

Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <http://orca.cf.ac.uk/112129/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Pinol, Josep, Senar, Miquel A. and Symondson, William O. C. 2019. The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology* 28 (2) , pp. 407-419. 10.1111/mec.14776 file

Publishers page: <http://dx.doi.org/10.1111/mec.14776> <<http://dx.doi.org/10.1111/mec.14776>>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



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

3 Josep Piñol^{1,2}, Miquel A Senar¹, William O C Symondson³

4

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

6 (2) CREA, Cerdanyola del Vallès, 08193, Spain

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

9

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

11

12 **Corresponding Author:**

13 Josep Piñol

14 CREA, Univ. Autònoma Barcelona, Cerdanyola del Vallès 08193, Spain

15 E-mail: Josep.Pinol@uab.es

16

17 **Running title:** Primers for quantitative metabarcoding

18

19 **Abstract**

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

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

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

40

41 **Introduction**

42 Ideally, metabarcoding should be a technique used to quantify species abundance in natural
43 communities (C_i in Figure 1) using high-throughput DNA sequencing (HTS). This is normally

44 accomplished by sampling the organisms in the community using a particular sampling method
45 (S_i ; Morinière et al., 2016) or by collecting fragments of DNA shed from organisms
46 (environmental DNA or eDNA, E_i ; Taberlet, Coissac, Hajibababei, & Rieseberg, 2012). The
47 target can also be the subset of the community consumed by a predator or an herbivore in
48 what is termed diet analysis (D_i ; Pompanon et al., 2012); the diet is estimated from the DNA
49 remains in faecal samples or in the gut contents of the consumer (G_i). In all cases, the DNA is
50 extracted into a solution with DNA of many species at relative abundances O_i . Then, the
51 extracted DNA can be directly sequenced (shotgun metagenomics) or sequenced following
52 amplification via PCR of one or more genomic regions (amplicon metabarcoding). Finally, the
53 obtained DNA reads R_i are assigned to species or OTUs (F_i). Every process described in
54 Figure 1 introduces its own biases (Leray & Knowlton, 2017; Pompanon et al., 2012), and so
55 the estimation of the community composition C_i (or D_i in diet analysis) from the final read
56 abundance (F_i) is a daunting task that we are now just beginning to grasp (Barnes & Turner,
57 2016; Elbrecht, Vamos, Meissner, Aroviita, & Leese, 2017). Only when all biases are avoided
58 or corrected, will it be possible to perform quantitative metabarcoding.

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

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

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

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

96 from the reads obtained after PCR amplification and HTS sequencing (R_i). We focus on this
97 for two reasons. First, the process $O_i \rightarrow R_i$ is an obligatory step for diet (G_i), eDNA (E_i), and
98 fresh or well conserved sample analyses (S_i), and so our contribution can potentially benefit
99 people in several fields. Second, whereas there are several causes that influence the number
100 of reads R_i (i.e. genome size, mitochondrial copy number, ...), the number of template-primer
101 mismatches is probably the most important one (Elbrecht & Leese, 2015; Mao, Zhou, Chen, &
102 Quan, 2012; Pinto & Raskin, 2012; Piñol, Mir, Gomez-Polo, & Agustí, 2015). We address the
103 problem using a simple model that simulates the process of PCR amplification in
104 heterogeneous mixtures. We test the model using the mitogenomes of ca. 1200 species of
105 insects available in RefSeq and 15 primer pairs targeting the COI region (Elbrecht & Leese,
106 2017a). Maybe the COI region is not the best suited for designing metabarcoding primers
107 (Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014; Elbrecht et al., 2016), but it remains
108 the region with most extensive information in genomic databases. The objectives are to
109 evaluate *in silico* which primer pairs and which characteristics of species mixtures provide a
110 quantitative relationship between the pre- and post-PCR marker abundance of the species in
111 the mixture.

112

113 **Material and Methods**

114 Rationale of the model

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

$$120 \quad F_i = \Lambda_i \cdot O_i \qquad \text{Eq. 1}$$

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

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

$$128 \quad f_i = \lambda_i \cdot o_i / a \qquad \text{Eq. 2}$$

129 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
130 amplification efficiency of species i and belongs to the interval (0, 1].

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

139 Primer pairs and species

140 We only considered the 15 COI primer pairs targeting the mitochondrial Folmer region (Folmer,
141 Black, Hoeh, Lutz, & Vrijenhoek, 1994) analysed by Elbrecht and Leese (2017a) (Tables 2 and
142 3). The selection includes the most common universal primer pairs currently used for DNA
143 metabarcoding of insects. We compiled a pool P of 1204 species of insects with an assembled

144 mitochondrial genome at RefSeq (<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/>
145 visited at 25-May-2017) (see the distribution of the species among orders in Figure S1).

