MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 543–556 doi: 10.1111/1755-0998.12338

# Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data

GENTILE F. FICETOLA, JOHAN PANSU, AURÉLIE BONIN, ERIC COISSAC, CHARLINE GIGUET-COVEX, MARTA DE BARBA, LUDOVIC GIELLY, CARLA M. LOPES, FRÉDÉRIC BOYER, FRANÇOIS POMPANON GILLES RAYÉ and PIERRE TABERLET
*Laboratoire d'Ecologie Alpine (LECA), Université Grenoble Alpes Savoie, F-38000 Grenoble, France*

## Abstract

**Environmental DNA (eDNA) metabarcoding is increasingly used to study the present and past biodiversity. eDNA analyses often rely on amplification of very small quantities or degraded DNA. To avoid missing detection of taxa that are actually present (false negatives), multiple extractions and amplifications of the same samples are often performed. However, the level of replication needed for reliable estimates of the presence/absence patterns remains an unaddressed topic. Furthermore, degraded DNA and PCR/sequencing errors might produce false positives. We used simulations and empirical data to evaluate the level of replication required for accurate detection of targeted taxa in different contexts and to assess the performance of methods used to reduce the risk of false detections. Furthermore, we evaluated whether statistical approaches developed to estimate occupancy in the presence of observational errors can successfully estimate true prevalence, detection probability and false-positive rates. Replications reduced the rate of false negatives; the optimal level of replication was strongly dependent on the detection probability of taxa. Occupancy models successfully estimated true prevalence, detection probability and false-positive rates, but their performance increased with the number of replicates. At least eight PCR replicates should be performed if detection probability is not high, such as in ancient DNA studies. Multiple DNA extractions from the same sample yielded consistent results; in some cases, collecting multiple samples from the same locality allowed detecting more species. The optimal level of replication for accurate species detection strongly varies among studies and could be explicitly estimated to improve the reliability of results.**

*Keywords*: detectability, earthworms, false negatives, false-positive detection, lake sediments, occupancy modelling, simulations, species occurrence

*Received 9 July 2014; revision received 27 September 2014; accepted 7 October 2014*

## Introduction

Environmental DNA (eDNA) metabarcoding, that is the simultaneous identification of multiple taxa from the DNA extracted from an environmental sample (e.g. soil, water and faeces samples), is an emerging approach for the study of the present and past biodiversity (Valentini *et al.* 2009b; Taberlet *et al.* 2012a). The analysis of eDNA has an increasing number of applications, such as the description of biodiversity of microbes, plants and animals from a wide range of environments (e.g. Bienert *et al.* 2012; Yoccoz *et al.* 2012; Zinger *et al.* 2012), the analysis of diet (e.g. Deagle *et al.* 2005, 2009; Valentini *et al.* 2009a; Pompanon *et al.* 2012; De Barba *et al.* 2014), the reconstruction of the past biodiversity and/or environmental changes (Jorgensen *et al.* 2012; Parducci *et al.* 2013; Boessenkool *et al.* 2014; Giguet-Covex *et al.* 2014; Willerslev *et al.* 2014) and environmental monitoring (Jerde *et al.* 2011, 2013; Hajibabaei *et al.* 2012; Darling 2014; Mahon *et al.* 2014; Nathan *et al.* 2014).

Analysis and identification of eDNA is used mainly for two purposes in ecological studies. First, considering a sample, we want to infer the list of all the taxa present in the sampled environment, like in many studies on microbial diversity (e.g. Zinger *et al.* 2012). Second, considering a set of samples, we want to infer in which of them a species or a set of species are present or absent (e.g. Giguet-Covex *et al.* 2014). These types of analysis are not faced with the same challenges. In the first case, a major problem is to determine within the full list of observed sequences which one corresponds to true species and which one corresponds to experimental artefacts (PCR or sequencing induced errors). Conversely,

Correspondence: Gentile F. Ficetola, Fax: +33 476 51 42 79;
E-mail: francesco.ficetola@gmail.com

the second case corresponds to studies in which we are interested in deciphering the distribution of a defined set of species, or in asserting the presence/absence of key taxa in a given environment. Like with other conventional inventory techniques, it is extremely important to take into account the possibility that the present taxa remain undetected (false negatives), but also the occurrence of false observations of the absent taxa (false positives) (Darling & Mahon 2011). Despite the high sensitivity of techniques based on eDNA, researchers are well aware that species detection using this approach is imperfect. Indeed, these analyses are usually performed with very little starting material and stochastic processes are often implicated as to whether a given DNA molecule is amplified. For instance, Ficetola *et al.* (2008) reported several PCRs in which the DNA of a target species was not detected, despite being present in the environmental sample. To cope with this issue, eDNA studies often perform replicated analyses: they can rely on multiple environmental samples, perform multiple extractions from the same environmental sample and/or multiple amplifications of the same extracted DNA. The number of replicates per environmental sample can be highly variable among studies, ranging from two to >10 (e.g. Ficetola *et al.* 2008; Jerde *et al.* 2011, 2013; Parducci *et al.* 2013; Giguet-Covex *et al.* 2014; Willerslev *et al.* 2014). Increasing the number of replicates certainly reduces the risk of missing the present taxa, but inflates costs and workload. Furthermore, false positives are always possible, and the risk of false positives might increase with the replication level.

False positives are a particularly crucial issue in ancient DNA metabarcoding studies performed on environmental samples because the low amount of highly degraded template DNA requires many PCR cycles. False positives may for instance arise because of PCR or sequencing errors not detected by dedicated software, or through sporadic contamination (Willerslev *et al.* 2014). To limit false positives, studies performed on degraded DNA use multiple quality assurance practices (e.g. sampling of localities where target taxa are notoriously absent, extraction blanks and equipment controls) and, additionally, sometimes consider a sequence only if it was confirmed by at least two independent PCRs, while those detected in one replicate only may be discarded or considered dubious (Giguet-Covex *et al.* 2014; Willerslev *et al.* 2014). Nevertheless, this approach has drawbacks, as it might overlook taxa that are actually present at low density or very difficult to detect.

Imperfect detection is an unavoidable feature of most data on species presence/absence and abundance: even during traditional ecological field studies, individuals and species that are present at one site are not always all detected, and failure in accounting for imperfect detection may result in biased inference (Yoccoz *et al.* 2001; MacKenzie *et al.* 2006; Kery & Schmidt 2008; Lahoz-Monfort *et al.* 2014). Several models have been proposed by ecologists to limit these issues. In the last decade, species occupancy models (SOMs) have been developed to analyse species distribution when detection probability is lower than one. In short, SOMs can use data on detection and nondetection at multiple occasions in a number of sites to evaluate the detection probability of species when they are present. SOMs can allow estimating the number of sites in which a target species is present but remain undetected and can be integrated with other analytical methods to better understand population dynamics and relationships between species and habitat (MacKenzie *et al.* 2006; Kéry & Schaub 2012). SOMs have been developed to analyse species distribution data obtained in the context of traditional field studies (MacKenzie *et al.* 2002, 2006), but recent work showed that SOMs can be successfully applied to the analysis of eDNA data (Pilliod *et al.* 2013; Schmidt *et al.* 2013). SOMs may therefore allow the evaluation of the probability of taxa detection through eDNA, and the estimation of the number of replicates required for reliable inference of taxon absence (Schmidt *et al.* 2013).

Species occupancy models were first developed to analyse the data in which a species may remain undetected (i.e. false absences are possible) but, when detected, a presence is always considered as genuine (i.e. false presences are impossible) (MacKenzie *et al.* 2002, 2006). However, misidentification of species is possible even with traditional ecological data, and SOMs have thus been expanded to account for potential false presences (Royle & Link 2006; Miller *et al.* 2011). These approaches have never been applied in the context of eDNA metabarcoding studies, but can be extremely valuable to obtain measures of confidence on taxa absences, to evaluate the possibility of false presences and to estimate the optimal number of replicates required for robust results.

