MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 604–607 doi: 10.1111/1755-0998.12508

NEWS AND VIEWS

PERSPECTIVE

How to limit false positives in environmental DNA and metabarcoding?

GENTILE FRANCESCO FICETOLA,*† PIERRE TABERLET*† and ERIC COISSAC*†
Universite Grenoble-Alpes, Laboratoire d'Ecologie Alpine (LECA), F-38000 Grenoble, France, †Centre National de la Recherche Scientifique, Laboratoire d'Ecologie Alpine (LECA), F-38000 Grenoble, France

Environmental DNA (eDNA) and metabarcoding are boosting our ability to acquire data on species distribution in a variety of ecosystems. Nevertheless, as most of sampling approaches, eDNA is not perfect. It can fail to detect species that are actually present, and even false positives are possible: a species may be apparently detected in areas where it is actually absent. Controlling false positives remains a main challenge for eDNA analyses: in this issue of Molecular Ecology Resources, Lahoz-Monfort *et al.* (2016) test the performance of multiple statistical modelling approaches to estimate the rate of detection and false positives from eDNA data. Here, we discuss the importance of controlling for false detection from early steps of eDNA analyses (laboratory, bioinformatics), to improve the quality of results and allow an efficient use of the site occupancy-detection modelling (SODM) framework for limiting false presences in eDNA analysis.

Keywords: bioinformatics, controls, eDNA, laboratory conditions, occupancy, sampling error

Received 24 December 2015; accepted 28 January 2016

When new approaches are at their infancy, there is a crucial need of methodological developments. The study of environmental eDNA, and the closely related metabarcoding, is quickly rising approaches for biodiversity analyses, and attract a growing interest. For instance, in 2013 just 2% of studies published by Molecular Ecology Resources focused on new developments relevant to metabarcoding or eDNA: the figure rose to 5% in 2014, to 10% in 2015, and will probably increase in 2016. New developments are flourishing in disparate directions, from the application of eDNA analyses to new substrates, to the optimization of laboratory toolkits and the development of new bioinformatics tools.

Nevertheless, some paths remain less explored. Field ecologists have always been aware that observations in nature are prone to detection mistakes: even if we survey a site multiple times, there are species that are present but may remain undetected. Even worse, we may go to the field, spot a bird and make an identification mistake. How can we account for such errors? Since the early 2000s, biostatistics has developed techniques allowing to take into account imperfect detection and, more recently, false detections (MacKenzie *et al.* 2006; Royle & Link 2006; Miller *et al.* 2011), but the first integration of these approaches with eDNA was only published in 2013 (Schmidt *et al.* 2013), and these tools remain underexploited in most eDNA studies. Until now, the application of SODM to environmental DNA has generally used standard approaches, developed for traditional ecological surveys. However, eDNA studies have specific features that may violate assumptions of some SODM analyses. The study by Lahoz-Monfort *et al.* (2016) is a new step to improve SODM analyses applied to eDNA, as it identifies the limitation of some SODM analyses performed until now, and discusses approaches that can provide more accurate estimation of detection probability, occupancy and (remarkably) false positives from eDNA data.

eDNA and metabarcoding give us the possibility of detecting taxa that remain unspotted using traditional approaches. Unfortunately, false positives may occur for multiple reasons, such as contamination during sampling or during laboratory work, PCR or sequencing errors (Ficetola *et al.* 2015). The 'cost' of a false positive may be particularly high. For instance, ancient eDNA can be used to reconstruct the history of agriculture (Giguet-Covex *et al.* 2014), but domestic animals and

Correspondence: Gentile Francesco Ficetola, Fax: +33 4 76 51 42 79; E-mail: francesco.ficetola@gmail.com

cultivated plants are often contaminating PCR reagents and kits (Champlot *et al.* 2010). False positives, leading to the conclusion that a given domestic species was exploited in an area centuries before its actual introduction, may undermine the overall confidence in eDNA to track environmental changes (Weiß *et al.* 2015). Similarly, if we use eDNA for the detection of invasive species, and falsely state its presence in a given area, resources may be wasted in the attempt of eradicating absent species.

In standard ecological analyses, the route from field surveys to a matrix of detection/nondetection is relatively straightforward, while more steps exist in eDNA research: a series of procedural and quality control measures must be adopted through all these steps to limit false-positive impact (Fig. 1). First, in the laboratory, it is essential to use an appropriate experimental setup, by following strict procedures to avoid contamination (Champlot *et al.* 2010), but also keeping controls and blanks at all the steps (extraction blanks, negative PCR controls, positive controls), which may later provide measures of the actual levels of contamination (De Barba *et al.* 2014). Furthermore, multiple analyses on the same sample are needed to obtain measures of reliability of results (e.g. Ficetola *et al.* 2015). Second, appropriate bioinformatics analyses are needed to translate the results of laboratory experiments into exploitable information on species distribution (Fig. 1c). Such phases are important for eDNA studies using approaches such as QPCR, but are even more critical for metabarcoding, as next-generation sequencing (NGS) results in millions of reads that must be appropriately treated to remove PCR and sequencing errors, chimaeras, sequences observed in just a few reads and so on. The application of all these steps ideally leads to matrices of species presence/absence (or, whenever possible, abundance; Evans *et al.* 2016). Nevertheless, false positives may still be present. Until now, ad hoc procedures have been proposed, such as not considering species detected in just one (and sometimes two) PCR replicates. However, the ad hoc removal is subjective, and may lead to severe underestimation of species occurrences. Lahoz-Monfort *et al.* (2016) demonstrate the performance of two approaches for the joint estimation of occupancy, detection probability and rate of false presences from eDNA data.