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

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

156 Generation of random communities

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

$$160 \quad o_i = C_k \cdot k \cdot (1 - k)^{i-1} \quad \text{Eq. 3}$$

161 Where o_i is the proportion of the species i in the mixture, k is the parameter of the model and
162 $C_k = [1 - (1 - k)^S]^{-1}$ is a constant that makes $\sum_{i=1}^S o_i = 1$. The parameter k belongs to the
163 interval $(0, 1)$. Small values of k produce communities in which the species have similar
164 abundances, whereas high values of k produce communities dominated by a few abundant
165 species.

166 Estimation of the PCR amplification efficiency

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

$$171 \lambda_i = \beta^{-m_i} \quad \text{Eq. 4}$$

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

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

183 Parameters of the simulation

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

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

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

$$202 \quad \log_{10}y = -0.25 + 0.61 \cdot (-m_T), r^2 = 0.73, F_{1,9} = 24.35, P = 0.0008$$

$$203 \quad \log_{10}y = -1.09 + 1.73 \cdot (-m_5), r^2 = 0.81, F_{1,9} = 38.24, P = 0.0002$$

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

208 The above empirical relationships are equivalent to equation 4, where β can be estimated as
209 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
210 estimated values of the slope translates to the estimate of β : the 95% interval of confidence
211 for β in model 1 is (2.1 to 7.8) and for model 2 is (12.6 to 234).

212 Simulations

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

- 215 1. We generated a random number of species S in the community between 5 and 100.
216 (The random number and all those below followed a uniform distribution).
- 217 2. We randomly chose S species from the pool P of species.
- 218 3. We generated a random value of k between 0.2 and 0.95.
- 219 4. We used this k value to estimate the initial relative DNA concentration o_i of each of the
220 S species in the community using equation 3.
- 221 5. We generated a random value of β in the interval (2, 8) for model 1 and in the interval
222 (12, 240) for model 2.
- 223 6. Using equation 4, we calculated the relative amplification efficiency λ_i of each species
224 in the community using the above β value and the previously calculated number of
225 mismatches between this primer pair and each species. For model 1 we used the
226 number of template-primer mismatches in the entire length of the primers; for model 2
227 we used only the mismatches occurring in the five 3'-end positions of both primers.
- 228 7. Finally, we calculated the relative DNA concentration of each species in the mixture f_i
229 from Equation 2 using the above estimates of λ_i and o_i .
- 230 8. Each simulation was summarized by the linear correlation coefficient r_i between f_i and
231 o_i . We also calculated whether the most abundant species at the beginning (o_i) was
232 also the most abundant at the end of the PCR reaction (f_i).

233

234 Analysis of the results of the simulations

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

236 $H_0: r = 0$ (there is not a linear relationship between o_i and f_i)

237 $H_1: r \neq 0$ (there is a linear relationship between o_i and f_i)

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

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

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

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

250

251 **Results**

252 Number of mismatches

253 Considering the entire length of both the forward and the reverse primers, the median of the
254 number of template-primer mismatches was 0 for primer pairs #10 to #14, 1 for primer pair #7,
255 and 3 or higher for the rest of primer pairs (Figure 3A). When only the five 3'-end positions of
256 each primer were considered, the median of the number of template-primer mismatches was
257 2 for primer pair #4, 1 for primer pairs #2 and #5, and 0 for the remainder (Figure 3B). Primer
258 pairs #10 and #14 were especially good, as more than 99% of the tested species (~1150) had
259 no template-primer mismatches in the five 3'-terminal positions.

260 Relationship $o-r-f$

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

270 The overall picture was slightly better when only the five 3'-terminal bases of both primers were
271 considered (Model 2). In this case, only primer pairs #4 and #5 generated a CI for r that
272 included the 0 value, while the opposite was true for the rest of them (Figure 4B). Primer pairs
273 #10, #11, #12, and #14 were again especially good, generating CI that were always above
274 the value of $r=0.9$. In addition, for primers pairs #10 to #14, it is safe to assume that the most
275 abundant species was correctly attributed (Table 4).