In this study, we applied the use of SOMs to the analysis of eDNA metabarcoding results and estimated the number of technical replicates required to confidently estimate the presence/absence of a taxon in a given environment. We also evaluated the performance of approaches currently used to control for false presence. First, we analysed simulated data with known properties to answer the following questions: (i) what is the optimal number of PCR replicates for present and ancient eDNA metabarcoding analyses?; (ii) what is the impact of imperfect detection, false presences and number of replicates on the results of eDNA metabarcoding studies?; (iii) what is the impact of dismissing taxa detected in only one of the replicated PCRs?; (iv) do occupancy models allow estimating the frequency of

false presences/false absences in eDNA metabarcoding data?

Second, we applied occupancy models to empirical eDNA metabarcoding data, to evaluate their detection probability and rate of false presences. We compared two different typologies of eDNA metabarcoding data (earthworm DNA from present-day soils and ancient DNA of mammals from lake sediment cores) to highlight the range of differences occurring among eDNA metabarcoding data sets. Finally, we evaluated whether different replication approaches (multiple samples per site, multiple DNA extractions per sample and multiple amplifications per DNA extract) may influence the study outcome.

## Materials and methods

### *Simulations*

We simulated data with known properties, mimicking patterns of taxa detection/nondetection in eDNA metabarcoding studies. We generated data sets representing 100 environmental samples for which one taxon was analysed, assuming that the taxon has a probability of presence in each sample = 0.3 (latent occupancy, i.e. the true, unobserved occupancy of the taxon; MacKenzie *et al.* 2006). For each data set, we repeated 100 simulations per combination of parameter sets ($p$, $fp$ and $Nr$, see below). Different values of taxon detection probability $p$ ($p$ = 0.25, 0.5 and 0.75) and different occurrences of false presences $fp$ ($fp$ = 0.002, 0.01 and 0.03) were used in simulations. These values of $p$ represent taxa ranging from very low to very high detection probability, which may represent differences of abundance/biomass among taxa, or differences between present and ancient DNA studies (see Results, section 'Analysis of empirical data'). Similarly, the $fp$ values represent error rates ranging from very low to moderate. Per each combination of the parameters $p$ and $fp$, we generated patterns of taxon detection/nondetection, with different numbers of replicates $Nr$ ($Nr$ = 4, 6, 8 and 12). This number of replicates reflects the range commonly observed in eDNA studies (e.g. De Barba *et al.* 2014; Giguet-Covex *et al.* 2014; Willerslev *et al.* 2014). In our analysis, the number of replicates represents the overall number of PCRs performed on each environmental sample. In empirical studies, a given level of replication may be reached through different ways (e.g. eight replicates may be obtained by performing one DNA extraction per sample and then eight PCRs on the extract, or by performing two extractions per sample and four PCRs per extraction). For simplicity, in our simulations, the number of replicates just represents the total number of PCRs, independently on how they were obtained, as analyses of empirical data suggest that differences between replicating at the amplification level or also at the extraction level are small (see Results section).

We then analysed the simulated data using different approaches:

1. *Naive approach* assuming perfect detection and no false positives, and presence of a taxon at one site if it was detected at least once. We estimated the frequency of false absences (i.e. the number of samples in which the taxon was present but remained undetected) and the frequency of false presences (i.e. the number of samples in which the taxon was absent, but was erroneously detected because of false positives) by comparing the patterns of detection/nondetection with the true simulated data set. The frequency of false absences and false presences was also calculated analytically using conditional probability, on the basis of the known values of $p$, $fp$ and $Nr$. Specifically, the overall probability of false presences across the $Nr$ replicates was calculated as 1 − the probability of not finding any false presence [i.e. $1 - (1 - fp)^{Nr}$], while the probability of false absences was calculated as the probability of obtaining $Nr$ false absences, conditional to the probability of not finding any false presence [i.e. $(1 - p)^{Nr} \times (1 - (1 - fp)^{Nr})$].
2. *Conservative approach* assuming perfect detection but possibility of false presences. Taxon presence was considered 'uncertain' if it was detected only once out of the $Nr$ replicates. Uncertain presences were discarded, and we then estimated the frequency of remaining false presences and the number of true presences incorrectly removed.
3. *MacKenzie occupancy model*: This approach analyses replicated data of taxon detection/nondetection assuming that detection probability is <1, that is false absences are possible, while supposing that all detections are correct (i.e. no false presences) (MacKenzie *et al.* 2002). Using this approach, we estimated the detection probability of the taxon and its occupancy (the frequency of samples where the taxon is actually present). As this approach hypothesizes that all detections are correct, we assumed that a taxon was absent if taxon presence was uncertain (i.e. if it was detected in one replicate only).
4. *Miller occupancy model*: This approach analyses data of taxon detection/nondetection assuming that detection probability is <1 and that false presences are possible (Miller *et al.* 2011). We classified taxon presences as certain if the taxon was detected in at least two replicates and uncertain if it was detected in one replicate only. Using this approach, we estimated the taxon detection probability, the rate of false presences and the occupancy.

Both occupancy models were run using the package unmarked in R 3.0 (Fiske & Chandler 2011; R Development Core Team 2013). The R code used for simulation is available as Supporting Information (Appendix S1).

### *Empirical data – molecular methods*

*Earthworm data.* We sampled 12 sites of 1 ha in the Vercors massif (Northern French Alps), of which five were situated in pasture, four in deciduous forests and three in coniferous forests. In each site, we collected 100 core samples (about 50 g each) of mineral soil and litter every 10 m over a regular grid of 100 × 100 m and pooled them together. The sampling procedure was repeated twice for each site. Extracellular DNA extraction was performed using the protocol described by Taberlet *et al.* (2012b). For each soil sample, we carried out two extractions (i.e. four extractions per site) and two different PCRs per extraction (total: eight PCR replicates per site). DNA was amplified using the ewD/ewE primers that target short sequences (about 70 bp) on the mitochondrial 16S gene (Bienert *et al.* 2012). Amplicons were then sequenced on a high-throughput sequencer (Illumina HiSeq 2000, 2 × 100 bp, pair-end reads; Illumina inc., www.illumina.com/). Finally, sequences were filtered to remove PCR/sequencing errors and chimeras using the OBITools software suite (http://metabarcoding.org/obitools) and assigned to the relevant taxon by comparison to a reference database (J. Pansu, S. De Danieli, J. Puissant, J.-M. Gonzalez, L. Gielly, L. Zinger, J.-J. Brun, P. Choler, P. Taberlet & L. Cecillon, unpublished). Only sequences assigned up to the species level with >95% similarity were used (J. Pansu, S. De Danieli, J. Puissant, J.-M. Gonzalez, L. Gielly, L. Zinger, J.-J. Brun, P. Choler, P. Taberlet & L. Cecillon, unpublished; see Appendix S2 for additional methodological details, Supporting information).

*Ancient DNA from lake sediment cores.* We analysed ancient DNA data previously published by Giguet-Covex *et al.* (2014). Giguet-Covex *et al.* (2014) extracted ancient DNA from a 20.2-m-long sediment core from Lake Anterne (2063 m asl, Northern French Alps); 47 slices of approximately 1 cm thickness, were sampled, corresponding to 10 160 cal. before present (i.e. 8210 BC) to nowadays. Each sample was divided in two subsamples, with two DNA extractions per subsample (i.e. four per sample) and two independent PCRs per extraction, resulting in eight amplification replicates per core slice. Mammal DNA was amplified using the MamP007 primers pair targeting a 60- to 84-bp fragment on the mitochondrial 16S gene (Giguet-Covex *et al.* 2014). Sequencing was performed using the Illumina HiSeq 2000 platform (2 × 100 bp, pair-end reads). PCR/sequencing errors and chimeras were also filtered out from the obtained data set, and sequences were assigned to the relevant taxa (see Giguet-Covex *et al.* 2014 for details).

