Received: 17 July 2019 | Revised: 11 February 2020 | Accepted: 12 February 2020

DOI: 10.1002/edn3.79

REVIEWS AND SYNTHESIS

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences

WILEY

Numerical methods for sedimentary-ancient-DNA-based study on past biodiversity and ecosystem functioning

Wentao Chen[1] | **Gentile Francesco Ficetola**[1,2]

[1]Laboratoire d'Écologie Alpine (LECA), UMR 5553, Université Grenoble Alpes & Centre National de la Recherche Scientifique, Grenoble, France

[2]Department of Environmental Science and Policy, Università degli Studi di Milano, Milano, France

Correspondence
Wentao Chen, Laboratoire d'Écologie Alpine (LECA), UMR 5553, Université Grenoble Alpes & Centre National de la Recherche Scientifique, Grenoble, France.
Email: chen.wentao.ecol@gmail.com

Funding information
H2020 European Research Council, Grant/Award Number: 772284

Abstract

Sedimentary ancient DNA (*Seda*DNA) is an emerging tool to reconstruct past biodiversity with high taxonomic resolution. Its growing popularity has stimulated an increasing complexity of *Seda*DNA data production (e.g., DNA extraction, amplification, and sequencing; authentication of molecules; bioinformatics). Conversely, less attention has been devoted to how appropriate statistical analyses can help to extract ecological information from *Seda*DNA. Until now, ecological studies based on *Seda*DNA have taken limited advantage of the multiple statistical and numerical methods available for analysis. Here, we present a range of numerical approaches that can be particularly useful to multispecies ecological analysis on *Seda*DNA, with a special focus on biodiversity studies on macroorganisms. We discuss the advantages and complexity of such methods and describe how some of them can be optimized for ecological analyses of *Seda*DNA-based metabarcoding data, with a special focus on *Seda*DNA studies. First, site occupancy-detection models can help to better ascertain the variation through time of the occurrence of target species and to identify the factors determining their detection through time. Second, several approaches can be used to estimate variation of relative abundance. Even though methods for abundance estimation have major limitations, they can provide useful information on temporal variation of ecosystem functions. Third, approaches exist to obtain better measures of species diversity, while taking into account the uncertainties of species abundance and identification. Fourth, techniques of clustering, ordination, and constrained ordination allow identification of temporal trends and testing of candidate drivers of community variation. Finally, structural equation models can be used to assess complex causal relationships among biodiversity, human activities, and environment. *Seda*DNA studies can make use of a broad panel of analytical approaches, which can improve our understanding of long-term biodiversity changes, maximizing the information we can obtain from past ecosystems.

KEYWORDS

biodiversity, DNA metabarcoding, ecosystem functioning, numerical methods, sedimentary ancient DNA, statistical analysis

1 | INTRODUCTION

The DNA of organisms living in the past can be successfully extracted and amplified from a variety of sediments, and this has greatly boosted our understanding of environmental changes, providing unprecedented reconstructions of past biodiversity at high taxonomic resolution (Bálint et al., 2018; Parducci, Bennett, Ficetola, Alsos, Suyama, Wood, Pedersen, et al., 2017). Thanks to the rapid technological advancement in high-throughput sequencing (HTS), the last decade has seen a growing number of studies based on sedimentary ancient DNA (*Seda*DNA), with increasingly complex approaches for data production and analysis. As paleoecological records, *Seda*DNA opens a new window on past biodiversity and offers invaluable opportunities to test ecological hypotheses over large temporal span and at multiple levels, ranging from population genetics (Parducci et al., 2012) to community ecology (Capo et al., 2016), to ecosystem functioning (Giguet-Covex et al., 2014).

Like other environmental DNA studies, *Seda*DNA studies follows a succession of six steps (Zinger et al., 2019): (a) sampling and conservation of the starting material; (b), extraction of DNA; (c) amplification or capture of target genomic regions; (d) high-throughput DNA sequencing; (e) bioinformatics analyses for sequence filtering and taxonomic assignation; and (f) application of appropriate statistical tools, often taken from numerical ecology, to understand patterns of biodiversity change and identify drivers and processes. It is increasingly evident that the quality of a study strongly depends on the application of appropriate approaches at each of these steps. The number of available techniques is quickly growing, with refined and dedicated tools particularly designed for the steps of molecular and bioinformatics analyses (e.g., Taberlet, Bonin, Zinger, & Coissac, 2018; Zinger et al., 2019). The final step (downstream analysis of data) is equally important, as it is essential for correct interpretation of data and ecological inference.

Biostatisticians are developing a rich and ever-growing arsenal of statistical techniques, which allow dealing with increasingly complex data analysis scenarios. Nevertheless, only a few of these techniques are routinely adopted for analyzing *Seda*DNA data. Failing to use the most appropriate tools may prevent harnessing the full potential of *Seda*DNA and/or lead to erroneous conclusions. For example, it is widely acknowledged that autocorrelation is a common feature of most spatial/temporal ecological data, and this must be taken into account for correct estimation of, for example, environmental drivers behind ecological patterns (Dormann, 2007; Ives & Zhu, 2011). Yet, the complexity of spatial/temporal autocorrelation is rarely considered in *Seda*DNA analyses (but see Chen & Ficetola, 2019; Dougherty et al., 2016; Ficetola et al., 2018). In the last years, a growing number of studies are showing the potential and benefits of appropriate numerical approaches to the analysis of environmental DNA. These studies have demonstrated how such combination can, for instance, provide better estimates of biodiversity parameters (Sommeria-Klein, Zinger, Taberlet, Coissac, & Chave, 2016), trace shifts of interspecific interactions (Zobel et al., 2018), and reveal the ecosystem-level consequences of biological invasion (Ficetola et al., 2018).

The recent blossoming of *Seda*DNA research has stimulated review papers summarizing ongoing trends and issues of this complex research topic (e.g., Bálint et al., 2018; Parducci, Bennett, Ficetola, Alsos, Suyama, Wood, & Pedersen, 2017). However, these reviews have mostly focused on the issues of *Seda*DNA data production (e.g., DNA extraction; amplification and sequencing; authentication of ancient molecules; bioinformatics analyses) and on the application of *Seda*DNA data within the ecological framework. The numerical aspects of ecological analyses based on *Seda*DNA have been rarely discussed, in spite of the needs and benefits discussed above. In this review, we focus on the ecological analysis of *Seda*DNA, which is performed after the bioinformatics steps (Parducci, Bennett, Ficetola, Alsos, Suyama, Wood, & Pedersen, 2017; Zinger et al., 2019). The aim of this work was to present a range of statistical approaches that can be particularly useful to multispecies ecological analysis on *Seda*DNA, especially for biodiversity studies on macroorganisms (e.g., animals, vascular plants), and to discuss how these methods can improve the analysis of *Seda*DNA. The approaches discussed are particularly relevant for *Seda*DNA studies (Figure 1), yet they can be also useful for other eDNA and metabarcoding approaches. We first describe the use of site occupancy-detection models to deal with imperfect detections (Section 2.1). We subsequently discuss how to map DNA data to taxon abundance (Section 2.2) and diversity (Section 2.3). We then focus on multivariate statistical methods to characterize multispecies assemblage changes, beginning with measuring the differences among objects (Section 3.1); then, we introduce some statistical methods for further ecological analyses: sample clustering (Section 3.2), detecting trends (Section 3.3), attributing change to drivers (Section 3.4), and testing complex causal networks (Section 3.5). In each section, we also present potential issues specific to *Seda*DNA data and potential solutions.

2 | FROM CALLING TAXON PRESENCE TO ESTIMATING SPECIES DIVERSITY

2.1 | Assessing the presence of taxa and the use of site occupancy-detection models

For many *Seda*DNA research projects, ecological analyses are based on a list of taxa and their presence/absence in each sample. Often, *Seda*DNA is used to ascertain whether a given taxon was present at some time (e.g., Giguet-Covex et al., 2014; Pedersen et al., 2016). However, species that were actually present can go undetected, so it is important to estimate the reliability of *Seda*DNA-derived presence/absence data. SedaDNA analyses usually start with a limited amount of DNA, and stochastic processes determine whether PCR amplifies a given DNA molecule; thus, it is essential to replicate analyses to validate patterns of species detection/ nondetection (Ficetola et al., 2015). In eDNA experiments with a multiple-replication setting, presence calling of a taxon [or more precisely, a

