emphasize that the RSA used here is based on the relative proportions of DNA of the species in the mixture, but what is needed in most ecological applications is the relative proportions of biomass (or individual counts) of the different species. The reason behind our choice was to simplify the problem of obtaining the actual RSA from high-throughput sequenced reads in several steps. There is one bias caused by the variable mitochondrial copy number of the different species and there is another, independent, bias caused by the variable DNA content of the biomass of different species. We addressed here the first bias and obviated the second one. From our results, the RSA based on the biomass could be obtained by multiplying our estimates by the biomass-to-DNA ratio of each species, if known. There are very little data in the literature about the proportion of DNA to biomass in different species, but it can be very variable; for instance, Pornon et al. (2016) reports a very different DNA yield from the same number of pollen grains of three plant species.

In metabarcoding applications, some authors use empirical correction factors based on mixtures of known relative biomass of several species rather than in mixtures of DNA (e.g., Matesanz et al., 2019; Thomas et al., 2016). In these cases, the correction factor solves for the mitochondrial copy number among species and also for the DNA-to-biomass ratio and the differential amplification efficiency caused by PCR. This method is undoubtedly practical but does not differentiate the relative importance of each source of bias.

4.3 | Mito-metagenomics in *real* samples

The present study is based on artificial mixtures of a low number of species whose mitogenomes are already assembled. Thus, it is fair to question the value of our proposal in real samples with many more species, with a limited amount of DNA, where the prior species composition is unknown, or when the reference mitogenomes are obtained in the same experiment and are only partially assembled.

4.3.1 | More complex mixtures

Real samples can contain hundreds of species, and that might affect the ability of the method to detect the less abundant ones. With the sequencing depth achieved here (~3.4 million raw reads per sample; Table S1) we were able to detect three (out of four) species with an expected RSA below 1‰ (Table S10). The subsequent rarefaction experiment showed that with fewer reads more species become undetected (Figure 4). In consequence, it seems that at least $3.4{\cdot}10^6$ reads are needed to detect most species with an RSA above 1‰. Having hundreds of species in the mixture would not hamper the quantitative ability of the method, as most species would be above a 1‰ abundance. However, ultra-rich samples with thousands of species would require a higher sequencing depth to ensure the detection of most species.

4.3.2 | Limited amount of DNA available

The Illumina TruSeq kit used here to prepare the libraries requires 1 μg of DNA and this might be a problem with small specimens or in DNA-poor samples. However, today there are alternative methods that provide good results with just 1 ng of DNA, like the Illumina Nextera DNA Flex kit (Sato et al., 2019), albeit potential biases should be tested in future experiments for these kits.

4.3.3 | Absence of mitogenomes in the reference database

Our results showed that the proposed methodology was robust in the absence of the mitogenomes of species in the reference database. Of course, the species without their mitogenome in the reference database will never be found, but their reads will not generate many false positives, even for species with close relatives in the reference database (Table 3). The presence of species without their reference genome in the mixture is likely to occur frequently in real samples. The unassigned reads (or also when the prior composition of the sample is unknown) can be further explored by mapping them against other databases, like COI barcodes from BOLD; thus, the identity of more species will be revealed, albeit not their relative abundance.

4.3.4 | Incomplete genomes

In several MMG studies, the reference mitogenomes are assembled from the same mixtures in which the RSA is intended to be quantified

TABLE 4 Summary table of the species in single-species libraries data used for the obtention of the correction factors N_{Mi} and $\overline{N}_{Mi}$. x_i is the number of reads mapping into the mitogenome of species *i* divided by the total number of reads from that single-species library; N_{Mi} is the estimated number of mitochondrial DNA copies for species *i*; $\overline{N}_{Mi}$ is as N_{Mi} but calculated using the mean length of whole genomes and mitogenomes of all species considered here