### *Empirical data – data analysis*

To evaluate the detection probability, false presences and occupancy, we analysed the empirical data using the MacKenzie *et al.* (2002) (earthworm data set) and the Miller *et al.* (2011) (ancient DNA data set) occupancy models, following the same procedure of simulated data. We used two different approaches for the two data sets because uncertain presences only occurred in the ancient DNA data sets (see Results).

For empirical data, three different approaches were used to obtain replicates: multiple samples from the same site/core sample, multiple DNA extractions from the same sample and multiple amplifications on the same DNA extract. We therefore used the analysis of similarity (ANOSIM) to evaluate whether there are significant differences among communities obtained from different environmental samples from the same site, or from different DNA extractions of the same sample (Legendre & Legendre 2012). Similarity among communities was evaluated using the Bray–Curtis distance (Legendre & Legendre 2012); significance of ANOSIM was assessed through 999 simulations using VEGAN (Oksanen *et al.* 2013; R Development Core Team 2013). This analysis was not performed for the ancient DNA data, because many PCRs did not detect the DNA of any taxon, therefore hampering most of pairwise comparisons among replicates.

## Results

### *Analysis of simulated data*

*Naive approach.* When assuming perfect detection, taxon occupancy was severely underestimated if the number of replicates was low, and if detection probability was low, results of simulations and analytical solutions showed high concordance (Table 1, Fig. 1a–d). For instance, if $p = 0.25$ and $Nr = 4$, the taxon remained undetected in about one-third of samples where it was actually present (Table 1, Fig. 1a). Six replicates were needed to avoid the false absences if detection probability = 0.5, and 12 replicates were needed if detection probability was very low (Fig. 1b–d). However, with this approach, a high number of false positives exist in the data, particularly if *fp* was high and many replicates were run (Fig. 1e–h, Table 1). With many replicates, the number of false positives remained limited only if *fp* was very low.

*Conservative approach.* If detection in one replicate only was considered uncertain and discarded, the false positive rate was very low (Fig. 2e–h), with false presences

**Table 1** Frequency of false absences and false presences obtained using the naïve approach, estimated analytically and on the basis of simulations

| Detection probability | *N* replicates | False presences | False absences | |
|---|---|---|---|---|
| | | | Analytical | Simulations |
| 0.25 | 4 | 0.03 | 0.280 | 0.267 |
| 0.25 | 4 | 0.02 | 0.304 | 0.267 |
| 0.25 | 4 | 0.002 | 0.314 | 0.317 |
| 0.25 | 6 | 0.03 | 0.148 | 0.133 |
| 0.25 | 6 | 0.02 | 0.168 | 0.167 |
| 0.25 | 6 | 0.002 | 0.176 | 0.167 |
| 0.25 | 8 | 0.03 | 0.078 | 0.067 |
| 0.25 | 8 | 0.02 | 0.092 | 0.100 |
| 0.25 | 8 | 0.002 | 0.099 | 0.100 |
| 0.25 | 12 | 0.03 | 0.022 | 0.000 |
| 0.25 | 12 | 0.02 | 0.028 | 0.033 |
| 0.25 | 12 | 0.002 | 0.031 | 0.033 |
| 0.50 | 4 | 0.03 | 0.055 | 0.067 |
| 0.50 | 4 | 0.02 | 0.060 | 0.067 |
| 0.50 | 4 | 0.002 | 0.062 | 0.067 |
| 0.50 | 6 | 0.03 | 0.013 | 0.000 |
| 0.50 | 6 | 0.02 | 0.015 | 0.000 |
| 0.50 | 6 | 0.002 | 0.015 | 0.000 |
| 0.50 | 8 | 0.03 | 0.003 | 0.000 |
| 0.50 | 8 | 0.02 | 0.004 | 0.000 |
| 0.50 | 8 | 0.002 | 0.004 | 0.067 |
| 0.50 | 12 | 0.03 | <0.001 | 0.000 |
| 0.50 | 12 | 0.02 | <0.001 | 0.000 |
| 0.50 | 12 | 0.002 | <0.001 | 0.000 |
| 0.75 | 4 | 0.03 | 0.003 | 0.000 |
| 0.75 | 4 | 0.02 | 0.004 | 0.000 |
| 0.75 | 4 | 0.002 | 0.004 | 0.000 |
| 0.75 | 6 | 0.03 | <0.001 | 0.000 |
| 0.75 | 6 | 0.02 | <0.001 | 0.000 |
| 0.75 | 6 | 0.002 | <0.001 | 0.000 |
| 0.75 | 8 | 0.03 | <0.001 | 0.000 |
| 0.75 | 8 | 0.02 | <0.001 | 0.000 |
| 0.75 | 8 | 0.002 | <0.001 | 0.000 |
| 0.75 | 12 | 0.03 | <0.001 | 0.000 |
| 0.75 | 12 | 0.02 | <0.001 | 0.000 |
| 0.75 | 12 | 0.002 | <0.001 | 0.000 |

| * | *N* replicates | False presences | False absences | |
|---|---|---|---|---|
| | | | Analytical | Simulations |
| | 4 | 0.03 | 0.115 | 0.114 |
| | 4 | 0.02 | 0.039 | 0.038 |
| | 4 | 0.002 | 0.008 | 0.000 |
| | 6 | 0.03 | 0.167 | 0.171 |
| | 6 | 0.02 | 0.059 | 0.057 |
| | 6 | 0.002 | 0.012 | 0.014 |
| | 8 | 0.03 | 0.216 | 0.210 |
| | 8 | 0.02 | 0.077 | 0.079 |
| | 8 | 0.002 | 0.016 | 0.005 |
| | 12 | 0.03 | 0.306 | 0.307 |
| | 12 | 0.02 | 0.114 | 0.110 |
| | 12 | 0.002 | 0.024 | 0.014 |

*Results are unaffected by detection probability, as the false absences are estimated on the sites where the species is actually absent.