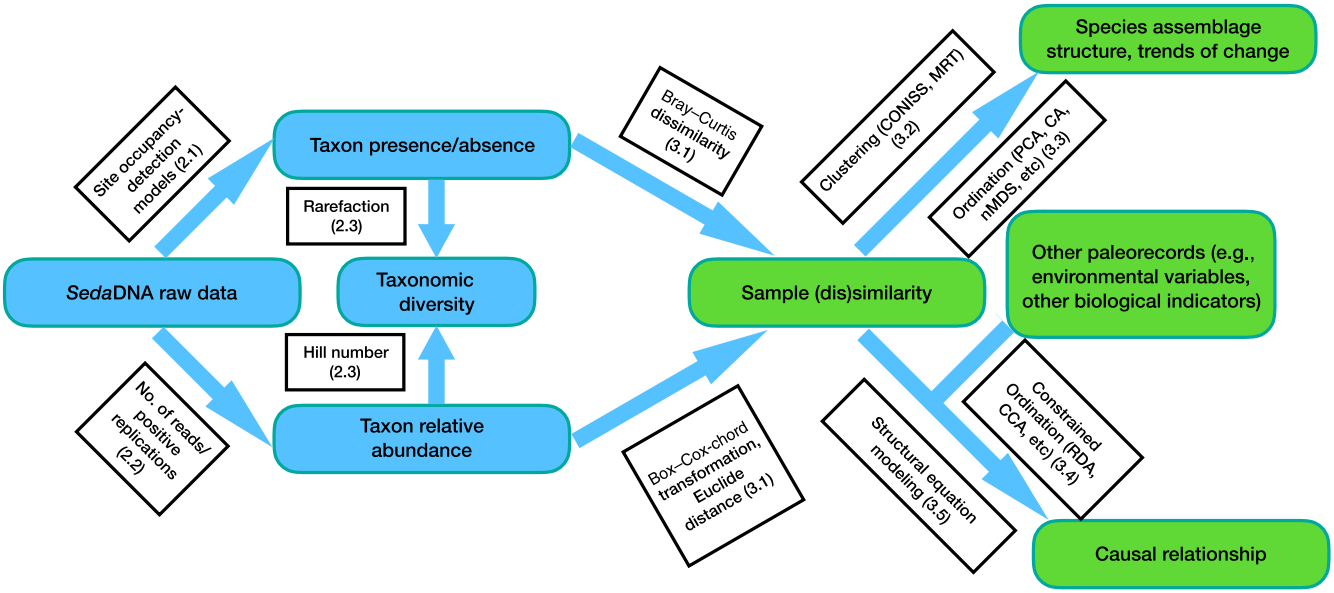


FIGURE 1 Ecological analysis to study past biodiversity based on sedimentary ancient DNA (*Seda*DNA). Rounded squares represent data forms that can be generated from *Seda*DNA data, with background colors indicating the related section (blue for those with Section 2 and green for Section 3). Arrows indicate data processing flow. Each arrow is accompanied by squares indicating possible approaches, with the related subsection numbers in parentheses at the end, to the respective data analysis step

molecular operational taxonomic unit (MOTU)] is often based on the number of positive replicates (Ficetola et al., 2015).

The reliability of species occurrence patterns can then be assessed in a probabilistic manner. The most often advocated approach is through site occupancy-detection models [SODMs, or site occupancy models, SOMs (Guillera-Arroita, 2017; MacKenzie et al., 2018)]. SODMs involve the modeling of one or several taxa's presence/absence (occupancy) in each sample (or site) as a probability φ. When present, a taxon is detected with probability p_{11} (i.e., the probability of detecting a true presence) and not detected with probability 1-p_{11} (i.e., the probability of obtaining false negatives). In *Seda*DNA, the situation is even more complex, because several processes can also lead to the detection of false positives, with probability p_{10}. False positives can arise from contamination, "tag jumps," or issues during the bioinformatics steps (e.g., inappropriate filtering) (Ficetola et al., 2015; Zinger et al., 2019). By fitting a statistical model that links these parameters to observed data (via maximum-likelihood or Bayesian approaches), one can estimate those parameters, as well as the probability of the actual presence/absence of each taxon in each sample (Ficetola et al., 2015; Kéry, 2010; Lahoz-Monfort, Guillera-Arroita, & Tingley, 2016; Schmidt, Kéry, Ursenbacher, Hyman, & Collins, 2013). Both false negatives and false positives can be determined by processes occurring during different steps of analyses (e.g., sampling, extraction, and PCR).

SODMs can be expanded with a multilevel structure taking into account both false positives and false negatives (i.e., modeling species occurrence and detection at different levels: site, sample, DNA extraction, PCR) (Davis, Williams, Snow, Pepin, & Piaggio, 2018; Guillera-Arroita, Lahoz-Monfort, van Rooyen, Weeks, & Tingley, 2017). However, since error rates occurring at the different stages are not identifiable without additional information beyond the observed—and imperfect—data, such multilevel SODMs require additional information from (a) an unambiguous survey method without false positives and/or (b) from a calibration experiment that provides direct estimations of the false positive/negative rates (Chambert, Miller, & Nichols, 2015; Guillera-Arroita, Lahoz-Monfort, Rooyen, Weeks, & Tingley, 2017). Unambiguous detections (approach a) are rarely available in paleoecological studies. On the other hand, calibration experiments (approach b) can be performed to estimate the false positive rate during each step. For example, the rate of false positive occurring during the DNA extraction process can be estimated by introducing extraction blanks and PCR controls (Ficetola et al., 2015; Pansu, Giguet-Covex, et al., 2015). The resulting data can then be used to inform prior distributions (e.g., through a binomial model setting) for the corresponding rates in a Bayesian parameter estimation approach. We note that it is not always clear how many control replicates are needed to provide a prior distribution that is sufficiently informative, for the purpose of resolving the identification issue [see (Griffin, Matechou, Buxton, Bormpoudakis, & Griffiths, 2019) for a discussion on using informative priors to address the identifiability issue). This may depend on the actual error rate; therefore, some "test drives" on simulated data may help prior to the actual experiments.

SODMs provide key ecological information on species distribution and on ecological processes. First, SODMs allow estimation of the probability that a site is occupied (occupancy, φ). Occupancy is a parameter of key ecological significance, which is used, to name a few, to trace the shift in a species' presence in a particular area through time (Schmidt et al., 2013) and to estimate the habitat patchiness of a species at multiple locations within a particular time span (Bailey et al., 2004). Second, the detection probability of target DNA generally increases with its concentration; thus, the frequency

of successful PCR amplifications increases with eDNA abundance. Therefore, the posterior probability of presence in each sample, a value that can be calculated from model parameters and observed data, may be used as a proxy of target DNA abundance (Ficetola, Manenti, & Taberlet, 2019; Ficetola, Taberlet, & Coissac, 2016; Furlan, Gleeson, Hardy, & Duncan, 2016; Lahoz-Monfort et al., 2016). Third, SODMs can incorporate covariates, which can be related to both the detection probability and the occupancy of taxa. This provides a means to take into account variation of detectability of species and might also help to identify the environmental factors that may affect the presence of species and/or the preservation of eDNA (e.g., Dougherty et al., 2016; Willoughby, Wijayawardena, Sundaram, Swihart, & DeWoody, 2016). Finally, additional random processes can be specified in SODMs to model ecological mechanisms. For instance, Olajos et al. (2018) used SODM to estimate the colonization date of whitefish in two Scandinavian lakes, based on lake sediment DNA. They modeled the colonization/extinction events of whitefish in each lake as random processes that eventually affected the presence/absence of the species and presented their results as posterior probability distributions of whitefish DNA presence.

Until now, SODMs of *Seda*DNA have often been based on techniques developed for other types of ecological studies. However, some aspects of *Seda*DNA may require the application of dedicated tools. In *Seda*DNA metabarcoding, as in all paleoecological records extracted from sediment cores, samples may be temporally autocorrelated. Furthermore, samples can be taken from multiple spatially interconnected sites, thus also leading to spatial autocorrelation. Autocorrelation can arise because intrinsic processes influence species distribution (e.g., dispersal between nearby patches; intrinsic autocorrelation) and because species distribution is related to environmental variables that are in turn autocorrelated (e.g., habitat, climate; extrinsic autocorrelation). The spatial and temporal dependence among *Seda*DNA samples must be addressed in order to obtain unbiased estimation of site occupancy, detection probability, rate of false positives, etc. On the other hand, population processes can be taken into account through extensions of SODMs explicitly estimating colonization and extinction processes (MacKenzie, Nichols, Hines, Knutson, & Franklin, 2003), as in the case of Olajos et al. (2018). Alternatively, Chen & Ficetola (2019) introduced a SODM setting with conditionally autoregressive model, or CAR model, adapted for eDNA metabarcoding studies, taking into account both temporal and spatial autocorrelations in a unified manner.

Another issue with occupancy inference is that the surface represented by the *Seda*DNA content may depend on the geomorphological settings of the study site, and might vary through time. The source area of SedaDNA is a key issue that very few studies have been dedicated to address. Studies in Scandinavian lakes showed that around 70% of contemporary sedimentary plant DNA records correspond to taxa found within 2 m from the lakeshore, whereas only a small part of the DNA records correspond to taxa only found far from the lake (Alsos et al., 2018; Edwards et al., 2018). Conversely, the analysis of Alpine lakes suggested that *Seda*DNA can represent the whole catchment from which the sediments originated, but erosive processes and structure of the hydrographic network may heavily affect the representation of *Seda*DNA within the sediments (Giguet-Covex et al., 2019). Insufficient knowledge of the precise area represented by *Seda*DNA can make data analysis and interpretation difficult, as multiple, spatially distinct communities can coexist in a catchment area. Therefore, it is important to carefully discuss the potential spatial coverage of DNA samples.

Several R packages and independent software are currently available to run SODM on eDNA data, such as unmarked (Fiske & Chandler, 2015) and EDNAOCCUPANCY (Dorazio & Erickson, 2018). EDNAOCCUPANCY was specifically designed for eDNA data analysis and provides tools to fit multiscale occupancy models with various settings, in a Bayesian framework. Furthermore, several authors provided scripts and examples to run Bayesian analyses in dedicated software such as WINBUGS (Schmidt et al., 2013), JAGS (Guillera-Arroita, Lahoz-Monfort, Rooyen, et al., 2017), or Stan (Chen & Ficetola, 2019). The great flexibility of these tools allows researchers to fine-tune their models in order to meet the specific aims, and to address the particular issues of their datasets.

2.2 | Measuring abundance

Measures of abundance can provide more complete information for ecological studies, compared to mere presence/ absence of taxa. However, the possibility to obtain reliable estimates of abundance through eDNA and metabarcoding remains highly debated. Some studies have found positive relationships between species abundance and eDNA concentration. For example, for eDNA researches targeting one single or a few species, a quantitative PCR approach can be used to quantify the DNA abundance of the species in question (e.g., Dougherty et al., 2016; Dunker et al., 2016; Ficetola et al., 2019), but see also Goldberg et al., 2016).