Species	Complete genome size (Mbp)	Mitochondrial genome size (kbp)	Number of raw reads (paired-end)	Number of reads mapped to the mitogenome	x_i (from Equation 3)	N_{Mi} (from Equation 2)	$\overline{N}_{Mi}$ (from Equation 4)
Papilio machaon	278	15.2	434,520	1,992	0.0046	84.4	95.5
Drosophila virilis	206	14.9	4,714,902	48,967	0.0104	144.6	217.6
Drosophila melanogaster	144	20.0	2,283,768	20,424	0.0089	66.4	187.1
Drosophila mojavensis	197	14.9	1,668,424	22,839	0.0137	180.5	287.8
Bactrocera oleae	472	15.8	580,996	4,052	0.0070	209.5	145.6
Linepithema humile	220	16.1	1,422,342	3,151	0.0022	30.3	46.0
Bombus terrestris	249	17.4	1,994,938	16,262	0.0082	117.4	170.4
Apis mellifera	250	16.3	1,262,388	64,448	0.0511	823.9	1115.6
Acyrthosiphon pisum	542	17.0	684,688	7,218	0.0105	340.1	220.9

(Crampton-Platt et al., 2016; Zhou et al., 2013). In these cases, the mitogenomes are assembled *de novo* and, normally, they are incomplete. We do not see any impediment in using mitogenomes assembled in this way if all of them have a similar length and quality. However, we would advise against the simultaneous use of mitogenomes with disparate length or quality for quantification purposes, because that would bias the RSA towards the species with better mitogenomes (Tang et al., 2015). On the contrary, the use of all available partial mitogenomes would be fine for identification purposes.

4.3.5 | Estimation of the mitochondrial DNA copy number (N_{Mi})

Perhaps the most difficult problem in real samples is the estimation of N_{Mi}. The rationale that we propose for the estimation of N_{Mi} (Equations 1–4) seems reasonable, but the devil is in the detail: the estimation of the variables needed to calculate N_{Mi} is paved with difficulties for species without a reference genome. First, the estimation of the proportion of reads that belong to the mitogenomes (x_i, Equation 3) is biased, because we assumed that all reads belong to the same species (R_{Gi}, Equation 3); however, there is always DNA that comes from other sources, like food, gut bacteria, parasites, etc. Consequently, the number of reads of the entire genome R_{Gi} is overestimated and, hence, x_i underestimated. Second, the size of the mitogenome and the whole genome is generally unknown for most species; even for the best studied species their whole genome is far from complete (e.g., Paris et al., 2020), and the estimated size (G_i) is an underestimation of the real size.

Nevertheless, despite the above problems, the correction factor N_{Mi} helped to reproduce the expected *across*-species RSA in our libraries. Similar results were obtained using the mean values of the mitogenome and whole genome sizes ($\overline{N}_{Mi}$, Equation 4). The apparent lack of effect of the species-specific genome and mitogenome sizes might be caused to the low variability of the mitogenome size (coefficient of variation, CV = 9%) and moderate variability of the whole genome size (CV = 47%; Table 4). On the contrary, the proportion of reads mapping into the mitogenome (x_i, Equation 3) was much more variable among species (CV = 113%).

Given the previous considerations, we suggest the use of the correction factor $\overline{N}_{Mi}$ instead of N_{Mi}, for species without a reference genome and to estimate the three necessary variables (x_i, M_i, G_i) in the following way.

- x_i. The proportion of reads belonging to the mitogenome of species *i* could be estimated by shotgun sequencing a single-species DNA extract. The value of x_i would be an underestimation of the real value but given the high interspecific variability, the obtained x_i values should still be useful for correction purposes.
- M_i. Ninety per cent of the 1794 mitogenomes used here have a length of 14.9 to 17.0 kpb (i.e., a rank of 2.1 kbp or 13% of the mean M_i; Table S2). Consequently, we recommend the use of the mean value $\overline{M}$ for the group of species of interest (Table S12).