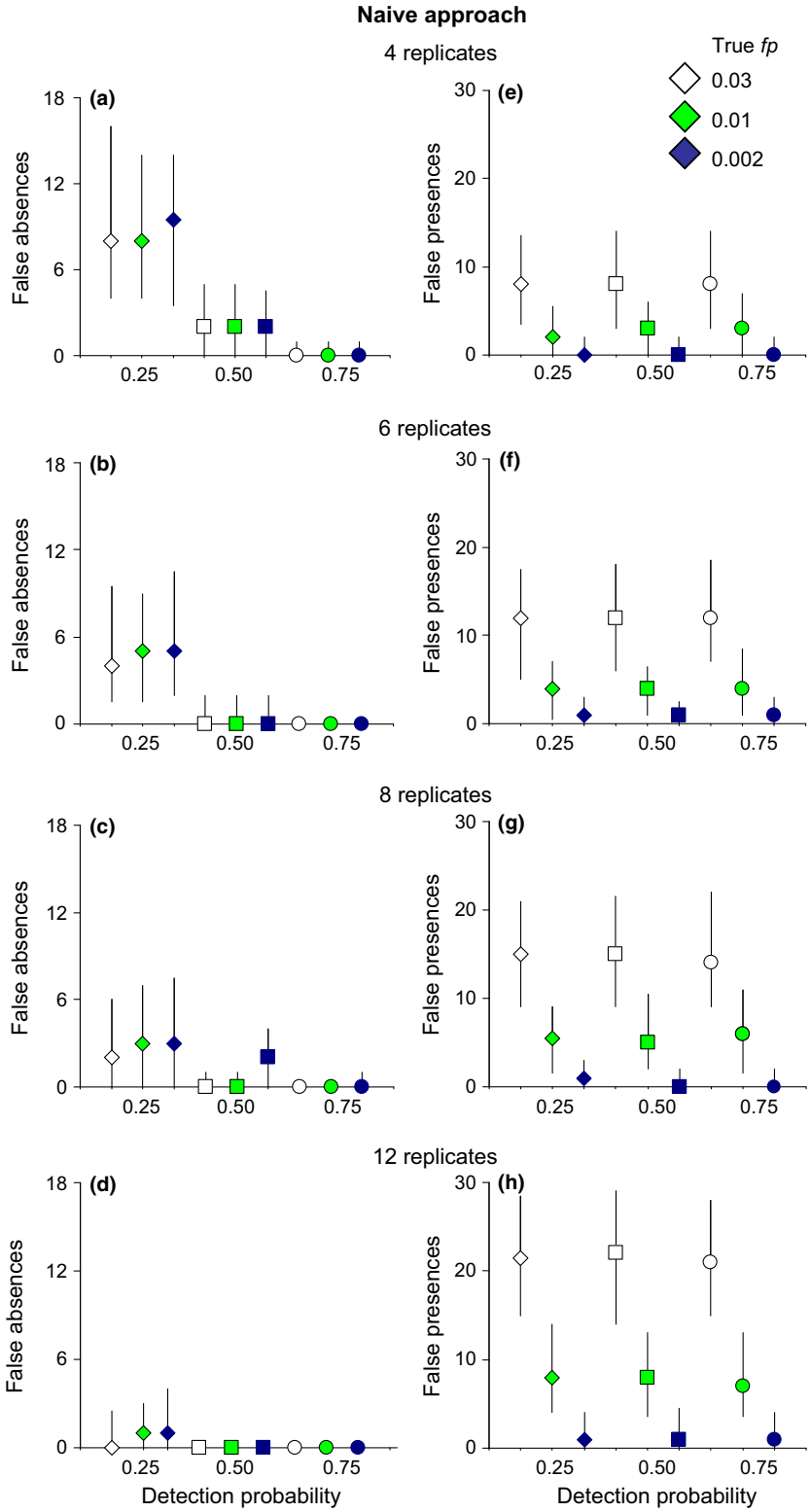


**Fig. 1** Analysis of simulated data using the Naive approach. (a–d) The number of simulated environmental samples in which the species was incorrectly assumed to be absent, depending on the number of replicates performed; (e–h) number of simulated environmental samples in which the species was incorrectly assumed to be present, depending on the number of replicates performed. Symbols represents medians ± 95% CI: *fp*, frequency of false presences; diamonds, detection probability = 0.25; squares, detection probability = 0.5; circles, detection probability = 0.75.

remaining in the data only if *fp* was high and many replicates were run (Fig. 2h). However, this conservative approach incorrectly removed many true presences (Fig. 2a–d). For instance, if only four replicates were run and detection probability was low, this approach removed two-third of true presences. The number of incorrectly removed presences was limited only if detection probability was very high ($p = 0.75$), or if many rep-

licates were run (8–12 replicates, depending on detection probability) (Fig. 2 a–d).

*MacKenzie occupancy model.* The accuracy of the MacKenzie *et al.* (2002) occupancy model in the estimation of detection probability strongly depended on the number of replicates and on the actual values of $p$ (Fig. 3a–d). With just four replicates, detection probability was correctly estimated only if it was high (0.75, Fig. 3a). Six replicates were needed to correctly estimate a detection

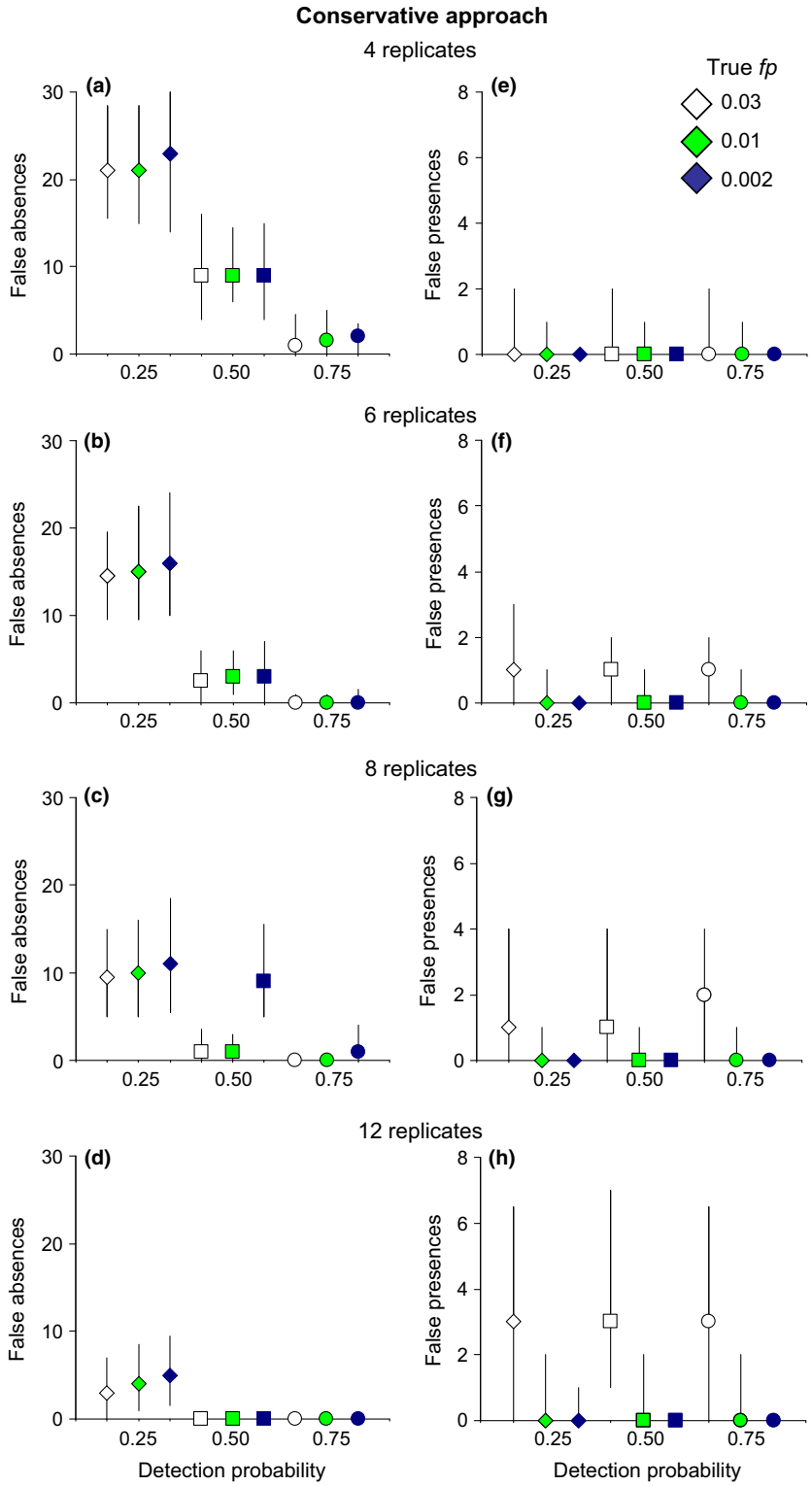


**Fig. 2** Analysis of simulated data using the conservative approach. (a–d) The number of simulated environmental samples in which the species was incorrectly assumed to be absent, depending on the number of replicates performed; (e–h) number of simulated environmental samples in which the species was incorrectly assumed to be present, depending on the number of replicates performed. Symbols represents medians ± 95% CI: *fp*, frequency of false presences; diamonds, detection probability = 0.25; squares, detection probability = 0.5; circles, detection probability = 0.75.