When real-time quantification is not an option, as in multispecies metabarcoding studies, an alternative approach to quantitatively analyze PCR results is to use the number of final sequence reads as an index of abundance. In fact, studies on modern eDNA have shown that the relative abundance of the eDNA of a taxon in a sample (measured as the proportion of reads) is often positively correlated to the relative abundance (e.g., biomass) of species within the community (Evans et al., 2016; Yoccoz et al., 2012); thus, some SedaDNA studies have used the proportion of reads as a measure of species abundance. Nevertheless, such relationship can be biased by multiple factors, such as gene copy number variation (Vasselon et al., 2018), different match with primers among species, differences in eDNA shedding, and multiple technical parameters (e.g., GC content, polymerase mixes, and number of PCR cycles); therefore, all the results must be interpreted with caution (Elbrecht & Leese, 2015; Fonseca, 2018; Nichols et al., 2018). Furthermore, the relationship between proportion of reads and abundance may vary across functional groups (Yoccoz et al., 2012); therefore, it is highly recommended to calibrate such relationship over the target groups

in each study before further analysis. In principle, the abundance of a given taxon can be compared across different samples if the stochastic nature of DNA sequence amplification by PCR is addressed by replication, and controlled through appropriate data transformation. Some authors used n^{th} (usually second or fourth) root transformation on the relative proportion of eDNA reads, which has been shown as an effective way to reduce the impact of unevenness in such data, and allowed the authors to characterize the vegetation changes in an arctic permafrost environment during Late Quaternary (Zimmermann et al., 2016).

A different approach to estimate abundance is to consider a multiple-replication experiment setting, using the number of positive replications as a measure of relative abundance, in addition to the number of reads (Pansu, Giguet-Covex, et al., 2015). The strength of this approach lies in the assumption that the probability of a sequence being amplified in a given PCR replication is positively correlated with its abundance (Furlan et al., 2016). Ficetola et al., (2018) used the number of PCR replicates obtained from lake core sediments as a proxy of rabbit abundance in the Kerguelen Island. This abundance measure was strongly correlated with the spore abundance of coprophilous fungi, which in turn is an indicator of the abundance of wild and domestic herbivores like rabbits (Richardson, 2001). Such correspondence between abundance measures originated from independent sources partially supports the idea that *Seda*DNA can provide some measure of abundance, at least in simple environmental settings.

Given the many uncertainties in the meaning of sequence abundance (Taberlet et al., 2018), any abundance measurement based on eDNA metabarcoding data should be interpreted with caution. Further studies are required to validate abundance obtained from *Seda*DNA with different approaches (e.g., macrofossils and historical records) and across a range of environmental conditions. Unfortunately, this can be particularly challenging, because *Seda*DNA is, in most cases, used to obtain information on past ecosystems, for which alternative measures of species abundance are rarely available. Furthermore, as discussed in Section 2.1, the spatial ranges that *Seda*DNA represents are not always clear, further complicating comparison of relative abundances through time.

2.3 | Quantifying species diversity

Species diversity is a major determinant of ecosystem stability and functioning (Loreau & de Mazancourt, 2013; Loreau et al., 2001) and is a key topic of ecological studies. Still using eDNA to estimate diversity is challenging. As previously discussed (see Section 2.2), uncertainties remain in the interpretation of sequence abundance and in the representativeness of the number of sequence reads. Several approaches can be adopted to mitigate the unfavorable consequences of those uncertainties. For species richness estimation from a limited number of *Seda*DNA samples, rarefaction analyses generate a curve relating the number of species to the number of samples through random resampling (e.g., Bellemain et al., 2013; Zimmermann et al., 2016). Recent statistical advances also provide opportunities to cope with uncertainties pertaining to eDNA metabarcoding data. For example, Hill numbers represent a parameterized diversity index that generalizes the three most important diversity indices—species richness (i.e., number of species), the Shannon index, and the Simpson index—by varying the value of a parameter *q*, which specifies the importance attributed to rare species (Chao, Chiu, & Jost, 2014; Hill, 1973). This can be very useful to limit the impact of artifact MOTUs, to weight common and rare species, and to systematically decompose and estimate diversity, in a systematic manner. It is even possible to calculate the Hill number from a species distance matrix (taxonomic, phylogenetic, or functional distance), with an extra parameter to adjust for the effect of closely related species, therefore allowing control of the impact of PCR and sequencing errors. All these features suggest that Hill number has a great potential to obtain fine-tuned biodiversity measures from eDNA data, despite that fact that this approach is not yet widely used in *Seda*DNA analysis.

3 | MULTIVARIATE METHODS TO CHARACTERIZE ASSEMBLAGE CHANGES

3.1 | Measuring sample (dis)similarity

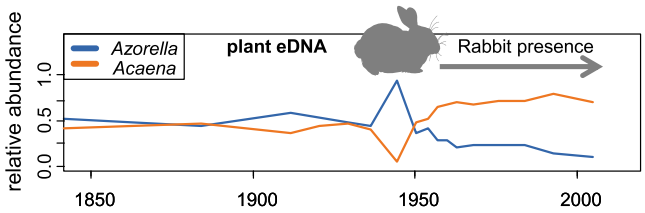
Measuring the similarity or dissimilarity between species (taxon) assemblages across different samples is a key step in many analyses of ecological communities. For instance, this is the basis of clustering samples, extracting trends of biodiversity change, and establishing causal relationships between environmental drivers and biodiversity. Some previous applications of *Seda*DNA used only presence/absence data in comparing samples (e.g., Pansu, Giguet-Covex, et al., 2015), as a conservative way to avoid the challenges of abundance estimation. However, comparisons based on relative abundance can often provide more insight to the biodiversity change than the mere presence–absences, because two species can remain present along the study period; still, their abundance can fluctuate in response to environmental stressors (Box 1). Therefore, researchers often use proxies of relative abundance of species, such as the proportion of read number of each taxon or the number of positive replicates.

In all cases, raw presence/absence or abundance data must be properly transformed before calculating (dis)similarities, in order to prevent the double-zero issue, wherein the simultaneous absence of a taxon in both samples is erroneously considered a sign of similarity between them (Legendre & Legendre, 2012). A variety of transformations and distance/(dis)similarity measures have been proposed to deal with different data types and purposes (Borcard et al., 2011; Legendre & Borcard, 2018). For presence/absence data, measures such as the Bray–Curtis dissimilarity (Bray & Curtis, 1957) can be used. For relative abundance data estimated from eDNA, a potentially favorable choice is the Euclidean distance computed on Box–Cox-chord-transformed data (Legendre & Borcard, 2018). This approach is a generalization of traditional data transformations

Box 1. When presence/absences are not enough

*Seda*DNA often just focuses on the detection of target taxa, given the complexity of obtaining reliable estimates of abundance. However, the ecological function of species is often correlated to abundance or biomass, therefore abundance estimate would ideally provide extremely useful information. A species can remain present for a very long period, even though environmental modifications can determine heavy abundance shifts. For instance, Ficetola et al. (2018) used *Seda*DNA to evaluate long-term trends of plants in the Kerguelen Island (Southern Indian Ocean), and how they are affected by the introduction of rabbits. The DNA of two plant species (*Azorella selago* and *Acaena magellanica*) was detected in all sediment samples from the period 1820-present. However, their relative abundance (measured as the proportion of reads) showed huge variation through time. The two species showed comparable abundance until 1945, when the *Seda*DNA of invasive rabbits was first detected. Rabbit detection corresponded with a spike of *Azorella* abundance (which in 1945 accounted for >90% of plant *Seda*DNA reads), and then quickly declined. In recent samples, *Azorella* accounts for just 10% of plant *Seda*DNA, and this matches well the rarity of this plant in the present landscape. Such strong abundance change has been interpreted as the effect of rabbit grazing and borrowing: rabbits directly consume *Azorella*, and heavy grazing and burrowing explain the *Azorella* peak around 1945, when rabbits first invaded the ecosystem. Then *Azorella* declined because of overgrazing, making this plant a threatened species (Chapuis, Boussès, & Barnaud, 1994). In this case, the mere presence/absence would not provide any information on changes of plant communities. However, it is also pivotal validating abundance estimates. In this example, both species showed no mismatches in the priming regions, suggesting that primers amplify well both species. Furthermore, analyses of present-day landscapes showed an excellent match between eDNA relative abundance and above-ground cover (Pansu, Winkworth, et al., 2015).

Figure I. Change of the relative abundance of *Seda*DNA of two native plant taxa, *Azorella selago* (blue) and *Acaena magellanica* (red), found in the Kerguelen Island through the last two centuries. Note the shift of the relative abundances between the two taxa after the rabbit invasion occurring around 1945.



designed to simultaneously solve the double-zero problem and improve the multinormality of data [or multivariate normality, which is assumed by some multivariate extensions of linear models such as multivariate analysis of variance and discriminant analysis (Williams, 1983), although many analyses use permutation tests to avoid parametric assumption (Legendre & Legendre, 2012)]. With a parameter to adjust for different degree of data skewness (Legendre & Borcard, 2018), this approach provides a flexible way to compare community compositions recorded in eDNA and avoids the double-zero issue.

3.2 | Clustering samples to find hidden structure

In many biodiversity studies, a key objective is to find hidden structure of species assemblages across multiple sites or samples, by comparing them along some spatial/temporal or environmental gradients. Such comparisons can be performed using two major types of techniques: clustering and ordination (Legendre & Legendre, 2012).