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

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

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

286 Relative importance of each factor on the magnitude of the correlation ρ_{r-f_i}

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

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

305

306 **Discussion**

307 The model is intended to establish whether the results of a part of the metabarcoding analysis,
308 but not of the entire metabarcoding pipeline (Figure 1), are *likely* to be quantitative. By
309 quantitative we more precisely mean that there exists a significant linear correlation (at a
310 certain significance level) between the relative DNA concentration before and after the PCR

311 reaction (O_i and F_i in Figure 1) using a particular primer set and a group of organisms. What
312 the model does not provide is the *certainty* of a significant relationship for a given analysis.
313 This approach may have greater utility for the analysis of eDNA samples or community DNA
314 than for gut content analyses, given the additional sources of error associated with digestion.
315 Also, the number of different species in the mix is likely to be much lower in predator/herbivore
316 gut samples than in eDNA samples.

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

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

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

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

346 Choice of primer pairs

347 The choice of primer pairs is the most important decision to make for DNA metabarcoding. The
348 model suggests that some of the primers tested in this study are better suited than others for
349 quantitative DNA metabarcoding. Among the primer pairs tested here, the best choice seems
350 to be the primer pair #10 (Gibson et al., 2014) and the primer pairs #11 to #14 (Elbrecht &
351 Leese, 2017a). All of them guarantee (with a probability of 0.95) a significant linear relationship
352 o_r-f_i ; moreover, the linear correlation coefficient between pre- and post-PCR DNA
353 concentrations is likely to be high (in the range 0.4 – 1) (Figure 4). The rest of the primer pairs,
354 except #4 and #5, also provide significant results, but with lower r values. Primers sets #10 to
355 #14 are highly degenerated, so, it is justified that they amplify better *in silico* than other sets
356 with a much lower degeneracy (e.g., #4 and #5).

357 The primer pair #10 (Gibson et al., 2014) was developed to amplify a 310 bp region of many
358 families of arthropods. This primer contains the universal base inosine (I); in our calculations,
359 we considered that inosine could pair any base, but, in reality, its capacity to pair with the four
360 bases is variable (Martin, Castro, Aboul-ela, & Tinoco, 1985); besides, the use of inosine
361 increases the price of the primers. The four primer pairs of Elbrecht and Leese (2017a) are all

362 possible combinations of two forward and two reverse primers that produce amplicons of
363 different length. As these primers pairs were developed recently, they have been hardly used
364 by other researchers for DNA metabarcoding (Kreherwinkel et al., 2017); considering our
365 results, we would recommend their use. In any case, *in vivo* validations of primers sets with
366 mock samples of the species of interest is still advisable before embarking on a metabarcoding
367 study.

368 The above recommendation does not imply that the rest of the primers are of no use in
369 metabarcoding. Some of them can have great coverage of some groups, like primers #5 that
370 are particularly good for Lepidoptera (Zeale, Butlin, Barker, Lees, & Jones, 2011) and have
371 been used with profit for the characterisation of the diet of bats (Clare, Symondson, & Fenton,
372 2014).

373 Characteristics of the mixture of species

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

388 insectivorous bird or bat) than for a small predator (e.g. an insect) that may be polyphagous
389 but have few prey in its guts at any moment in time.

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

396 Relationship between the number of mismatches and amplification efficiency

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

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

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

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

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

427 Limitations of the model

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

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

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

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

461

462 **Acknowledgements**

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

467

468 **Literature**

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

472 Angly, F.E, Dennis, P.G., Skarshewski, A., Vanwonderghem, I., Hugenholtz, P., & Tyson, G.W.
473 (2014). CopyRighter: a rapid tool for improving the accuracy of microbial community profiles
474 through lineage-specific gene copy number correction. *Microbiome*, *2*, 11.

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

477 Bista, I., Carvalho, G.R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., ... Creer, S. (2018)
478 Performance of amplicon and shotgun sequencing for accurate biomass estimation in
479 invertebrate community samples. *Molecular Ecology Resources*, DOI:10.1111/1755-
480 0998.12888.

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

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

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

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

493 Clarke, L.J., Beard, J.M., Swadling, K.M., & Deagle, B.E. (2017). Effect of marker choice and
494 thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution*,
495 *7*, 873–883.

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

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

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

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

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

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

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

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

518 Evans, N.T., Olds, B.P., Renshaw, M.A., Turner, C.R., Li, Y., Jerde, C.L., ... Lodge, D.M.
519 (2016). Quantification of mesocosm fish and amphibian species diversity via DNA
520 metabarcoding. *Molecular Ecology Resources*, *16*, 29-41.

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

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

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

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

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

537 Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M.J., Baselga, A., & Vogler, A.P.
538 (2015). Validating the power of mitochondrial metagenomics for community ecology and
539 phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, *6*, 883-894.

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

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

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

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

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

555 Klymus, K.E., Marshall, N.T., & Stepien, C.A. (2017). Environmental DNA (eDNA)
556 metabarcoding assays to detect invasive invertebrate species in the Great Lakes. *PLoS ONE*,
557 *12*, e0177643.