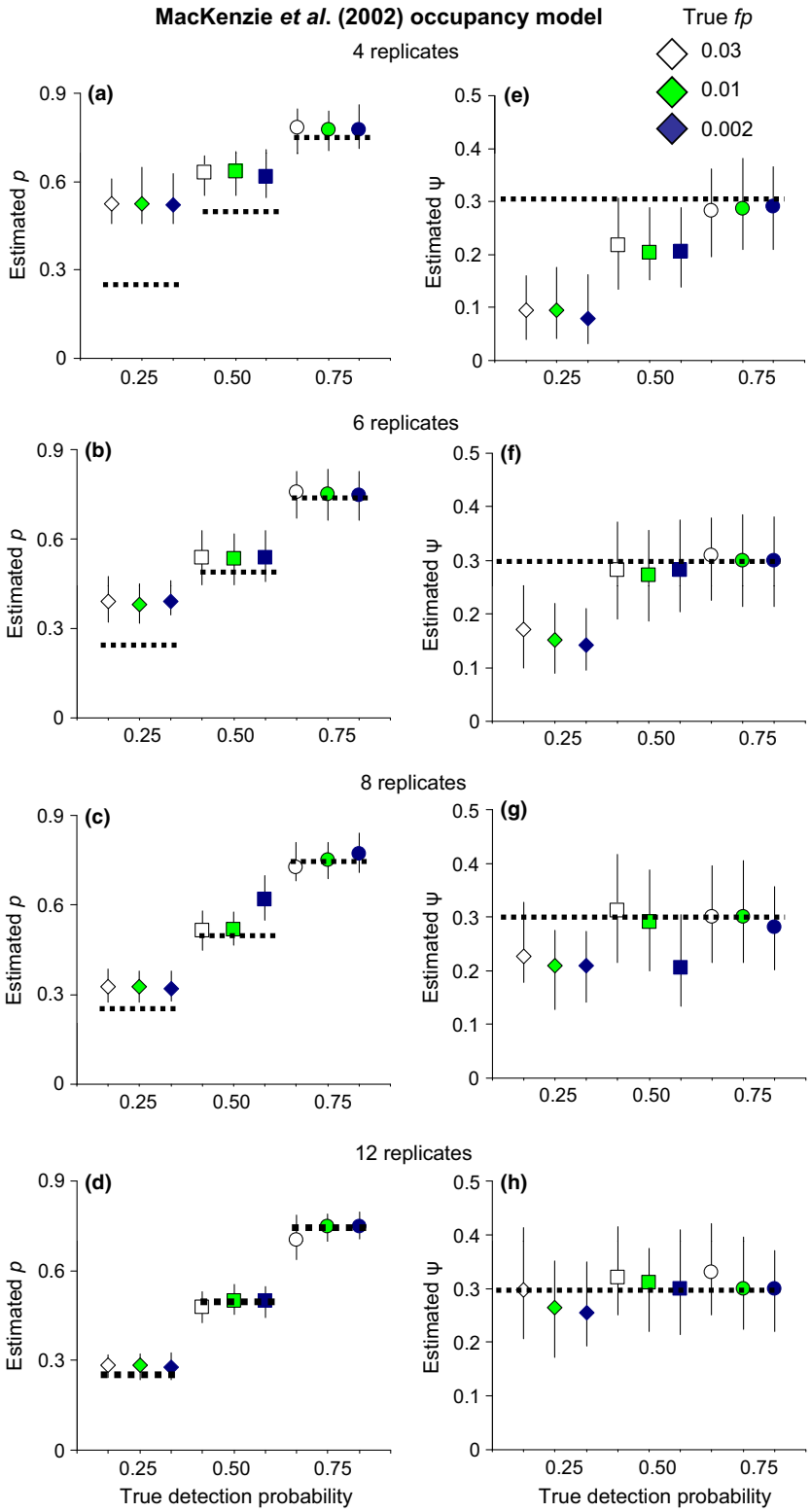


**Fig. 3** Analysis of simulated data using the MacKenzie *et al.* (2002) occupancy model. (a–d) Estimated detection probability ($p$); (e–h) estimated occupancy ($\psi$). Symbols represents medians ± 95% CI: *fp*, frequency of false presences. Bold dotted lines represent the true values of the parameters, used to generate the simulated data. Diamonds. detection probability = 0.25; squares, detection probability = 0.5; circles, detection probability = 0.75.

probability = 0.5, and 8–12 replicates were needed if detection probability was low (0.25, Fig. 3c–d). If the number of replicates was not large enough, the MacKenzie model tended to overestimate detection probability. Similarly, this approach correctly estimated true occupancy if detection probability was high, or if many replicates were performed. Four replicates were sufficient to estimate occupancy if detection probability was high

(0.75), while 12 replicates were needed if $p$ was very low (0.25). If the number of replicates was not large enough, the MacKenzie model underestimated occupancy (Fig. 3e–h).

*Miller's occupancy model.* Estimates of detection probability and occupancy using the Miller *et al.* (2011) model were highly consistent with those of the MacKenzie model. If detection probability was high, four replicates were sufficient to estimate occupancy and detection probability, while 12 replicates were needed if $p$ was very low (Fig. 4a–h). Furthermore, if a sufficient number of replicates (8–12) were run, the Miller's model estimated false presences with a good accuracy. Estimates of false presence were overestimated if detection probability was low, or if too few replicates were run (Fig. 4i–l).

### *Analysis of empirical data*

*Earthworm data.* Nine earthworm species were identified in the data set (Table 2). In all the samples and for all the species, three or more replicated PCRs confirmed species presence; therefore, no presences were considered uncertain, and data were thus analysed using the MacKenzie *et al.* (2002) model. For all the earthworm species, the estimated detection probability was high, ranging from 0.52 to >0.9, and observed frequency of species corresponded very well to their estimated occupancy (Table 2a).

Distinct soil samples from the same locality sometimes yielded slightly different communities. Specifically, ANOSIM detected significant differences between soil samples from the same site (at $P < 0.05$) in six of the twelve study sites. In the cases where ANOSIM detected differences between soil samples, one or two species were detected in only one of the two samples. Conversely, we never detected significant differences between extractions performed on the same soil sample (for all comparisons, $P \geq 0.33$).

*Ancient DNA from lake sediments.* Two mammal taxa (*Bos* and *Ovis*) were identified; sequences perfectly matched published sequences of domestic cow and sheep (Giguet-Covex *et al.* 2014). For the two taxa, in a few cases, we obtained a single positive amplification not confirmed by PCR replication (from three samples for *Bos* and four samples for *Ovis*). Consequently, we used the Miller *et al.* (2011) approach that takes into account uncertain presence. For both taxa, the estimated detection probability was moderate (0.36–0.38). Nevertheless, the observed and estimated occupancies were extremely similar, because the number of replicates was high (eight replicates per sample). The estimated rate of false presences remained limited (about 1.3%) (Table 2b).