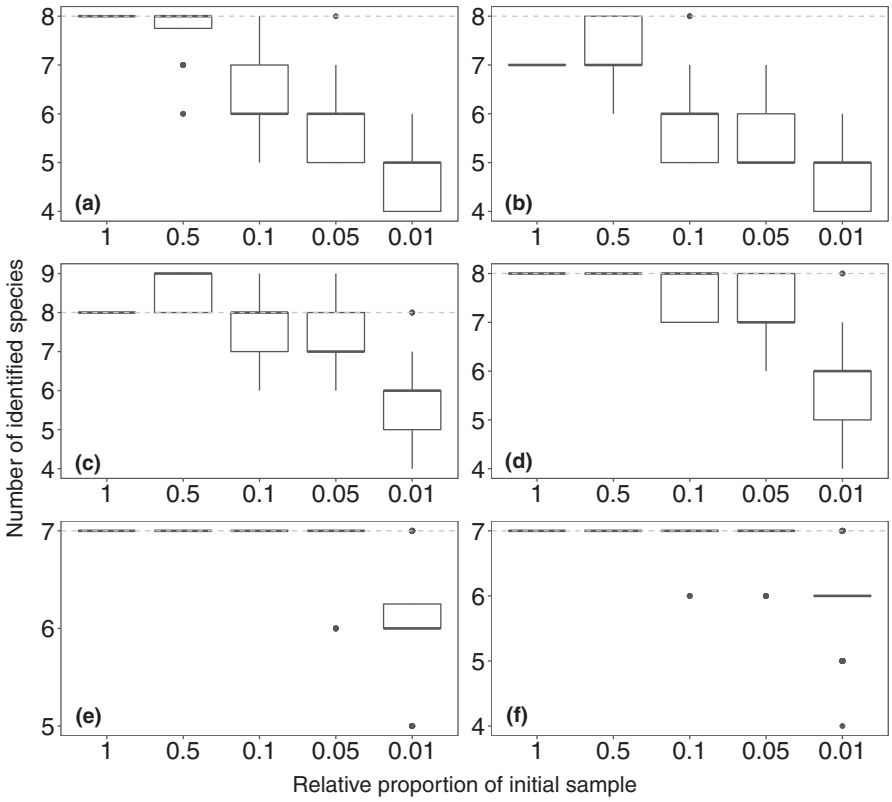

FIGURE 4 Number of identified species using different proportions of reads in the mixed-species libraries. Each simulation was performed 100 times with different subsets, except when the entire library was used. Letters from a to f indicate mixed-species libraries from 1 to 6. Grey dashed lines indicate the expected number of recovered species in each library

- G_i. The length of the whole genome is more variable across species than the length of the mitogenomes: 90% of the 115 whole genomes of insects available at RefSeq (Table S13A) have a length between 0.14 and 0.98 Gbp. However, if the insects are split by orders the variability of G_i is smaller for most insect orders (Table S13B). Consequently, we would advise using the mean value $\bar{G}$ for each group of taxa (e.g., insect orders).

4.4 | Concluding remarks

The approach presented here to identify insect species and to estimate their relative abundance in complex mixtures using MMG worked well with artificial samples of known composition for a select group of species whose mitogenomes are sequenced to an advanced degree. The key for the accurate estimation of the *across*-species RSA was a correction factor for the mitochondrial copy number of each species. We are aware that the proposed methodology is not immediately applicable to most real samples, so its real value should be tested on more of such samples.

ACKNOWLEDGEMENTS

We are grateful to several entomologists who provided the specimens: Xavier Espadaler, Nicolás Pérez, Alfredo Ruíz, Francesc Mestres, Aleix Valls, Francisco Beitia, Carlos Hernández-Castellano, Joan Josep Ibañez, Carlos Pradera, Rasmus S. Larsen, Jacobus J. Boomsma, Luis Calcaterra and Misato O. Miyakawa. We also thank Anna Barceló and Roger Lahoz of the Genomics facilities of the UAB for the preparation of the DNA libraries and sequencing the DNA and Mario Cáceres for his advice on a preliminary version of this manuscript. Financial support was provided by Spanish Government grant TIN2017-84553-C2-1-R and Generalitat de Catalunya grant AGAUR 2017-SGR-1001.

AUTHOR CONTRIBUTIONS

Lidia Garrido-Sanz and Josep Piñol conceived the experiment. Lidia Garrido-Sanz performed the bioinformatic analyses and the other two authors supervised it. All authors were involved in the analysis, interpretation of data and writing of the manuscript.