Cluster analysis is the partitioning of a collection of objects (e.g., species assemblages) under study, with the aim to find discontinuities in a dataset. Generally speaking, the main purpose of cluster analysis is to bring out unseen structures in data, although the ecological relevance of the result clusters is subject to the user's interpretation (Borcard et al., 2011). There is a plethora of clustering methods, each of which adapts to some specific data types and meets specific needs. In *Seda*DNA metabarcoding studies, the stratigraphically ordered nature of samples usually poses a temporal constraint to the clustering. A solution widely used in pollen analysis is the Constrained Incremental Sums of Squares (CONISS) algorithm (Grimm, 1987), while it has seldom been applied to *Seda*DNA data (e.g., Zimmermann et al., 2016). This algorithm works by iteratively merging the two stratigraphically neighboring clusters that give the least increase in within-cluster variance (compared to the sum of the dispersions of the two original clusters), which is calculated as sum of squares of a chosen dissimilarity coefficient (e.g., a distance measure). The result of a CONISS analysis is a hierarchical structure of sample clusters, which can be represented as a tree. One of the advantages of CONISS is that the clusters always comprise samples representing neighbor periods, thus maximizing interpretability of results. Like most clustering methods, CONISS does not provide a test to determine the optimal number of clusters (or zones). This task is usually done by comparing the variance reduction by each

level of clustering to those expected from a randomly splitting with the same number of zones [e.g., those resulted from a broken-stick model (Bennett, 1996)].

A more recent technique, namely the multivariate regression tree (MRT) analysis (De'ath 2002), allows clustering samples under the control of one or multiple, quantitative or categorical, explanatory variables. Like the output of CONISS, the resulting MRT is also a hierarchical tree, of which the "leaves" are subsets of samples chosen to minimize the within-group variance, but with each consecutive partitioning defined by a threshold value (for the case of a quantitative explanatory variable) or a binary state (for the case of a categorical variable) (Borcard et al., 2011). A cross-validation process can be applied to determine to the most reasonable partition size, that is, to decide at which level to retain the clustering. The process is a prediction-oriented method that randomly splits the set of samples into a training set to construct a MRT, and a smaller testing set to validate the predictive power of the constructed MRT. The optimal partition size is defined as the level that minimizes the predictive error. Alternatively, a smaller partition size can be retained, if its predictive errors are one standard error above the minimal predictive error, which is called the one-standard-error rule (Breiman, Friedman, Olshen, & Stone, 1984). The MRT analysis is an efficient tool to partition stratigraphic samples and has been successfully applied to *Seda*DNA data (Ficetola et al., 2018). Although the MRT analysis has been originally developed as a prediction-oriented machine-learning technique (Borcard et al., 2011), it can also serve as explanatory tool to investigate biodiversity–environment relationships.

3.3 | Looking for main trends by ordination

In contrast to cluster analysis, which identifies discontinuities in a dataset, ordination analysis arranges object points along one or more continuous axes, each of which representing an ordered relationship (Legendre & Legendre, 2012). Ordination methods are a group of multivariate techniques that, in ecology, can be applied to a species-site matrix or on a distance/dissimilarity matrix. These methods arrange objects (i.e., sites or samples) along one or more axes based on their compositions and then display them in a coordinate frame so that the higher-dimensional relationships among them are easy to inspect (Legendre & Legendre, 2012; Pielou 1984). Therefore, ordination methods provide a convenient way to visualize multidimensional biodiversity datasets such as a species abundance-site matrices. As for clustering analysis, plenty of ordination methods are available, adapted for different data types and for different assumptions on the data. For instance, principal component analysis (PCA) works on species-site matrices, with the aim to find a series of mutually orthogonal linear combinations of species (i.e., axes) that successively maximize the variance of scatter points, while preserving the Euclidean distance among sites. Therefore, PCA brings out linear relationships among variables. On the other hand, correspondence analysis (CA) works on non-negative, frequency-like matrices and aims to find successive axes that maximize the similarity (measured as "correspondence"), but instead preserves the chi-square distance. Accordingly, CA can be a good choice when a unimodal response of species to the environmental gradient is expected (Legendre & Legendre, 2012; ter Braak, 1985). Both PCA and CA, together with other eigenvalue-based methods such as principal coordinates analysis (PCoA), seek to represent a matrix in a series of axes ordered by their "importance," at the same time preserving the distances of a particular type (e.g., Euclidean in the case of PCA) among objects.

Researchers need to choose among these methods according to the data type and the expected relationships among variables. The choice is, however, not necessarily restricted by the nature of the raw data, because a large number of transformation techniques enable adjustments of raw data to meet the requirement of particular ordination methods; Legendre & Gallagher (2001) discussed the data transformation techniques most relevant to ecological analysis based on species data. There are nevertheless some specific considerations for the case of *Seda*DNA. First, *Seda*DNA detection can be difficult for rare species, and thus, their estimation may be biased (Alsos et al., 2018); therefore, it may not be appropriate to use methods that emphasize the role of rare species in the ordination plot, such as CA (Legendre & Legendre, 2012; but see Greenacre, 2013). Second, appropriate transformations are needed to mitigate the effect of overdispersion among taxa, especially if it is applied to complex measures of relative abundance such as PCR-based read number data [e.g., 4th root transformation (Zimmermann et al., 2016); the Box–Cox-chord family of transformation, i.e., the combination of a Box–Cox transformation with exponent within [0, 1], and the chord transformation (Legendre & Borcard 2018)]. When presence/absence data are preferred over read numbers, Hellinger distance or chord distance transformation before ordination analysis is often recommended (Legendre & Borcard, 2018; Legendre & Gallagher, 2001; see, e.g., Epp et al., 2015).

In contrast, nonmetric multidimensional scaling (nMDS) works by an optimization process that tries to place objects in a space of a given number of dimensions, in a way that minimizes the total squared differences between the between-object distances in this space and those in the original space. In other words, nMDS does not preserve any particular between-object distance and finds instead a configuration of points (objects) in a space of a lower number of dimensions with least changes possible in the distance relationships among points (Kruskal, 1964). Therefore, nMDS works on resemblance (including distance, dissimilarity, or similarity) matrices and is not optimized for any particular relationship among variables, in contrast to PCA and CA. In all, nMDS is suitable for analyses that aim to reconstruct community composition as well as possible, especially when one does not desire to preserve any particular distance measure among objects (Borcard et al., 2011).

Given its flexibility and robustness, nMDS is often used in *Seda*DNA metabarcoding studies as the first step to characterize the similarity of communities (e.g., Ficetola et al., 2018; Pansu, De Danieli, et al., 2015; Pansu, Winkworth, et al., 2015). Most importantly, since the relationship between DNA data (represented by number of reads or of number of positive replicates) and taxon

abundance remains unclear, there is to date no formal evaluation on how it will interfere with linear (by PCA) or unimodal (by CA) responses. nMDS is thus favored over parametric methods such as PCA. Despite its advantages, a successful nMDS analysis should be conducted with care.

First, an appropriate similarity or dissimilarity coefficient is needed to transform the raw matrix, and the choice of such coefficient is dependent on the data and the scientific context. In some metabarcoding applications, the Bray–Curtis dissimilarity is preferred, since it avoids the double-zero problem, and can deal with both presence/absence and abundance data (e.g., Pansu, De Danieli, et al., 2015; Pansu, Giguet-Covex, et al., 2015). Some other options are available. For instance, Legendre & Borcard, 2018 suggested the abovementioned Box–Cox-chord family of transformations to deal with the double-zero problem. The Renkonen (dis)similarity is an alternative option when dealing with relative abundances, because of its density invariance [i.e., not being sensitive to raw abundances, but only to relative ones (Jost et al., 2011)]. This feature may be favorable for comparing taxonomic compositions of sediment samples with different amounts of DNA, as different degrees of DNA degradation that create such disparity are common in sediment samples. Second, there is no simple rule to determine in advance how many dimensions need to be retained in nMDS. Usually, two to three axes are enough, but it is recommended to check the stress (the penalty function minimized by the nMDS algorithm) in several alternatives in order to help deciding. Third, although its goodness of fit can be assessed by the abovementioned stress, nMDS provides no intuitive measure of the representativeness of its axes upon the original dataset, in contrast with PCA, which comes with the amount of variance explained by each of its axes (indicated by the respective eigenvalues). Therefore, the interpretation of the axes of nMDS requires different strategies. For example, environmental variables can be fitted on the nMDS by linear regression. The passive explanation of axes using environmental variables is a post hoc approach that can be applied to both nMDS and PCA, and allows the interpretation of an ordination using external variables. This procedure also allows testing the significance of environmental variables fitted on the ordination plots, as well as measuring the amount of variation they explain (see Borcard et al., 2011 for details). Finally, in most of cases the axes of an initial nMDS result do not lie in parallel with any environmental gradient. If needed, one can rotate the axes so that the first axis is parallel to an external variable without changing the configuration of points [e.g., using the MDSrotate function in vegan (Oksanen et al., 2017)]. This provides an effective way to compare the order of sites (samples) and the external variable.

Finally, due to the noisy nature of eDNA (Taberlet et al., 2018), the results of ordination might be overwhelmed by random dispersions. It is therefore crucial to effectively reduce the noise and keep the signals at the same time, by taking appropriate data transformation and dissimilarity measure. In the special case of *Seda*DNA, an even greater challenge is to deal with the unevenness of data quality along time, usually caused by DNA degradation (Parducci, Bennett, Ficetola, Alsos, Suyama, Wood, & Pedersen, 2017; Pedersen et al., 2015). A potential solution could be to statistically model the degradation process and compensate (e.g., weight the data along a degradation gradient) its effects on the final dataset. Unknown taphonomy of *Seda*DNA may also confound the interpretation of results (Bálint et al., 2018; Giguet-Covex et al., 2019); therefore, inspecting the sedimentological and geochemical data of the sediment core(s) is a necessary (but not always sufficient) step to control for such effects.