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

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

564 Kwok, S., Kellogg, D.E., McKinney, N., Spasic, D., Godal, L., Levenson, C., & Sninsky, J.J.
565 (1990). Effects of primer-template mismatches on the polymerase chain reaction: Human
566 immunodeficiency virus type 1 model studies. *Nucleic Acids Research*, *18*, 999-1005.

567 Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., ... Machida, R.J.
568 (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region
569 for metabarcoding metazoan diversity; application for characterizing coral reef fish gut
570 contents. *Frontiers of Zoology*, *10*, 34.

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

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

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

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

578 Meusnier, I., Singer, G.A.C., Landry, J.F., Hickey, D.A., Hebert, P.D.N., & Hajibabaei, M.
579 (2008). A universal mini-barcode for biodiversity analysis. *BMC Genomics*, *9*, 214.

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

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

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

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

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

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

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

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

600 Pompanon, F., Deagle, B.E., Symondson, W.O.C., Brown, D.S., Jarman, S.N., & Taberlet, P.
601 (2012). Who is eating what: diet assessment using next generation sequencing. *Molecular
602 Ecology*, *21*, 1931-1950.

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

606 Pornon, A., Escaravage, N., Burrus, M., Holota, H., Khimoun, A., Mariette, J., ... Andalo, C.
607 (2016). Using metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific
608 Reports*, *6*, 27282.

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

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

614 Saitoh, S., Aoyama, H., Fujii, S., Sunagawa, H., Nagahama, H., Akutsu, M., ... Nakamori, T.
615 (2016). A quantitative protocol for DNA metabarcoding of springtails (Collembola). *Genome*,
616 *59*, 705-723.

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

620 Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K., & Nikolausz, M. (2007).
621 Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-
622 targeting bacterial community analysis. *FEMS Microbiol. Ecol.*, *60*, 341–350.

623 Stadhouders, R., Pas, S.D., Anber, J., Voermans, J., Mes, T.H., & Schutten, M. (2010). The
624 effect of primer-template mismatches on the detection and quantification of nucleic acids using
625 the 5' nuclease assay. *The Journal of Molecular Diagnostics*, *12*, 109-117.

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

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

- 631 Thomsen, P.F., Møller, P.R., Sigsgaard, E.E., Knudsen, S.W., Jørgensen, O.A., & Willerslev,
632 E. (2016). Environmental DNA from seawater samples correlate with trawl catches of
633 subarctic, deepwater fishes. *PLoS ONE*, *11*, e0165252.
- 634 Van Houdt, J.K.J., Breman, F.C., Virgilio, M., & De Meyer, M. (2010). Recovering full DNA
635 barcodes from natural history collections of Tephritid fruitflies (Tephritidae, Diptera) using mini
636 barcodes. *Molecular Ecology Resources*, *10*, 459-465.
- 637 Wintzingerode, F.V., Göbel, U.B., & Stackebrandt, E. (1997). Determination of microbial
638 diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology*
639 *Reviews*, *21*, 213-229.
- 640 Wright, E.S., Yilmaz, L.S., Ram, S., Gasser, J.M., Harrington, G.W. & Noguera, D.R. (2014).
641 Exploiting extension bias in polymerase chain reaction to improve primer specificity in
642 ensembles of nearly identical DNA templates. *Environmental Microbiology*, *16*, 1354-1365.
- 643 Zeale, M.R.K., Butlin, R.K., Barker, G.L.A., Lees, D.C., & Jones, G. (2011). Taxon-specific
644 PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources*, *11*, 236–
645 244.
- 646 Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., ... Huang, Q. (2013). Ultra-deep sequencing
647 enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR
648 amplification. *GigaScience*, *2*, 4.

649

650 **Authors contribution**

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

653

654 **Data Accessibility**

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

657

658 **Supporting information**

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

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

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

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

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

666

667
668

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, Vizcaino, 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)