DATA AVAILABILITY STATEMENT

Sequenced DNA samples have been made available at the Dryad repository: https://doi.org/10.5061/dryad.t1g1jwsz7. The γ–δ algorithm script is available at GitHub: https://github.com/LidiaGS/g-d_algorithm.

ORCID

Lidia Garrido-Sanz https://orcid.org/0000-0003-2622-9674
Miquel Àngel Senar https://orcid.org/0000-0002-0316-5420
Josep Piñol https://orcid.org/0000-0002-4067-3301

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Andrews, S. (2015). *FastQC: A quality control tool for high throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bista, I., Carvalho, G. R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., Shokralla, S., Seymour, M., Bradley, D., Liu, S., Christmas, M., & Creer, S. (2018). Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources*, *18*, 1020–1034. https://doi.org/10.1111/1755-0998.12888

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

CBoL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Crampton-Platt, A., Timmermans, M. J. T. N., Gimmel, M. L., Kutty, S. N., Cockerill, T. D., Khen, C. V., & Vogler, A. P. (2015). Soup to tree: The phylogeny of beetles inferred by mitochondrial metagenomics of a borean rainforest sample. *Molecular Biology and Evolution*, *32*(9), 2302–2316. https://doi.org/10.1093/molbev/msv111

Crampton-Platt, A., Yu, D. W., Zhou, X., & Vogler, A. P. (2016). Mitochondrial metagenomics: letting the genes out of the bottle. *Gigascience*, *5*(1), 15. https://doi.org/10.1186/s13742-016-0120-y

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*(7), e0130324. https://doi.org/10.1371/journal.pone.0130324

Faber, J. E., & Stepien, C. A. (1998). Tandemly repeated sequences in the mitochondrial DNA control region and phylogeography of the pike-perches *Stizostedion*. *Molecular Phylogenetics and Evolution*, *10*(3), 310–322. https://doi.org/10.1006/mpev.1998.0530

Garrido-Sanz, L., Senar, M. À., & Piñol, J. (2020). Estimation of the relative abundance of species in artificial mixtures of insects using low-coverage shotgun metagenomics. *Metabarcoding and Metagenomics*, *4*, e48281. https://doi.org/10.3897/mbmg.4.48281

Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M. J. T. N., 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*(8), 883–894. https://doi.org/10.1111/2041-210X.12376

Gregory, T. R. (2020). *Animal genome size database*. http://www.genomesize.com

Gueuning, M., Ganser, D., Blaser, S., Albrecht, M., Knop, E., Praz, C., & Frey, J. E. (2019). Evaluating next-generation sequencing (NGS) methods for routine monitoring of wild bees: Metabarcoding, mitogenomics or NGS barcoding. *Molecular Ecology Resources*, *19*(4), 847–862. https://doi.org/10.1111/1755-0998.13013

Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*, S96–S99. https://doi.org/10.1098/rsbl.2003.0025

Ji, Y., Huotari, T., Roslin, T., Schmidt, N. M., Wang, J., Yu, D. W., & Ovaskainen, O. (2020). SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Molecular Ecology Resources*, *20*, 256–267. https://doi.org/10.1111/1755-0998.13057

Kassambara, A. (2018). *ggpubr: 'ggplot2' Based publication ready plots*. R package version 0.2. https://CRAN.R-project.org/package=ggpubr

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. https://doi.org/10.1038/s41598-017-17333-x

Lang, D., Tang, M., Hu, J., & Zhou, X. (2019). Genome-skimming provides accurate quantification for pollen mixtures. *Molecular Ecology Resources*, *19*(6), 1433–1446. https://doi.org/10.1111/1755-0998.13061

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997v1 [q-bio.GN].