## Discussion

Careful planning is a key phase for all ecological studies. The optimal level of replication is an important parameter that often results from complex trade-offs. Inadequate replication may yield inconclusive and inconsistent results, while performing many replicates is costly and time-consuming. DNA metabarcoding is an emerging area of research where the optimization of a replication strategy is particularly important. Species detection from small amounts of degraded DNA is clearly imperfect, still laboratory and sequencing costs limit replication. In recent years, researchers are increasingly combining analyses of simulated and empirical data to assess the effect of incomplete sampling and for the evaluation of the performance of analytical methods (Guillera-Arroita *et al.* 2010; Zurell *et al.* 2010; Ficetola *et al.* 2014). In this study, the analysis of simulated data allowed the identification of optimal strategies that limit the frequency of false presences and false absences in eDNA metabarcoding data, and we found that SOMs can help the analysis of eDNA data by providing accurate information on detection probability, false presences and true occupancy.

How many replicates should we perform? Increasing the number of replicated PCRs quickly reduces the risk of false negatives, improving the reliability of results. In practice, if only four replicates are run, the number of false absences can be very high, particularly if detection probability is low, and species occupancy may be severely underestimated (Fig. 1, Table 1). To limit the false absences, at least six replicates are needed when detection probability is about 0.5, and at least eight are needed if detection probability is lower. In many cases, metabarcoding studies target multiple organisms, with wide variation of detection probability. For instance, $p$ varied between 0.5 and >0.9 for the earthworm species analyses, while it was clearly lower for the mammals of the ancient DNA data set. If we want a reliable detection of the community, replication level should match the requirements of species with both low and high $p$. Eight replicates may represent an appropriate level of replication if no a priori information is available on the detection probability of species, as they may be suitable even for detecting species under difficult conditions, such as with ancient DNA. Nevertheless, it should be remarked that the optimal level of replication needed will also depend on the particular study and research objectives. For example, studies aiming at screening the possible biodiversity present in the samples (e.g. Bienert *et al.* 2012; Thomsen *et al.* 2012; Yoccoz *et al.* 2012; Parducci

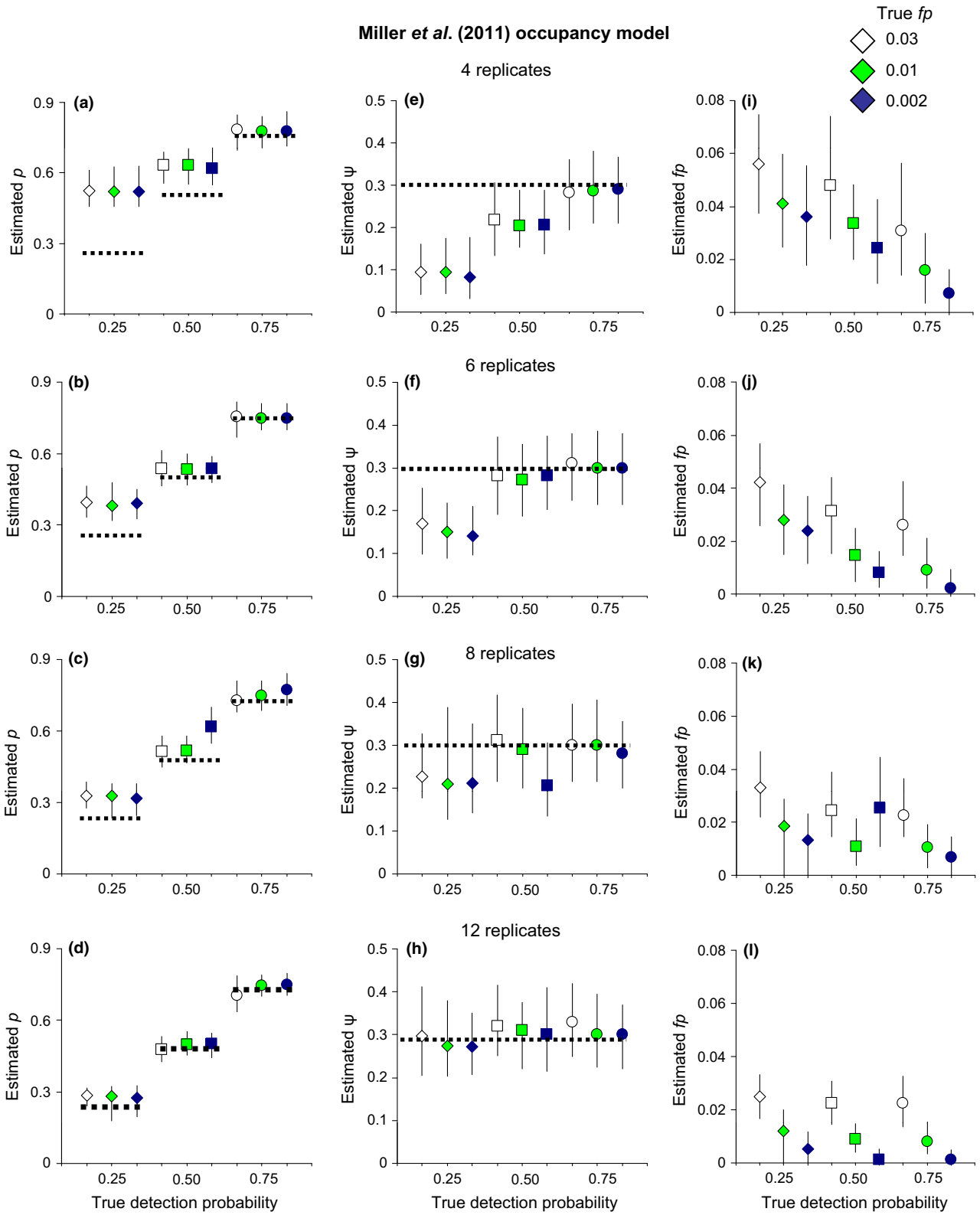


**Fig. 4** Analysis of simulated data using the Miller *et al.* (2011) occupancy model. (a–d) Estimated detection probability ($p$); (e–h) estimated occupancy ($\psi$). (i–l) Estimated rate of false presences ($fp$). Symbols represents medians ± 95% CI. Bold dotted lines represent the true values of the parameters, used to generate the simulated data. Diamonds, detection probability = 0.25; squares, detection probability = 0.5; circles, detection probability = 0.75.

**Table 2** Results of occupancy models applied to present-day soil DNA metabarcoding data (earthworms) and to ancient DNA of mammals from lake sediments

| Taxon | Observed frequency | $\psi$ | $p$ | $fp$ |
|---|---|---|---|---|
| (a) Earthworms | | | | |
| *Aporrectodea caliginosa* | 0.750 | 0.750 | 0.939 | — |
| *Aporrectodea icterica* | 0.167 | 0.167 | 0.867 | — |
| *Aporrectodea longa* | 0.583 | 0.583 | 0.849 | — |
| *Aporrectodea rosea* | 0.833 | 0.833 | 0.892 | — |
| *Dendrodrilus rubidus* | 0.250 | 0.274 | 0.519 | — |
| *Lumbricus castaneus* | 0.750 | 0.750 | 0.826 | — |
| *Lumbricus rubellus* | 0.333 | 0.334 | 0.565 | — |
| *Lumbricus terrestris* | 0.417 | 0.417 | 0.919 | — |
| *Octolasion cyaneum* | 0.917 | 0.917 | 0.963 | — |
| (b) Ancient DNA from lake sediments | | | | |
| *Bos* sp. | 0.273* | 0.283 | 0.377 | 0.013 |
| *Ovis* sp. | 0.113* | 0.118 | 0.363 | 0.014 |

$\Psi$, estimated occupancy; $p$, estimated detection probability; $fp$, estimated rate of false presences.
*Calculated using samples with two detections or more.