669

670 **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)
HCO2198	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	AGATATTGGAACWTTATATTTTATTTTGG	Zeale et al. (2010)
ZBJ-ArtR2c	R	WACTAATCAATTWCCAAATCCTCC	Zeale et al. (2010)
mICOIntF	F	GGWACWGGWTGAACWGTWTAYCCYCC	Leray et al. (2013)
mICOIntR	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)
III_C_R	R	GGIGGRTAIACIGTTCAICC	Shokralla et al. (2015)
III_B_F	F	CCIGAYATRGCIITYCCICG	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	TCDDGGRTGNCCRAARAAYCA	Elbrecht and Leese (2017a)
ArF5	F	GCICIGAYATRCIITYCCICG	Gibson et al. (2014)
ArR5	R	GTRATIGCICIGCIARIACIGG	Gibson et al. (2014)
jpgLCO1490	F	TITCIACIAAYCAYAARGAYATTGG	Geller, Meyer, Parker, & Hawk (2013)
jpgHCO2198	R	TAIACYTCIGGRTGICCRARAAYCA	Geller et al. (2013)
L499	F	ATTAATATACGATCAACAGGAAT	Van Houdt, Breman, Virgilio, & De Meyer (2010)
H2123d	R	TAWACTTCWGGRTGWCCAAARAATCA	Van Houdt et al. (2010)

671

672 **Table 3.** Primer pairs used in this study. Primer names as in table 2. Primer pair #15 was
 673 not further considered because it provided much fewer species with useful data than the
 674 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	mICOLintR	319	944
#7	mICOLintF	jgHCO2198	313	1162
#8	LCO1490	III_C_R	325	1014
#9	III_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

675

676

677 **Table 4.** Proportion of runs in which the same species is most abundant both before and
678 after the simulated PCR reaction for Model 1 and Model 2 simulations. Proportion based
679 on 10000 simulation runs per primer pair. It is indicated in bold face whether it is safe (at
680 $\alpha = 0.95$) 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

681

682

683 **Table 5.** Percentage of the variance in the linear correlation coefficient r explained by
684 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

685

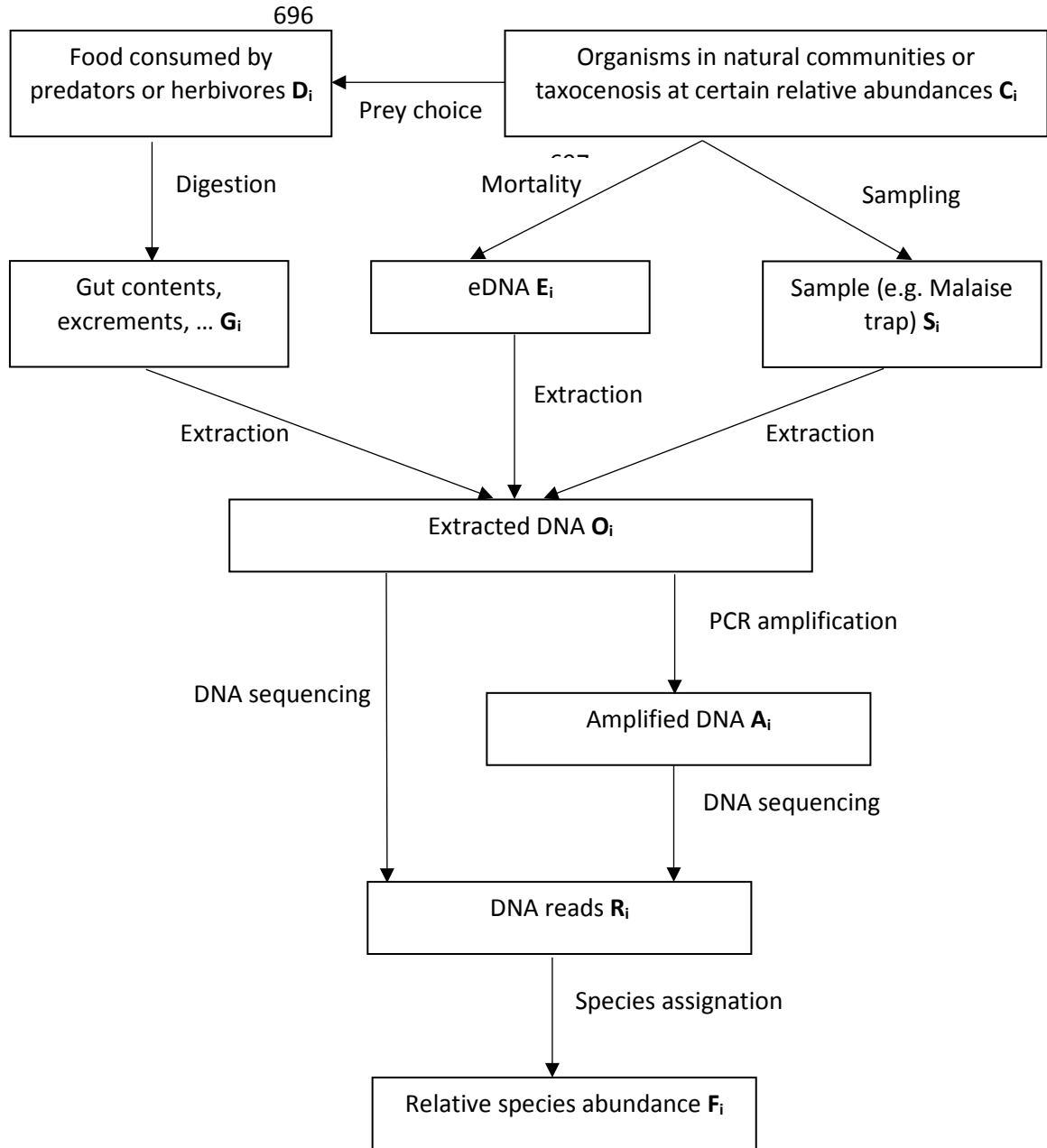
686 **Table 6.** Two hypothetical mixtures of ten species, with the number of mismatches of
 687 each species, its original and final DNA concentration (o_i and f_i), and its amplification
 688 efficiency ($\beta = 4$). The mixture B is the same as the mixture A, but for the swap of o_i of
 689 species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