Sometimes, unambiguous data can be obtained in a subset of sites through different methods (e.g. direct observation). In this case, the combination of eDNA with such unambiguous data allows estimating the rate of false presences (Miller *et al.* 2011). This approach is powerful, and can for instance help when eDNA is used to search for a target species in water, as some direct observations are possible. Nevertheless, unambiguous detections cannot be obtained for some typologies of eDNA studies. For instance, eDNA metabarcoding is

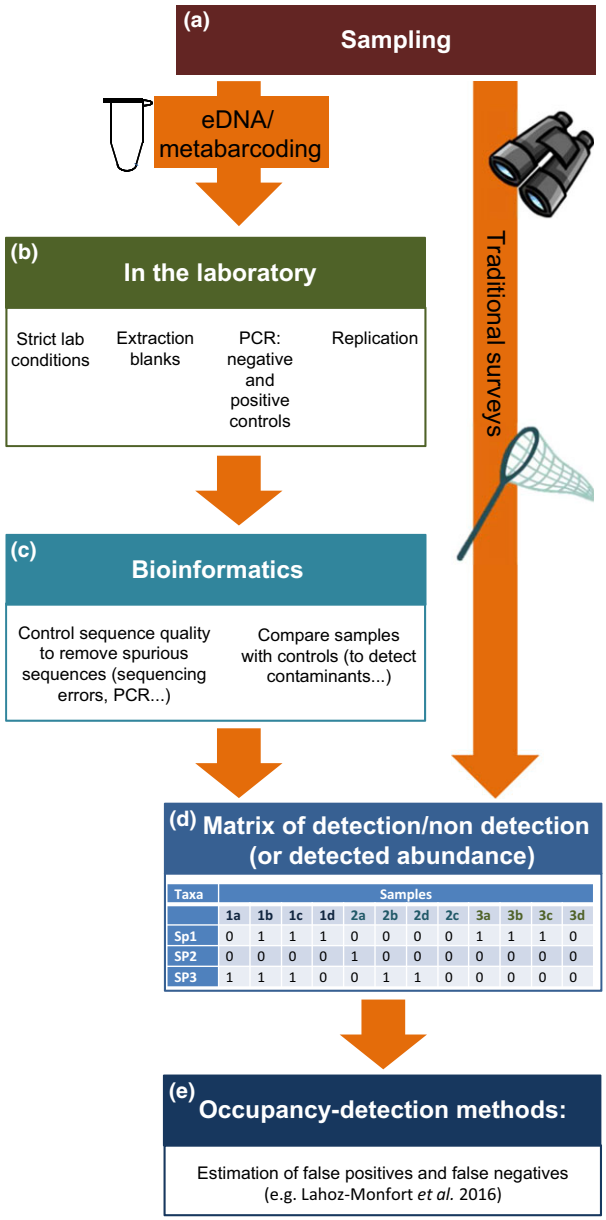

Fig. 1 Procedures to control for false positives at different steps of the environmental DNA (eDNA) and metabarcoding pipeline. From sampling to data, eDNA analysis requires more steps than traditional surveys: each step is a potential source of errors.

often used to analyse environmental samples for which traditional methods are challenging (e.g. microbiota, ancient environments, soil organisms), thus hampering the access to confirmatory data.

In such case, the generalized occupancy models developed by Royle & Link (2006) may be the best choice to successfully estimate false-negative and false-positive errors, because they do not require true detections data. Nevertheless, this approach is not without problems (e.g. may confound heterogeneity in detection and

false-positive errors; Royle & Link 2006). Furthermore, to apply these models we must assume that false detection rate is smaller than true detection rate: this precludes their application for species that are frequent contaminant. The only way to undoubtedly qualify (or exclude) a species as a contaminant is to set up a full set of controls and blanks at all the experimental steps. Furthermore, if laboratory practices are not sufficiently strict, the number of contaminants increases, thus multiplying the species that are not analysable. In practice, as the Royle & Link (2006) models are used at the last step of the eDNA pipeline, their efficiency is strongly linked to the application of an appropriate molecular procedures at the previous steps.