3.4 | Testing the effects of environmental variables on community composition

The abovementioned unconstrained ordination methods do not provide statistical tests of factors related to community variation; they only serve to represent the major features of the data within a reduced number of dimensions. To statistically test specific hypotheses about biodiversity–environment relationships, other approaches can be used. Constrained ordination is a popular choice to assess the impact of environmental factors on community composition. Constrained ordinations are a class of statistical methods that generally combines the analysis of eigenvalues and eigenvectors (as in PCA, PCoA, CA, etc.) and regression (Legendre & Legendre, 2012). They produce explanatory vectors that are related in a certain manner to the matrix of explanatory factors. Such manner is the defining feature of different types of analysis. In the case of redundancy analysis (RDA), a constrained ordination technique often used in eDNA data analysis, the ordination vectors are the PCA axes of the fitted values from a linear regression of environmental variables on (species) response data (Borcard et al., 2011). RDA can therefore be considered as an extension of PCA. On the other hand, in constrained correspondence analysis [CCA, (ter Braak, 1986)], fitted values from the linear regression are further ordinated by CA. Again, the expected relationship between explanatory variables and community composition is essential in choosing between RDA and CCA, and appropriate data transformations are needed before performing either one on *Seda*DNA data (see Section 3.3). Usually, the significance of the ordination vectors resulting from a constrained ordination cannot be tested with classical parametric tests, because there is in general no reference distribution for the obtained statistics ("pseudo-*F*") (Legendre & Legendre, 2012). Therefore, permutation tests are needed, which randomly permute certain elements of the data to generate a distribution of the chosen statistic under the null hypothesis. The RDA/CCA approach has provided valuable insights on the environment–community relationships in the past through *Seda*DNA. For example, using RDA on *Seda*DNA data, Pansu, Giguet-Covex, et al., 2015 showed that plant community changes, occurring in an alpine lake area through the past six thousand years, were mainly driven by livestock farming and not by temperature changes.

Ecological phenomena often involve multiple tangled links. It is therefore crucial to resolve their confounding effects in analysis. In linear regression, it is possible to estimate the effects of some explanatory variables, while controlling for the effects of some

covariates. Similar approaches exist in constrained ordination, for example, partial RDA and partial CCA (Legendre & Legendre, 2012; ter Braak, 1988). Especially, *Seda*DNA data are often temporally structured, so that some parts of the variation may have originated from intrinsic temporal processes. Such processes may generate effects that confound with those caused by extrinsic environmental factors, that is, resulting from niche-based processes (Legendre & Legendre, 2012). Using partial constrained ordinations, it is possible to partition the variance of the composition dataset into four categories: (a) environmental variance independent of temporal structure; (b) variance shared between temporal variation and environmental drivers; (c) variance explained by the temporal structure only; and (d) variance explained neither by the temporal structure nor by the environmental factors (Borcard et al., 1992). Different approaches have been proposed to take into account the temporal (or spatial) autocorrelation structure. (Borcard et al., 1992) suggested to include the coordinates of samples as additional predictors, but this approach has drawbacks since it does not directly address the problem of autocorrelation, but merely takes into account trends in the data across the largest distances (Dormann et al., 2007).

Furthermore, ecological processes often have a scale-dependent structure that cannot be captured by a simple analysis on the coordinates (Legendre & Legendre, 2012). In the case of paleoecology, it is widely acknowledged that even the same environmental factor can have different effects on ecosystem on different time scale (Mills et al., 2016). Recent developments in spatial/temporal analyses provide a series of methods to assess the multiscale spatial/temporal patterns in multivariate data. Moran's eigenvector maps (MEMs) (Legendre & Gauthier, 2014) are family of models that can be used as a tool to assess nondirectional patterns. MEMs work by extracting eigenvectors that maximize autocorrelation, out of the spatial/temporal weighting matrix that summarizes the spatial/temporal relationships between samples. On the other hand, directional patterns across space/time can also be assessed; with asymmetric eigenvector maps (AEMs) modeling, the directional counterpart of MEMs (Blanchet et al., 2008). In practice, the resulting eigenvectors from both MEMs and AEMs can be tested as explanatory variables in a constrained ordination analysis, in which the most relevant eigenvectors can be selected (Legendre & Gauthier, 2014). Selecting the appropriate number of eigenvectors to include into analyses can be nevertheless challenging. Several approaches have been proposed to the selection of appropriate eigenvectors, and the outcome can be strongly different among approaches. Importantly, adding many eigenvectors can inflate the amount of variance explained by temporal (or spatial) structure, while selecting eigenvectors correlated to the independent variables can lead to overfitting, and in turn reduce the power to detect the effects of other environmental variables (Bauman et al., 2018). Therefore, special care should be devoted to the selection of the eigenvectors to be incorporated into constrained ordination. Suggested approaches include forward selection with a double-stopping criterion after testing the significance of the global model to prevent overfitting (the FWD approach), and selecting the eigenvectors minimizing the autocorrelation in the model residuals (the MIR approach). The FWD approach is preferred when an accurate description of the spatial/temporal patterns is the main purpose, whereas the MIR approach is more appropriate when one wants to control for autocorrelation (Bauman et al., 2018). Unfortunately, to our knowledge, these methods are not yet adopted in *Seda*DNA-based studies.

Species–environment relationships may hold true in a certain timescale but not so in others, or their strength may vary accordingly (Mills et al., 2016). An effective way to highlight the essential scales of spatial/temporal variation is to look into the ecological variability through different "scale filters" (Jombart, Dray, & Dufour, 2009). This can be done by multiscale pattern analysis (MSPA), which applies RDA to the MEMs or AEMs [as in Legendre & Gauthier (2014)] and then performs a variation of PCA (called %PCA) on the matrix of resulting centered R^2 (Jombart et al., 2009). Analyzing such scale-dependent relationships with *Seda*DNA data however demands finer, and eventually more regular, temporal sampling intervals than simple constrained ordinations, in order to detect as much scale-dependent effects as possible. Furthermore, inappropriate use of these multivariate-partitioning methods to model spatial/temporal or environmental processes can produce statistical artifact, such as inflated R^2 statistics with raw-data-based ordinations, and under-fitting with distance-based regressions (Gilbert & Bennett, 2010). Therefore, they must be applied with caution.

The methods presented so far in this section essentially are extensions of multivariate analysis of variance (MANOVA) (Borcard et al., 2011), which partitions multivariate variance into components that are attributable to different explanatory variables, through the comparison of multivariate sample means. This approach by itself is not able to account for all potential response modes of ecosystem to environmental variables. For example, the variability of community composition can be altered as a result of change in disturbance regime (Collins, 2000). Insuch cases, environmental changes can lead to changes in the variance of some properties of the system, in addition to the average properties themselves. Tests for homogeneity of variance such as Levene's test (Levene, 1961) can be applied to evaluate such effects for univariate response data, given that the explanatory variables divide the samples into several distinct groups. When the focus is on multivariate responses (e.g., community composition), multivariate analogues of these tests can be used, such as the test for homogeneity of multivariate dispersions (Anderson, 2006). This can be done by firstly performing principle coordinate analysis [or PCoA, an ordination technique to illustrate similarities or dissimilarities among objects with an Euclidean representation (Borcard et al., 2011)] on a distance matrix. Subsequently, the centroid or spatial median of each object group in the representation is calculated, and the method calculates the distance from each object to its group centroid or spatial median. Finally, the distances are compared across groups with a multivariate version of Levene's test (Anderson, 2006). Although promising, this approach has not been frequently applied to *Seda*DNA data. Importantly, *Seda*DNA data often are noisy, and the degree of "noisiness" may depend

on multiple factors such as frequency of geomorphological events and DNA degradation, leading to confounded inference of variance change (Giguet-Covex et al., 2019). One must therefore carefully check the data to control any confounding effects before comparing.

Here, we have presented some of the most promising applications of multivariate statistical methods to *Seda*DNA data. Paliy & Shankar 2016 provided a further guidance on the application of abovementioned techniques (PCA, CA, nMDS, CCA, RDA, etc.), as well as some others not discussed here, to the ecological analysis of data obtained through metabarcoding, with a chart to guide the choice of methods (see figure 8 in Paliy & Shankar 2016). Although their main focus was on the applications for microbial ecology, these discussions are broadly applicable to data from high-throughput sequencing and can also be relevant for *Seda*DNA data.

3.5 | Testing complex driver–response relationships

Ecosystems function in complex ways, involving multiple biotic/abiotic factors simultaneously at work, which are woven into a network of causality. Ecological explanations are therefore often both inherently multivariate and causal in nature (Shipley, 2016). Long-term ecological records such as *Seda*DNA, combined with other proxies recording environmental conditions, provide excellent opportunities to investigate such complex causal network. Instead of testing particular causal relationships one-by-one, it is possible to investigate a complex causal network as a whole through a combination of both graphical and mathematical models, that is, structural equation modeling [*SEM*, (Grace, 2006; Shipley, 2016)]. Treating causal network as a whole brings benefits to the interpretation of data. Most notably, it enables assessment of directed relationships between variables and not merely correlations. This possibility comes from the fact that the ambiguous causality direction [either A causes B, or B causes A, or both A and B are caused by an unobserved C? (Shipley, 2016)] in a correlation can be resolved by testing all the possibilities based on (a) a priori knowledge of possible causal relationships within the system in question and (b) the observational data at hand. In fact, hypothesized causal relationships can be represented as paths in a directed acyclic (loop-free) graph, in which variables are represented as vertices and can appear simultaneously as predictors and responses (Lefcheck, 2016). The size and significance of each causal relationship can be then evaluated by generalized linear models (GLMs), generalized additive models (GAMs) (Shipley, 2000a), or even generalized linear mixed models (GLMMs), if both fixed and random effects are considered, for a wide range of distributional assumptions. The overall goodness of fit of the *SEM* can also be evaluated: In fact, by analyzing the structure of the graph, one can formulate a set of claims on the conditional independence among variables, and a *P*-value can be estimated for each one of these claims. Combining all these *P*-values of a *SEM* by Fisher's *C* statistic provides an indicator of the consistency between the hypothesized causal relationships and the data. The procedure described above (Shipley's d-separation test) and its extensions allow researchers to deal with models characterized by complex hierarchical or multilevel structures (Shipley, 2000a, 2009).