*et al.* 2013; Boessenkool *et al.* 2014), as the two real data sets analysed here, would differ from diet studies, where the main goal is to detect food items of biological importance for the animals (Pompanon *et al.* 2012; De Barba *et al.* 2014). In this latter case, nondetecting food items consumed only in traces or small quantities, and therefore present with low detection probabilities in some samples, would not affect the study conclusions, and a limited replication will most likely be sufficient. Sequencing depth, choice of primer and sequencing platforms are additional parameters that influence taxa detection and should be also considered during the planning of studies (Tang *et al.* 2012; Zinger *et al.* 2012; Deagle *et al.* 2013; Smith & Peay 2014).

Researchers working with eDNA metabarcoding are well aware that false positives are always a risk, because of contamination or because of errors during PCR and sequencing, and it is not unusual to detect sequences of taxa that are actually exotic to the study sites. Blank and positive controls are key tools to identify these taxa (Cooper & Poinar 2000; De Barba *et al.* 2014). An additional approach is to exclude as 'uncertain' the taxa that have been detected only once out the *Nr* replicates (Giguet-Covex *et al.* 2014; Willerslev *et al.* 2014). Our simulations showed that, with moderate levels of false positives, this method can successfully remove all the false presences (Fig. 2), but again it requires a sufficient number of replicates: if replicates are not enough, this approach may be too conservative and would remove taxa with low detection probability.

Species occupancy methods have excellent performance in the estimation of detection probability, true occupancy and even false presences, and can be successfully applied to eDNA data (Schmidt *et al.* 2013). SOMs can be a key resource for metabarcoding studies, but unfortunately they are only seldom used. In our simulations, two major approaches to SOMs (MacKenzie *et al.* 2002; Miller *et al.* 2011) yielded similar results in the estimation of occupancy and detection probability. Their performance was generally good and improved with the number of replicates and if target taxa showed high detection probability (Mackenzie & Royle 2005; Guillera-Arroita *et al.* 2010). If the number of replicates and detection probability were too low, these approaches tended to overestimate the detection probability and underestimated occupancy, while they required at least eight replicates for a robust inference over a wide range of $p$ values (Figs 3 and 4). Furthermore, SOMs can be successfully applied to eDNA data for the estimation of error rate (Miller *et al.* 2011), if taxa that have been detected only once out of the *Nr* replicates are considered 'uncertain'. This is excellent news for eDNA metabarcoding studies, in which dubious presences occur, as it may be possible to provide a measure of the reliability of such dubious presences. Importantly, if we know the error rate, we can eventually change the minimum number of positives required to consider taxon presence as 'genuine', and therefore identify taxon-specific or study-specific thresholds for the filtering of uncertain presences. For instance, at least three positive PCRs might be required to confirm the presence of species for which the risk of false detection is high.

Nevertheless, the approach described here has some limitations. First, PCR and sequencing errors may result in highly reproducible sequences. These errors may be found in multiple replicates and therefore incorrectly assumed to be genuine. We stress the importance of using the appropriate procedures of bioinformatic filtering, to limit the occurrence of these artefacts; the approach described here does not ensure removal of this kind of error. Second, this approach may be particularly suited for metabarcoding studies focusing on a restricted number of potential taxonomic units, or for which it is particularly relevant to evaluate whether a given taxon is present or absent in a sample. Examples of this application of metabarcoding include studies on bioindicators, or targeting taxa for which a relatively good proportion of species is well known. However, many metabarcoding studies try to describe the biodiversity of cryptic, poorly known taxa such as soil microorganisms. In this case, thousands of potential operational taxonomic units are potentially present, many of which have very low occurrence. In this case, other approaches should be developed, for instance, for the estimation of the richness of present taxa using accumulation curves (Lundberg *et al.* 2013).

In addition, occupancy modelling works well only if detection probability is reasonably high. Occupancy modelling is thus applicable to taxa with moderately low detection probability (about 0.25) only if at least eight replicates are performed. However, biodiversity is dominated by many extremely rare species that likely have very low detection probability. An unrealistically high number of replicates would be needed to apply occupancy modelling to these taxa. Inappropriately applying occupancy modally to taxa that are very rare and difficult to detect may lead to inaccurate conclusions (e.g. Fig. 3e).

DNA metabarcoding data are increasingly used for ecological inference (Ji *et al.* 2013). Objective measures of data reliability, such as rate of false presences and false absences, can also be used to improve the outcome of ecological analysis. For instance, strategies exist to integrate measures of detection probability within models relating taxa occurrence to environmental variables (MacKenzie *et al.* 2006; Fiske & Chandler 2011; Gómez-Rodríguez *et al.* 2012). If detection probability is not perfect, these approaches (e.g. mixture models) allow better ecological inference and help identifying the factors determining biodiversity (MacKenzie *et al.* 2006; Kéry *et al.* 2009). Furthermore, in our study, we simplistically assumed that detection probability and false presences are constant across the samples. However, these parameters may be influenced by environmental, biological (e.g. sample age, substrate, environmental temperature and differences among species) and technical factors (e.g. differences among operators and laboratories, preferential detection of shorter sequences, bias linked to the specific PCR and sequencing protocol). Occurrence of potential sources of biases may be tested and, if needed, integrated into models to limit their impact on the conclusion of studies. Finally, some studies are comparing the effectiveness of eDNA sampling with more traditional methods (e.g. amphibian calls, electroshocking and pit traps) (Dejean *et al.* 2012; Tréguier *et al.* 2014). These comparisons should take into account that both eDNA and traditional approaches are imperfect methods affected by false presences and false absences, and require the application of appropriate occupancy models.

Through this study, we broadly used the term 'replication levels' to indicate the total number of PCR replicates for each sample, but a given number of replicates may be reached through multiple ways (multiple samples per localities, multiple extractions per sample and multiple PCRs per extraction). Replication strategies are extremely different among studies: some researchers favour multiple PCRs on the same sample, while others favour multiple samples per site. Are these different approaches to replication equivalent? In the analysis of earthworm communities, we never found differences between distinct DNA extractions performed on the same environmental sample. This suggests that technical reproducibility is high when studying present-day samples (De Barba *et al.* 2014). Conversely, in some cases, two samples from the same locality yielded slightly different results, suggesting that microhabitat heterogeneity may be strong, and stressing that DNA metabarcoding results often represent very well local communities. Collecting multiple environmental samples per site may help limiting the effects of microhabitat/spatial heterogeneity, allowing more exhaustive results.

Replication level may have a strong impact on eDNA metabarcoding studies. Researchers should adjust replication level, depending on their aims and on the features of their study system. Before performing biological inference from eDNA metabarcoding data, we suggest (i) running occupancy models to evaluate the detection probability and rate of false presences; (ii) evaluating whether the current level of replication is appropriate to control for false negatives; (iii) if needed, removing 'uncertain presences', that is positives not confirmed by multiple PCRs. These three steps may allow improving the robustness of conclusions based on eDNA metabarcoding data.

## Acknowledgements

The comments of K. Deiner, S. Creer and two anonymous reviewers greatly improved an early version of the manuscript. We thank L. Chalmandrier for useful discussions on statistics. This project was funded by: FRB, Fondation pour la Recherche sur la Biodiversité, in collaboration with Irstea Grenoble (earthworm data set); Retromont program (Université Grenoble Alpes Savoie and CNRS in the framework of the DiPEE Chambéry-Grenoble) (ancient DNA data set) and by French National Research Agency (Metabar research program; ANR 11 BSV7 020 01). We specifically thank J.J. Brun and S. De Danieli to coordinate soil sampling; P. Choler, J. Poulenard and F. Arnaud for coordinating the Retromont program with P. Taberlet.

## References

Bienert F, De Danieli S, Miquel C *et al.* (2012) Tracking earthworm communities from soil DNA. *Molecular Ecology*, **21**, 2017–2030.