How can these approaches help answer our question? Reliable estimations of detection probability and rate of false detections can be used in multiple ways. First, when we use eDNA data for ecological inference (e.g. relationships between environmental features and biodiversity metrics), information on detection probability and false presences can be directly integrated into regressions or multivariate models (Royle & Dorazio 2008). Until now, this approach is rarely used, but would allow more accurate inferences. Furthermore, in many cases the question remains the same: is this fish species present in this lake? Even combining the most advanced tools, we still have uncertain cases, such as when we have sporadic detection of a target species just in one of many replicated samples from a given site (Jerde & Mahon 2015). If our estimations of false presences are correct, we can use them to calculate how many positives we need, to be confident that we do not falsely identify our fish as present at a site where it is absent (Box 1).

Approaches to control for false positives are quickly improving, still there are many areas where methodological advancements are needed. For instance, until now biostatistics has been applied to detection/nondetection matrices, obtained from sequences produced by NGS (Fig. 1d). How many NGS reads are needed to consider that a species have been detected by PCR? We know that species with just one (singletons) or two reads in one sample are probably artefacts, but if sequencing depth is high in some cases we can find tens of reads, assigned to absent species (De Barba *et al.* 2014; Elbrecht & Leese 2015). Providing rules-of-thumb is impossible, and perhaps appropriate analyses can help us to better transform NGS reads into community information. Second, eDNA data have a highly hierarchical nature (multiple samples per site, multiple PCRs per sample, etc.), but multiscale SODM taking into account false presences are yet to be developed (Lahoz-Monfort *et al.* 2016). Third, currently available SODM can use 'confirmed presences' as unambiguous data to better estimate false presences. In metabarcoding analyses, we sometime have 'confirmed false presences', such as the contaminants detected in the controls, which might perhaps inform SODM about the frequency of false presences. We hope that growing collaboration between ecologists, molecular biologists and biostatisticians will allow solving these and the many forthcoming challenges.

Box 1. How many positives do we need, to avoid stating that a species is present at a site where it is actually absent?

Let p_{10} be the false-positive rate and K the number of replicated analyses on one sample. Say that we consider that a taxon is present if detected in at least i PCR replicates from the same sample. The probability of falsely stating that a taxon is present where absent can be calculated using the binomial distribution as the probability of obtaining at least i positives with probability p_{10}, in K Bernoulli trials:

$$\Pr(x \geq i | \text{site not occupied}) = \sum_{j=i}^{k} \binom{k}{j} \times p_{10}^{j} \times (1-p_{10})^{k-j}$$

For instance, imagine that a fish species is absent from a given site, but it has a false-positive rate of 0.02. In this case, the probability of detecting false positives in 1 of 8 replicates is 0.149, the probability of detecting false positives in 2 of 8 replicates is 0.01, and the probability of detecting false positives in 3 of 8 replicates is 0.0004. In this case, we would need 3 independent positives to be highly confident on species occurrence.

Alternatively, instead of establishing presence/absence categorically, we can compute the probability that the species is present given the observed number of detections (x), and the estimated probabilities of occupancy (ψ), true detection (p_{11}) and false detection (p_{10}) (eq1 in Lahoz-Monfort *et al.* 2016):

$$\Pr(\text{site occupied} | x) = \frac{\psi p_{11}^{x} (1-p_{11})^{k-x}}{\psi p_{11}^{x} (1-p_{11})^{k-x} + (1-\psi) p_{10}^{x} (1-p_{10})^{k-x}}$$

References

Champlot S, Berthelot C, Pruvost M *et al.* (2010) An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS ONE*, **5**, e13042.

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.

Elbrecht V, Leese F (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass - sequence

relationships with an innovative metabarcoding protocol *PLoS ONE*, **10**, e0130324.

Evans NT, Olds BP, Renshaw MA *et al.* (2016) Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, **16**, 29–41.

Ficetola GF, Pansu J, Bonin A *et al.* (2015) Replication levels, false presences, and the estimation of presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, **15**, 543–556.

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.

Jerde CL, Mahon AR (2015) Improving confidence in environmental DNA species detection. *Molecular Ecology Resources*, **15**, 461–463.

Lahoz-Monfort JJ, Guillera-Arroita G, Tingley R (2016) Statistical approaches to account for false positive errors in environmental DNA samples. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12486.

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

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.

Royle JA, Dorazio RM (2008) *Hierarchical Modeling and Inference in Ecology: The Analysis of Data From Populations, Metapopulations and Communities*. Academic Press, London.

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.

Weiß CL, Dannemann M, Prüfer K, Burbano HA (2015) Contesting the presence of wheat in the British Isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data. *eLife*, **4**, e10005. 10.7554/eLife.10005.

G.F.F., P.T. and E.C. are actively involved in the use of environmental DNA and metabarcoding to understand biodiversity patterns, and in the development of new approaches along the whole eDNA pipeline. The authors developed together the ideas in this manuscript. G.F.F. wrote the first draft, followed by comments by all the authors.