To date, there are only a few applications of structural equation models on sedimentary DNA data (an example of such studies based on modern sedimentary DNA is presented in Box 2), and they are mostly focused on microbiome functioning in modern lake environment (e.g., Orland et al., 2019; Zhang, Ji, Wang, Zhang, & Xu, 2019). We argue that such application will provide invaluable insight to the drivers of biodiversity change and ecosystem functioning through time, even for terrestrial environments. A possible approach to do so is to use an ordination method to extract a few major axes, as representation of the main trends of biodiversity change along time, and then to integrate these axes into a *SEM* as variables among other environmental variables. However, as in other approaches to seek causal relationships with paleorecords, caution must be taken when dealing with the different time intervals among records, because causal relationships may hold in one scale and break in another (Mills et al., 2016, also see Section 3.4). Adding to this challenge are the irregular sampling time intervals, which can be extremely difficult to avoid in sediment analysis, especially when sedimentation rate is not constant (Birks et al., 2012; Mills et al., 2016; Simpson & Anderson, 2009). Finally, statistical tests of consistency such as Shipley's d-separation test do not guarantee that the tested *SEM*'s structure is adequate, and solid knowledge of the ecological processes involved is essential in order to formulate an appropriate *SEM*. Additional details and guidance to the application of *SEM* in ecology are discussed in Fan et al. (2016).

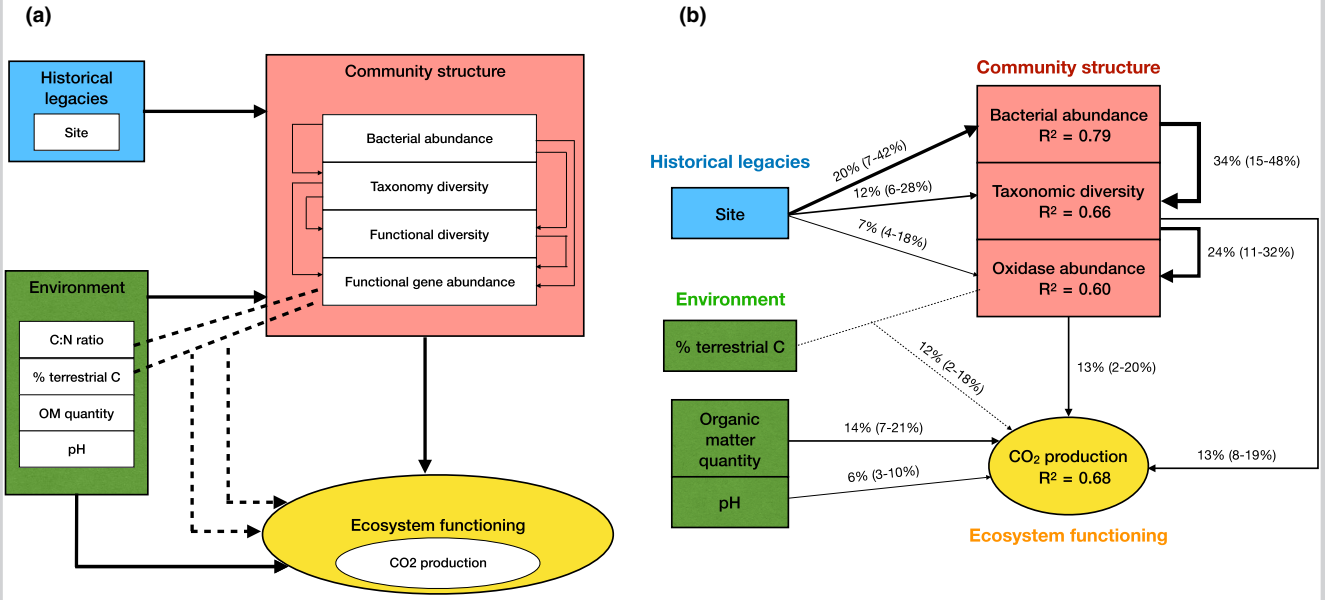
4 | CONCLUSIONS

Despite *Seda*DNA's relatively complicated data generation process (Taberlet et al., 2018) and the uncertainties in defining biodiversity units based on it, it provides an exceptionally large amount of information to past environmental changes. Existing numerical/statistical methods in the ecological literature can be applied to such data directly or with minor adjustments, yielding more in-depth insights to the long-term ecological processes in the past (Figure 1). Nevertheless, researchers must be aware that *Seda*DNA analysis is complicated by multiple factors, such as the uneven DNA quality due to degradation, the spatial and temporal autocorrelation, and the uneven sampling intervals. Furthermore, the origin, transport, deposition, and preservation of *Seda*DNA must be better understood, as they can have complex consequences on the *Seda*DNA detection and thus on the ecological signal obtained. Therefore, numerical approaches often need to be optimized or even developed to account for these issues, along with calibration processes, if possible. Molecular ecologists, sedimentologists, and biostatisticians should therefore closely collaborate for effective advances of *Seda*DNA-based ecological studies.

Box 2. Using structural equation modeling (SEM) to investigate the drivers of ecosystem functioning based on modern sedimentary DNA

SEM, or path analysis [*sensu* (Grace et al., 2012)], provides an effective way to assess direct and indirect causal relationships among environment drivers, biodiversity, and ecosystem functioning. A recent study based on modern sedimentary DNA and geochemical data investigated how microbial community structure, presentday environment and historical legacies influenced ecosystem functioning (measured as the CO_2 production of the sediment sample), in a freshwater lake in Ontario, Canada (Orland et al., 2019). To do so, the authors estimated for each site the bacterial abundance, microbial taxonomic diversity functional diversity, and functional gene abundance in top-layer sediment samples, based on their DNA contents. They proposed a conceptual model of pathways by which microbial community structure, present-day environment and historical legacies affect ecosystem functioning (Figure I-a), and then used path analysis based on linear mixed models to estimate the relative importance and direction of the proposed causal linkages (Figure I-b). The path analysis results (Figure I-b) highlighted the large contribution of both microbial community structure [median (95% CI): 26% (16–33%)] and environment [20% (13–29%)] to the variation in ecosystem functioning. Besides, the interaction between community structure and environment explained 12% (2–18%) of variation in ecosystem functioning. Namely, the positive effect of oxidase on CO_2 production was stronger when terrestrial C was relatively abundant. Finally, the influence of historical legacies was suggested by the variation of community structure among different sites, as a significant part of the variance in taxonomic diversity was either directly explained by site [12% (6–28%)], or indirectly through change in bacterial abundance [34% (15–48%)], which in turn was also influenced by site [20% (7–42%)].

Figure I. a. Conceptual model specifying how different ecosystem properties (community structure, environment and historical legacies) affect ecosystem functioning (CO_2 production). Solid arrows represent hypothetical causal relationships, and dashed arrows represent potential interactions between ecosystem properties. b. SEM results based on linear mixed models showing that a large part of ecosystem functioning explained by the individual and interactive effects of the environment and microbial community structure. Numbers accompanying each arrow are median (95% CI) percentage of variance explained. A variable is shown in boxes only if it has a direct or indirect effect on ecosystem function with 95% CI that exclude zero. Both a. and b. were redrawn based on Orland et al., 2019.



ACKNOWLEDGMENTS

We thank N. Yoccoz, one anonymous reviewer, and associate editor D. Frisch for their constructive comments on the previous versions of this manuscript. GFF has received funding from the European Research Council under the European Community Horizon 2020 Program, Grant Agreement no.772284 (IceCommunities). The Laboratoire d'Écologie Alpine is part of Labex OSUG@2020 (ANR10 LABX56). Both authors declare that they have no competing interests.

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

W.C. and G.F.F. jointly designed this review. W.C. wrote the first version of the manuscript, with subsequent contribution by G.F.F.

DATA AVAILABILITY STATEMENT

The main text of the article contains no original analysis of any new data. The data analyzed in Box 1 are available at https://advances.sciencemag.org/content/4/5/eaar4292/tab-figures-data.

ORCID

Wentao Chen https://orcid.org/0000-0002-2665-581X

REFERENCES

Alsos, I. G., Lammers, Y., Yoccoz, N. G., Jørgensen, T., Sjögren, P., Gielly, L., & Edwards, M. E. (2018). Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. *PLoS One*, *13*(4), e0195403. https://doi.org/10.1371/journal.pone.0195403

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, *62*(1), 245–253. https://doi.org/10.1111/j.1541-0420.2005.00440.x

Bailey, L. L., Simons, T. R., & Pollock, K. H. (2004). Estimating site occupancy and species detection probability parameters for terrestrial salamanders. *Ecological Applications*, *14*(3), 692–702. https://doi.org/10.1890/03-5012

Bálint, M., Pfenninger, M., Grossart, H.-P., Taberlet, P., Vellend, M., Leibold, M. A., … Bowler, D. (2018). Environmental DNA time series in ecology. *Trends in Ecology and Evolution*, *33*, 945–957. https://doi.org/10.1016/j.tree.2018.09.003

Bauman, D., Drouet, T., Dray, S., & Vleminckx, J. (2018). Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. *Ecography*, *41*(10), 1638–1649. https://doi.org/10.1111/ecog.03380

Bellemain, E., Davey, M. L., Kauserud, H., Epp, L. S., Boessenkool, S., Coissac, E., … Brochmann, C. (2013). Fungal palaeodiversity revealed using high-throughput metabarcoding of ancient DNA from arctic permafrost. *Environmental Microbiology*, *15*(4), 1176–1189. https://doi.org/10.1111/1462-2920.12020

Bennett, K. D. (1996). Determination of the number of zones in a biostratigraphical sequence. *New Phytologist*, *132*(1), 155–170. https://doi.org/10.1111/j.1469-8137.1996.tb04521.x

Birks, H. J. B., Lotter, A. F., Juggins, S., & Smol, J. P. (2012). Tracking environmental Change using lakes sediment: Volume 5 data handling and numerical techniques. *Library*.