A				
#sp	m_{π_i}	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_{π_i}	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

690

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



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

706

707 run once run many times

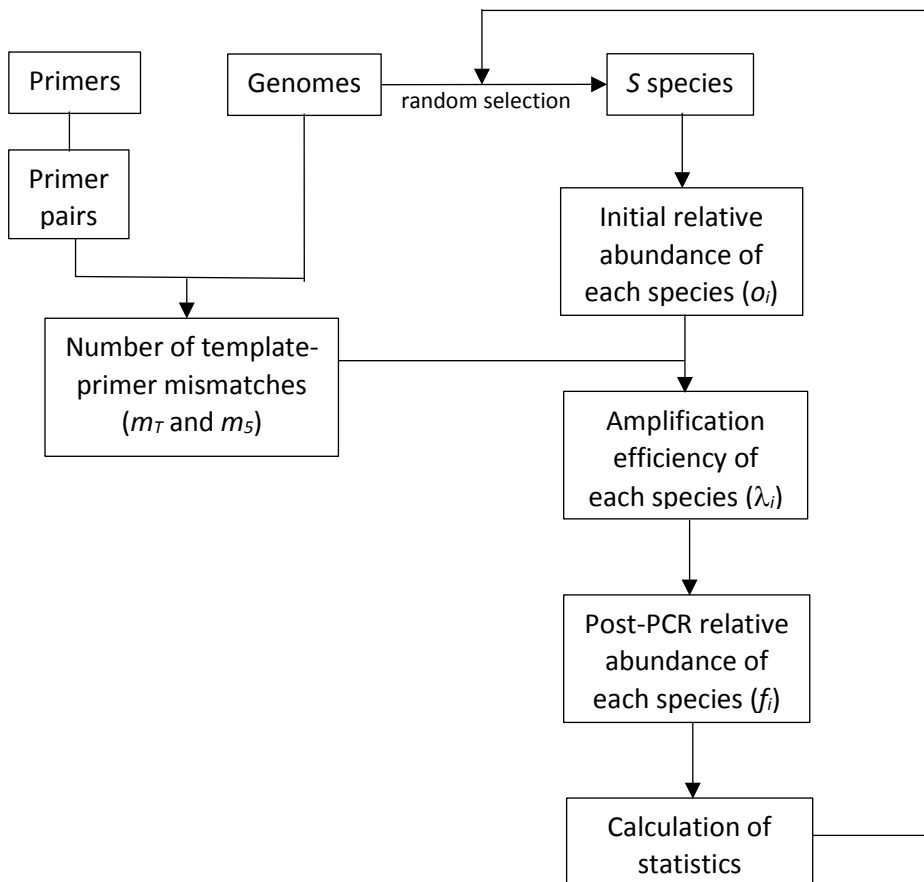
708

709

710

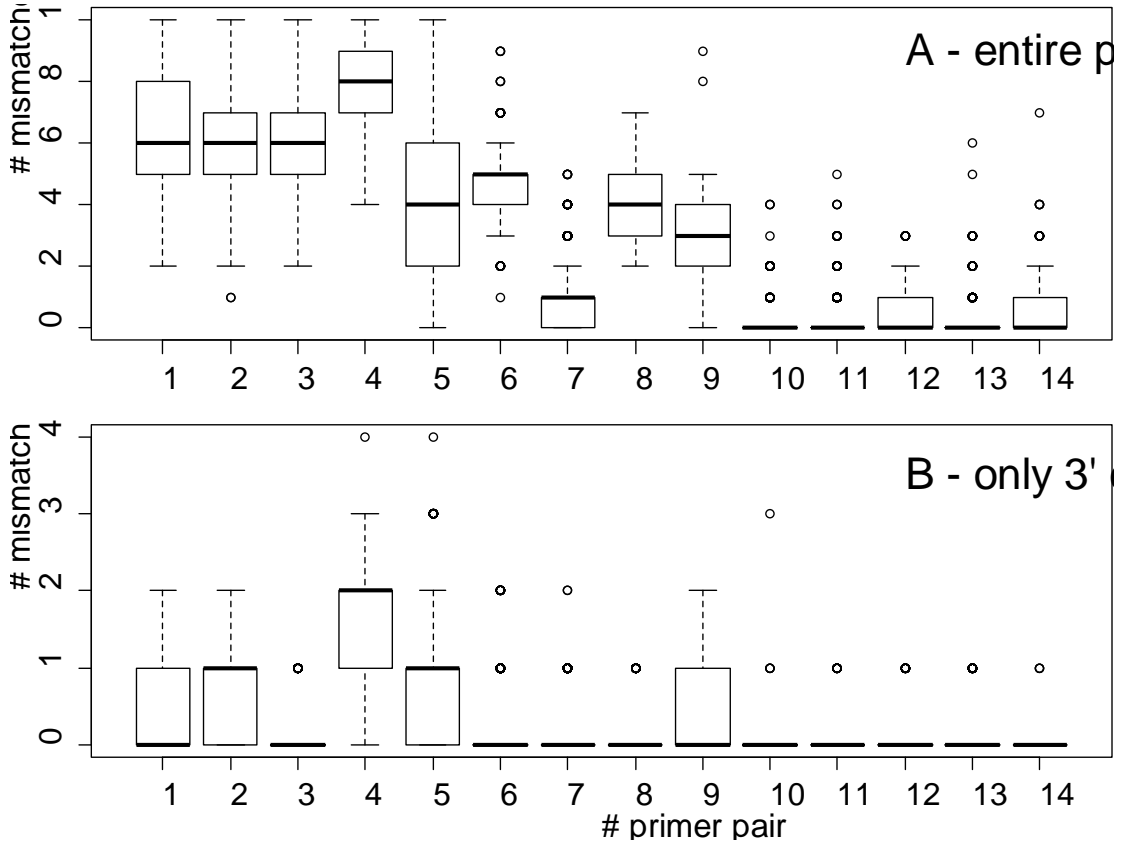