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, R., Jin, L., Long, K., Tang, Q., Ma, J., Wang, X., Zhu, L. I., Jiang, A., Tang, G., Jiang, Y., Li, X., & Li, M. (2018). Analysis of mitochondrial DNA sequence and copy number variation across five high-altitude species and their low-altitude relatives. *Mitochondrial DNA Part B*, *3*(2), 847–851. https://doi.org/10.1080/23802359.2018.1501285

Lu, J., & Salzberg, S. L. (2018). Removing contaminants from databases of draft genomes. *PLoS Computational Biology*, *14*(6), e1006277. https://doi.org/10.1371/journal.pcbi.1006277

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

Morgulis, A., Gertz, E. M., Schäffer, A. A., & Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of Computational Biology*, *13*(5), 1028–1040. https://doi.org/10.1089/cmb.2006.13.1028

Paris, M., Boyer, R., Jaenichen, R., Wolf, J., Karageorgi, M., Green, J., Cagnon, M., Parinello, H., Estoup, A., Gautier, M., Gompel, N., & Prud'homme, B. (2020). Near-chromosome level genome assembly of the fruit pest *Drosophila suzukii* using long-read sequencing. *Scientific Reports*, *10*, 11227. https://doi.org/10.1038/s41598-020-67373-z

Pearman, W. S., Freed, N. E., & Silander, O. K. (2019). The advantages and disadvantages of short- and long-read metagenomics to infer bacterial and eukaryotic community composition. *bioRxiv*. https://doi.org/10.1101/650788

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*(4), 819–830. https://doi.org/10.1111/1755-0998.12355

Pornon, A., Escaravage, N., Burrus, M., Holota, H., Khimoun, A., Mariette, J., Pellizzari, C., Iribar, A., Etienne, R., Taberlet, P., Vidal, M., Winterton, P., Zinger, L., & Andalo, C. (2016). Using metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific Reports*, *6*, 27282. https://doi.org/10.1038/srep27282

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

RStudio Team (2015). *RStudio: Integrated development for R*. RStudio Inc. http://ww.rstudio.com/

Saitoh, S., Aoyama, H., Fujii, S., Sunagawa, H., Nagahama, H., Akutsu, M., Shinzato, N., Kaneko, N., & Nakamori, T. (2016). A quantitative protocol for DNA metabarcoding of springtails (Collembola). *Genome*, *59*(9), 705–723. https://doi.org/10.1139/gen-2015-0228

Sato, M. P., Ogura, Y., Nakamura, K., Nishida, R., Gotoh, Y., Hayashi, M., Hisatsune, J., Sugai, M., Takehiko, I., & Hayashi, T. (2019). Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Research*, *26*(5), 391–398. https://doi.org/10.1093/dnares/dsz017

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

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Bolchacova, E., Voigt, K., Crous, P. W., Miller, A. N., Wingfield, M. J., Aime, M. C., An, K.-D., Bai, F.-Y., Barreto, R. W., Begerow, D., Bergeron, M.-J., Blackwell, M., … Schindel, D. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

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

Tang, M., Hardman, C. J., Ji, Y., Meng, G., Liu, S., Tan, M., Yang, S., Moss, E. D., Wang, J., Yang, C., Bruce, C., Nevard, T., Potts, S. G., Zhou, X., & Yu, D. W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, *6*, 1034–1043. https://doi.org/10.1111/2041-210X.12416

Tang, M., Tan, M., Meng, G., Yang, S., Su, X. U., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A., & Zhou, X. (2014). Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, *42*(22), e166. https://doi.org/10.1093/nar/gku917

Thomas, A. C., Deagle, B. E., Everson, J. P., Harsch, C. H., & Trites, A. W. (2016). 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

Thyagarajan, B., Wang, R., Barcelo, H., Koh, W.-P., & Yuan, J.-M. (2012). Mitochondrial copy number is associated with colorectal cancer risk. *Cancer Epidemiology, Biomarkers & Prevention*, *21*(9), 1574–1581. https://doi.org/10.1158/1055-9965.EPI-12-0138-T

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

Wolff, J. N., Shearman, D. C. A., Brooks, R. C., & Ballard, J. W. O. (2012). Selective enrichment and sequencing of whole mitochondrial genomes in the presence of nuclear encoded mitochondrial pseudogenes (Numts). *PLoS One*, *7*(5), e37142. https://doi.org/10.1371/journal.pone.0037142

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X. U., Zhou, L., Tang, M., Fu, R., Li, J., & Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience*, *2*(1), 4. https://doi.org/10.1186/2047-217X-2-4

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Garrido-Sanz, L., Senar, M. À., & Piñol, J. (2022). 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, 153–167. https://doi.org/10.1111/1755-0998.13464