Blanchet, F. G., Legendre, P., & Borcard, D. (2008). Modelling directional spatial processes in ecological data. *Ecological Modelling*, *215*(4), 325–336. https://doi.org/10.1016/j.ecolmodel.2008.04.001

Borcard, D., Gillet, F., & Legendre, P. (2011). *Numerical ecology with R (Use R!)*. New York, NY: Springer.

Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, *73*(3), 1045–1055. https://doi.org/10.2307/1940179

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*(4), 325–349.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). Boca Raton, FL: Chapman & Hall.

Capo, E., Debroas, D., Arnaud, F., Guillemot, T., Bichet, V., Millet, L., … Domaizon, I. (2016). Long-term dynamics in microbial eukaryotes communities: A paleolimnological view based on sedimentary DNA. *Molecular Ecology*, *25*, 5925–5943. https://doi.org/10.1111/mec.13893

Chambert, T., Miller, D. A., & Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, *96*(2), 332–339. https://doi.org/10.1890/14-1507.1

Chao, A., Chiu, C.-H., & Jost, L. (2014). Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through hill numbers. *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 297–324. https://doi.org/10.1146/annurev-ecolsys-120213-091540

Chapuis, J. L., Boussès, P., & Barnaud, G. (1994). Alien mammals, impact and management in the French subantarctic islands. *Biological Conservation*, *67*, 97–104.

Chen, W., & Ficetola, G. F. (2019). Conditionally autoregressive models improve occupancy analyses of autocorrelated data: An example with environmental DNA. *Molecular Ecology Resources*, *19*, 163–175. https://doi.org/10.1111/1755-0998.12949

Collins, S. L. (2000). Disturbance frequency and community stability in native tallgrass prairie. *The American Naturalist*, *155*(3), 311–325. https://doi.org/10.1086/303326

Davis, A. J., Williams, K. E., Snow, N. P., Pepin, K. M., & Piaggio, A. J. (2018). Accounting for observation processes across multiple levels of uncertainty improves inference of species distributions and guides adaptive sampling of environmental DNA. *Ecology and Evolution*, *8*, 10879–10892. https://doi.org/10.1002/ece3.4552

De'ath, G. (2002). Multivariate regression tree: A new technique for modeling species-environment relationships. *Ecology*, *83*(4), 1105–1117. https://doi.org/10.1890/0012-9658(2002)083[1105:MRTANT]2.0.CO;2

Dorazio, R. M., & Erickson, R. A. (2018). ednaoccupancy: An r package for multiscale occupancy modelling of environmental DNA data. *Molecular Ecology Resources*, *18*(2), 368–380. https://doi.org/10.1111/1755-0998.12735

Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, *16*(2), 129–138. https://doi.org/10.1111/j.1466-8238.2006.00279.x

Dormann, C. F., M. McPherson, J., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., … Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Dougherty, M. M., Larson, E. R., Renshaw, M. A., Gantz, C. A., Egan, S. P., Erickson, D. M., & Lodge, D. M. (2016). Environmental DNA (eDNA) detects the invasive rusty crayfish *Orconectes rusticus* at low abundances. *Journal of Applied Ecology*, *53*(3), 722–732. https://doi.org/10.1111/1365-2664.12621

Dunker, K. K. J., Sepulveda, A. J. A. A. J. A., Massengill, R. L. R., Olsen, J. J. B. J., Russ, O. L., Wenburg, J. J. K. J., … Shea, C. (2016). Potential of environmental DNA to evaluate northern pike (*Esox lucius*) eradication efforts: An experimental test and case study. *PLoS ONE*, *11*(9), 1–21. https://doi.org/10.5061/dryad.16m53.Funding

Edwards, M. E., Alsos, I. G., Yoccoz, N., Coissac, E., Goslar, T., Gielly, L., … Taberlet, P. (2018). Metabarcoding of modern soil DNA gives a highly local vegetation signal in Svalbard tundra. *The Holocene*, *28*(12), 2006–2016. https://doi.org/10.1177/0959683618798095

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*, *10*(7), 1–16. https://doi.org/10.1371/journal.pone.0130324

Epp, L. S., Gussarova, G., Boessenkool, S., Olsen, J., Haile, J., Schrøder-Nielsen, A., … Brochmann, C. (2015). Lake sediment multi-taxon DNA from North Greenland records early post-glacial appearance of vascular plants and accurately tracks environmental changes. *Quaternary*

Science Reviews, *117*(0318), 152–163. https://doi.org/10.1016/j.quascirev.2015.03.027

Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y., Jerde, C. L., … Lodge, D. M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 29–41. https://doi.org/10.1111/1755-0998.12433

Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: An updated review. *Ecological Processes*, *5*(1), 1–12. https://doi.org/10.1186/s13717-016-0063-3

Ficetola, G. F., Manenti, R., & Taberlet, P. (2019). Environmental DNA and metabarcoding for the study of amphibians and reptiles: Species distribution, the microbiome, and much more. *Amphibia-Reptilia*, *40*, 129–148. https://doi.org/10.1163/15685381-20191194

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., … Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*(3), 543–556. https://doi.org/10.1111/1755-0998.12338

Ficetola, G. F., Poulenard, J., Sabatier, P., Messager, E., Gielly, L., Leloup, A., … Arnaud, F. (2018). DNA from lake sediments reveals long-term ecosystem changes after a biological invasion. *Science Advances*, *4*, 1–9. https://doi.org/10.1126/sciadv.aar4292

Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Molecular Ecology Resources*, *16*(3), 604–607. https://doi.org/10.1111/1755-0998.12508

Fiske, I., & Chandler, R. (2015). unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, *43*(10), 1–23. https://doi.org/10.18637/jss.v043.i10

Fonseca, V. G. (2018). Pitfalls in relative abundance estimation using eDNA metabarcoding. *Molecular Ecology Resources*, *18*(5), 923–926. https://doi.org/10.1111/1755-0998.12902

Furlan, E. M., Gleeson, D., Hardy, C. M., & Duncan, R. P. (2016). A framework for estimating the sensitivity of eDNA surveys. *Molecular Ecology Resources*, *16*(3), 641–654. https://doi.org/10.1111/1755-0998.12483

Giguet-Covex, C., Ficetola, G. F., Walsh, K., Poulenard, J., Bajard, M., Fouinat, L., … Arnaud, F. (2019). New insights on lake sediment DNA from the catchment: Importance of taphonomic and analytical issues on the record quality. *Scientific Reports*, *9*(1), 1–21. https://doi.org/10.1038/s41598-019-50339-1

Giguet-Covex, C., Pansu, J., Arnaud, F., Rey, P.-J., Griggo, C., Gielly, L., … Taberlet, P. (2014). Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nature Communications*, *5*, 3211. https://doi.org/10.1038/ncomms4211

Gilbert, B., & Bennett, J. R. (2010). Partitioning variation in ecological communities: Do the numbers add up? *Journal of Applied Ecology*, *47*(5), 1071–1082. https://doi.org/10.1111/j.1365-2664.2010.01861.x

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., … Taberlet, P. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307. https://doi.org/10.1111/2041-210X.12595

Grace, J. B. (2006). *Structural equation modeling and natural systems*. Cambridge, UK: Cambridge University Press.

Grace, J. B., Schoolmaster Jr, D. R., Guntenspergen, G. R., Little, A. M., Mitchell, B. R., Miller, K. M., & Schweiger, E. W. (2012). Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere*, *3*(8), 1–44.

Greenacre, M. (2013). The contributions of rare objects in correspondence analysis. *Ecology*, *94*(1), 241–249. https://doi.org/10.1890/11-1730.1

Griffin, J. E., Matechou, E., Buxton, A. S., Bormpoudakis, D., & Griffiths, R. A. (2019). Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*, 377–392. https://doi.org/10.1111/rssc.12390

Grimm, E. C. (1987). Constrained cluster analysis By the method of incremental sum of squares. *Computers and Geosciences*, *13*(1), 13–35. https://doi.org/10.1016/0098-3004(87)90022-7

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: Advances, challenges and opportunities. *Ecography*, *40*(2), 281–295. https://doi.org/10.1111/ecog.02445

Guillera-Arroita, G., Lahoz-Monfort, J. J., Rooyen, A. R., Weeks, A. R., & Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution*, *8*, 1081–1091. https://doi.org/10.1111/2041-210X.12743

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*(2), 427–432. https://doi.org/10.2307/1934352

Ives, A. R., & Zhu, J. (2011). Statistics for correlated data: Phylogenies, space, and time. *Ecological Applications*, *16*(1), 20–32.

Jombart, T., Dray, S., & Dufour, A. B. (2009). Finding essential scales of spatial variation in ecological data: A multivariate approach. *Ecography*, *32*(1), 161–168. https://doi.org/10.1111/j.1600-0587.2008.05567.x

Jost, L., Chao, A., & Chazdon, R. L. (2011). Compositional similarity and beta diversity. In A. E. Magurran & B. J. McGill (Eds.), *Biological diversity: Frontiers in measurement and assessment* (pp. 66–84). New York, NY: Oxford University Press Inc.