Boessenkool S, McGlynn G, Epp LS *et al.* (2014) Use of ancient sedimentary DNA as a novel conservation tool for high-altitude tropical biodiversity. *Conservation Biology*, **28**, 446–455.

Cooper A, Poinar HN (2000) Ancient DNA: do it right or not at ALL. *Science*, **289**, 1139–1139.

Darling JA (2014) Genetic studies of aquatic biological invasions: closing the gap between research and management. *Biological Invasions*, doi:10.1007/s10530-014-0726-x (in press).

Darling JA, Mahon AR (2011) From molecules to management: adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environmental Research*, **111**, 978–988.

De Barba M, Miquel C, Boyer F *et al.* (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources*, **14**, 306–323.

Deagle BE, Tollit DJ, Jarman SN *et al.* (2005) Molecular scatology as a tool to study diet: analysis of prey DNA in scats from captive Steller sea lions. *Molecular Ecology*, **14**, 1831–1842.

Deagle BE, Kirkwood R, Jarman SN (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, **18**, 2022–2038.

Deagle BE, Thomas AC, Shaffer AK, Trites AW, Jarman SN (2013) Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular Ecology Resources*, **13**, 620–633.

Dejean T, Valentini A, Miquel C *et al.* (2012) Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. *Journal of Applied Ecology*, **49**, 953–959.

Ficetola GF, Miaud C, Pompanon F, Taberlet P (2008) Species detection using environmental DNA from water samples. *Biology Letters*, **4**, 423–425.

Ficetola GF, Cagnetta M, Padoa-Schioppa E *et al.* (2014) Sampling bias inverts ecogeographical relationships in island reptiles. *Global Ecology and Biogeography*, **23**, 1303–1313.

Fiske I, Chandler R (2011) unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, **43**, 1–23.

Giguet-Covex C, Pansu J, Arnaud F *et al.* (2014) Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nature Communications*, **5**, 3211.

Gómez-Rodríguez C, Bustamante J, Díaz-Paniagua C, Guisan A (2012) Integrating detection probabilities in species distribution models of amphibians breeding in Mediterranean temporary ponds. *Diversity and Distributions*, **18**, 260–272.

Guillera-Arroita G, Ridout M-S, Morgan BJT (2010) Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution*, **1**, 131–139.

Hajibabaei M, Spall JL, Shokralla S, van Konynenburg S (2012) Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *Bmc Ecology*, **12**, 10.

Jerde CL, Mahon AR, Chadderton WL, Lodge DM (2011) "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters*, **4**, 150–157.

Jerde CL, Chadderton WL, Mahon AR *et al.* (2013) Detection of Asian carp DNA as part of a Great Lakes basin-wide surveillance program. *Canadian Journal of Fisheries and Aquatic Sciences*, **70**, 522–526.

Ji Y, Ashton L, Pedley SM *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Jorgensen T, Kjaer KH, Haile J *et al.* (2012) Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA Meta-barcoding. *Molecular Ecology*, **21**, 1980–1988.

Kéry M, Schaub M (2012) *Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective*. Academic Press, Waltham, Massachusetts.

Kery M, Schmidt BR (2008) Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, **9**, 207–216.

Kéry M, Dorazio RM, Soldaat L *et al.* (2009) Trend estimation in populations with imperfect detection. *Journal of Applied Ecology*, **46**, 1163–1172.

Lahoz-Monfort JJ, Guillera-Arroita G, Wintle BA (2014) Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, **23**, 504–515.

Legendre P, Legendre L (2012) *Numerical Ecology*, 3rd edn. Elsevier, Amsterdam.

Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL (2013) Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, **10**, 999–1002.

Mackenzie DI, Royle JA (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105–1114.

MacKenzie DI, Nichols JD, Lachman GB *et al.* (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

MacKenzie DI, Nichols JD, Royle JA *et al.* (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, Burlington, Massachusetts.

Mahon AR, Nathan LR, Jerde CL (2014) Meta-genomic surveillance of invasive species in the bait trade. *Conservation Genetics Resources*, **6**, 563–567.

Miller DA, Nichols JD, McClintock BT *et al.* (2011) Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology*, **92**, 1422–1428.

Nathan LR, Jerde CL, Budny ML, Mahon AR (2014) The use of environmental DNA in invasive species surveillance of the great lakes commercial bait trade. *Conservation Biology*, doi:10.1111/cobi.12381.

Oksanen J, Blanchet FG, Kindt R *et al.* (2013) *vegan: Community Ecology Package version 2.0-8*, www.r-project.org.

Parducci L, Matetovici I, Fontana SL *et al.* (2013) Molecular- and pollen-based vegetation analysis in lake sediments from central Scandinavia. *Molecular Ecology*, **22**, 3511–3524.

Pilliod DS, Goldberg CS, Arkle RS, Waits LP (2013) Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences*, **70**, 1123–1130.

Pompanon F, Deagle BE, Symondson WOC *et al.* (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, **21**, 1931–1950.

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Royle JA, Link WA (2006) Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, **87**, 835–841.

Schmidt BR, Kéry M, Ursenbacher S, Hyman OJ, Collins JP (2013) Site occupancy models in the analysis of environmental DNA presence/absence surveys: a case study of an emerging amphibian pathogen. *Methods in Ecology and Evolution*, **4**, 646–653.

Smith DP, Peay KG (2014) Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS One*, **9**, e90234.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012a) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

Taberlet P, Prud'homme SM, Campione E, et al. (2012b) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology*, **21**, 1816–1820.

Tang CQ, Leasi F, Obertegger U *et al.* (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences*, **109**, 16208–16212.

Thomsen PF, Kielgast JOS, Iversen LL *et al.* (2012) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, **21**, 2565–2573.

Tréguier A, Paillisson J-M, Dejean T *et al.* (2014) Environmental DNA surveillance for invertebrate species: advantages and technical limitations to detect invasive crayfish Procambarus clarkii in freshwater ponds. *Journal of Applied Ecology*, **51**, 871–879.

Valentini A, Miquel C, Nawaz N *et al.* (2009a) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trn*L approach. *Molecular Ecology Resources*, **9**, 51–60.

Valentini A, Pompanon F, Taberlet P (2009b) DNA barcoding for ecologists. *Trends in Ecology & Evolution*, **24**, 110–117.

Willerslev E, Davison J, Moora M *et al.* (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, **506**, 47–51.

Yoccoz NG, Nichols JD, Boulinier T (2001) Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution*, **16**, 446–453.

Yoccoz NG, Brathen KA, Gielly L *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, **21**, 3647–3655.

Zinger L, Gobet A, Pommier T (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, **21**, 1878–1896.

Zurell D, Berger U, Cabral JS *et al.* (2010) The virtual ecologist approach: simulating data and observers. *Oikos*, **119**, 622–635.

This work is the outcome of multiple discussions of researchers working on metabarcoding at the LECA. All authors contributed to research design through discussions. J.P., C.G.C., M.D.B., L.G. and P.T. performed sampling and generated the molecular data. J.P., E.C. and F.B. performed bioinformatics analyses. G.F.F. performed statistical analyses, developed simulations and wrote the first version of the manuscript, with the contribution of all co-author.

## Data Accessibility

The R code used for simulation is available as Supporting Information (Appendix S1). The complete description of methods used for the generation of the earthworm data set is available as Supporting Information (Appendix S2). Sequences of the earthworm data set are deposited in the DRYAD database under accession no. http://dx.doi.org/10.5061/dryad.k31d4. Tags and primers were removed from this data set during the data filtering process. Sequences of the mammal data set are deposited in the DRYAD database under accession no. http://doi.org/10.5061/dryad.h11h7.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Appendix S1** R code used for simulations.

**Appendix S2** Earthworm dataset protocol.