711

712



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

717

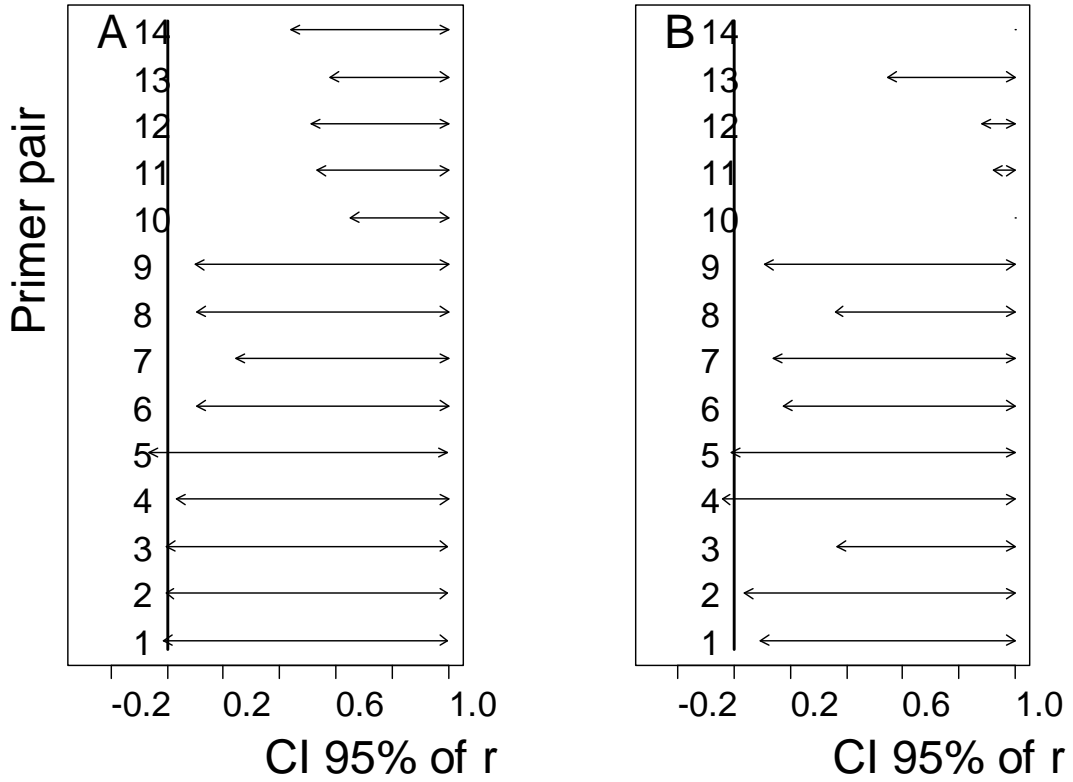


718

719

720

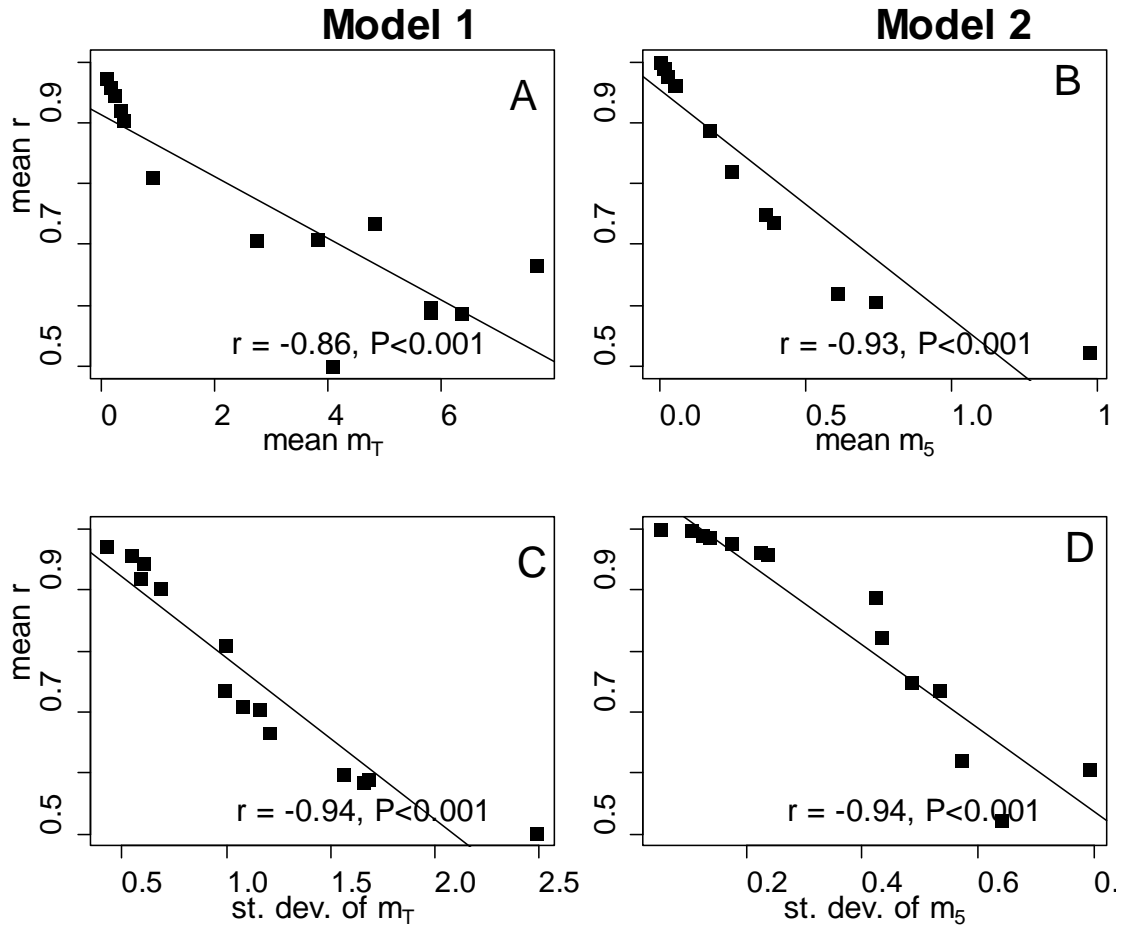
721 **Figure 4.** Ninety-five percent confidence interval (95% CI) for the Pearson correlation r
 722 for each primer pair analysed for Model 1 (A) and Model 2 (B). For primer pairs #10 and
 723 #14 in B, the CI is so small that the arrowheads could not be plotted. When the CI cuts
 724 the vertical line at $r = 0$ it indicates that it is not possible to consider that $r > 0$ (with a
 725 probability of 0.95). Primer pair numbering is the same as in Table 3.



726

727

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



733