Kéry, M. (2010). *Introduction to WinBUGS for ecologists*. Burlington: Academic.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular Ecology Resources*, *16*(3), 673–685. https://doi.org/10.1111/1755-0998.12486

Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modelling in r for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, *7*(5), 573–579. https://doi.org/10.1111/2041-210X.12512

Legendre, P., & Borcard, D. (2018). Box-Cox-chord transformations for community composition data prior to beta diversity analysis. *Ecography*, *41*, 1820–1824. https://doi.org/10.1111/ecog.03498

Legendre, P., & Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, *129*(2), 271–280. https://doi.org/10.1007/s004420100716

Legendre, P., & Gauthier, O. (2014). Statistical methods for temporal and space – Time analysis of community composition data. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1778), 20132728. https://doi.org/10.1098/rspb.2013.2728

Legendre, P., & Legendre, L. F. J. (2012). *Numerical ecology* (3rd ed.). Amsterdam, The Netherlands: Elsevier.

Levene, H. (1961). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics. Essays in Honor of Harold Hotelling* (pp. 279–292). Stanford, CA: Stanford University Press.

Loreau, M., & de Mazancourt, C. (2013). Biodiversity and ecosystem stability: A synthesis of underlying mechanisms. *Ecology Letters*, *16*, 106–115. https://doi.org/10.1111/ele.12073

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., Hector, A., … Wardle, D. A. (2001). Biodiversity and ecosystem functioning: Current knowledge and future challenges. *Science*, *294*(5543), 804–808. https://doi.org/10.1126/science.1064088

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local

extinction when a species is detected imperfectly. *Ecology*, *84*(8), 2200–2207. https://doi.org/10.1002/cpa.3160360305

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., & Hines, J. E. (2018). *Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence* (2nd ed.). San Diego, CA: Academic Press.

Mills, K., Schillereff, D., Saulnier-Talbot, É., Gell, P., Anderson, N. J., Arnaud, F., … Mcgowan, S. (2016). Deciphering long – Term records of natural variability and human impact as recorded in lake sediments: A palaeolimnological puzzle. *Wiley Interdisciplinary Reviews*, *4*(2), e1195. https://doi.org/10.1002/wat2.1195

Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M. K., … Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, *18*(5), 927–939. https://doi.org/10.1111/1755-0998.12895

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2017). *vegan: Community ecology package*. Retrieved from https://cran.r-project.org/package=vegan

Olajos, F., Bokma, F., Bartels, P., Myrstener, E., Rydberg, J., Öhlund, G., … Englund, G. (2018). Estimating species colonization dates using DNA in lake sediment. *Methods in Ecology and Evolution*, *9*(3), 535–543. https://doi.org/10.1111/2041-210X.12890

Orland, C., Emilson, E. J. S., Basiliko, N., Mykytczuk, N. C. S., Gunn, J. M., & Tanentzap, A. J. (2019). Microbiome functioning depends on individual and interactive effects of the environment and community structure. *ISME Journal*, *13*(1), 1–11. https://doi.org/10.1038/s41396-018-0230-x

Paliy, O., & Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*, *25*(5), 1032–1057. https://doi.org/10.1111/mec.13536

Pansu, J., De Danieli, S., Puissant, J., Gonzalez, J.-M., Gielly, L., Cordonnier, T., … Cécillon, L. (2015). Landscape-scale distribution patterns of earthworms inferred from soil DNA. *Soil Biology and Biochemistry*, *83*, 100–105. https://doi.org/10.1016/j.soilbio.2015.01.004

Pansu, J., Giguet-Covex, C., Ficetola, G. F., Gielly, L., Boyer, F., Zinger, L., … Choler, P. (2015). Reconstructing long-term human impacts on plant communities: An ecological approach based on lake sediment DNA. *Molecular Ecology*, *24*(7), 1485–1498. https://doi.org/10.1111/mec.13136

Pansu, J., Winkworth, R. C., Hennion, F., Gielly, L., Taberlet, P., & Choler, P. (2015). Long-lasting modification of soil fungal diversity associated with the introduction of rabbits to a remote sub-Antarctic archipelago. *Biology Letters*, *11*(9), 20150408. https://doi.org/10.1098/rsbl.2015.0408

Parducci, L., Bennett, K. D., Ficetola, G. F., Alsos, I. G., Suyama, Y., Wood, J. R., … Pedersen, M. W. (2017). Ancient plant DNA in lake sediments. *New Phytologist*, *214*(3), 924–942. https://doi.org/10.1111/NPH.14470

Parducci, L., Jorgensen, T., Tollefsrud, M. M., Elverland, E., Alm, T., Fontana, S. L., … Willerslev, E. (2012). Glacial survival of boreal trees in northern Scandinavia. *Science*, *335*(6072), 1083–1086. https://doi.org/10.1126/science.1216043

Pedersen, M. W., Ruter, A., Schweger, C., Friebe, H., Staff, R. A., Kjeldsen, K. K., … Willerslev, E., 2016). Postglacial viability and colonization in North America's ice-free corridor. *Nature*, *537*(7618), 45–49. https://doi.org/10.1038/nature19085

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., … Willerslev, E. (2015). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *370*(1660), 20130383. https://doi.org/10.1098/rstb.2013.0383

Pielou, E. C. (1984). *The interpretation of ecological data: A primer on classification and ordination*. New York, NY: John Wiley & Sons.

Richardson, M. J. (2001). Diversity and occurrence of coprophilous fungi. *Mycological Research*, *105*(4), 387–402. https://doi.org/10.1017/S0953756201003884

Schmidt, B. R., Kéry, M., Ursenbacher, S., Hyman, O. J., & Collins, J. P. (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*(7), 646–653. https://doi.org/10.1111/2041-210X.12052

Shipley, B. (2000a). A new inferential test for path models based on directed acyclic graphs. *Structural Equation Modeling*, *7*(2), 206–218. https://doi.org/10.1207/S15328007SEM0702_4

Shipley, B. (2009). Confirmatory path analysis in a generalized multilevel context. *Ecology*, *90*(2), 363–368. https://doi.org/10.1890/08-1034.1

Shipley, B. (2016). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference with R*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511605949

Simpson, G. L., & Anderson, N. J. (2009). Deciphering the effect of climate change and separating the influence of confounding factors in sediment core records using additive models. *Limnology and Oceanography*, *54*(6part2), 2529–2541. https://doi.org/10.4319/lo.2009.54.6_part_2.2529

Sommeria-Klein, G., Zinger, L., Taberlet, P., Coissac, E., & Chave, J. (2016). Inferring neutral biodiversity parameters using environmental DNA data sets. *Scientific Reports*, *6*, 35644. https://doi.org/10.1038/srep35644

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford, UK: Oxford University Press.

ter Braak, C. J. F. (1985). Correspondence analysis of incidence and abundance data: Properties in terms of a unimodal response model. *Biometrics*, *41*(4), 859. https://doi.org/10.2307/2530959

ter Braak, C. J. F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis stable. *Ecology*, *67*(5), 1167–1179.

ter Braak, C. J. F. (1988). Partial canonical correspondence analysis. In *Classification and related methods of data analysis: proceedings of the First Conference of the International Federation of Classification Societies (IFCS), Technical University of Aachen, F.R.G., 29 June-1 July, 1987* (pp. 551–558). Retrieved from http://edepot.wur.nl/241165%5Cnhttps://books.google.nl/books/about/Classification_and_related_methods_of_da.html?id=gVUZAQAAIAAJ&redir_esc=y

Vasselon, V., Bouchez, A., Rimet, F., Jacquet, S., Trobajo, R., Corniquel, M., … Domaizon, I. (2018). Avoiding quantification bias in metabarcoding: Application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods in Ecology and Evolution*, *9*(4), 1060–1069. https://doi.org/10.1111/2041-210X.12960

Williams, B. K. (1983). Some observations of the use of discriminant analysis in ecology. *Ecology*, *64*(5), 1283–1291. https://doi.org/10.2307/1937836

Willoughby, J. R., Wijayawardena, B. K., Sundaram, M., Swihart, R. K., & DeWoody, J. A. (2016). The importance of including imperfect detection models in eDNA experimental design. *Molecular Ecology Resources*, *16*(4), 837–844. https://doi.org/10.1111/1755-0998.12531

Yoccoz, N. G., Bråthen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., … Taberlet, P. (2012). DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, *21*(15), 3647–3655. https://doi.org/10.1111/j.1365-294X.2012.05545.x

Zhang, Y., Ji, G., Wang, C., Zhang, X., & Xu, M. (2019). Importance of denitrification driven by the relative abundances of microbial communities in coastal wetlands. *Environmental Pollution*, *244*, 47–54. https://doi.org/10.1016/j.envpol.2018.10.016

Zimmermann, H. H., Raschke, E., Epp, L. S., Stoof-Leichsenring, K. R., Schwamborn, G., Schirrmeister, L., … Herzschuh, U. (2016). Sedimentary ancient DNA and pollen reveal the composition of plant organic matter in Late Quaternary permafrost sediments of the Buor Khaya Peninsula (north-eastern Siberia). *Biogeosciences Discussions*, *14*(3), 1–50. https://doi.org/10.5194/bg-2016-386

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., … Taberlet, P. (2019). DNA metabarcoding – Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*, 1857–1862.

Zobel, M., Davison, J., Edwards, M. E., Brochmann, C., Coissac, E., Taberlet, P., … Moora, M. (2018). Ancient environmental DNA reveals shifts in dominant mutualisms during the late Quaternary. *Nature Communications*, *9*(1), 139. https://doi.org/10.1038/s41467-017-02421-3

How to cite this article: Chen W, Ficetola GF. Numerical methods for sedimentary-ancient-DNA-based study on past biodiversity and ecosystem functioning. *Environmental DNA*. 2020;2:115–129. https://doi.org/10.1002/edn3.79