Environmental DNA metabarcoding

Deiner, Kristy; Bik, Holly M.; Machler, Elvira; Seymour, Mathew; Lacoursiere-Roussel, Anais; Altermatt, Florian; Creer, Simon; Bista, Iliana; Lodge, David M.

Molecular Ecology Resources

DOI:
10.1111/mec.14350

Published: 01/11/2017

Peer reviewed version

Environmental DNA metabarcoding: transforming how we survey animal and plant communities

Kristy Deiner [1*], Holly M. Bik [2] Elvira Mächler [3,4], Mathew Seymour [5], Anaïs Lacoursière-Roussel [6], Florian Altermatt [3,4], Simon Creer [5], Iliana Bista [5,7] David M. Lodge [1], Natasha de Vere [8,9], Michael E. Pfrender [10], Louis Bernatchez [6]

Addresses:

[1] Cornell University, Atkinson Center for a Sustainable Future, and Department of Ecology and Evolutionary Biology, Ithaca, NY 14850 USA

[2] Department of Nematology, University of California, Riverside, CA 92521, USA

[3] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department of Aquatic Ecology, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland.

[4] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland.

[5] Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Environment Centre Wales Building, Bangor University, Deiniol Road, Bangor, Gwynedd LL57 2UW, UK

[6] IBIS (Institut de Biologie Intégrative et des Systèmes), Université Laval, Québec, QC G1V 0A6 Canada

[7] Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK

[8] Conservation and Research Department, National Botanic Garden of Wales, Llanarthne, Carmarthenshire, SA32 8HG, UK

[9] Institute of Biological, Environmental and Rural Sciences, Aberystwth University, SY23 3FL, UK

[10] Department of Biological Sciences and Environmental Change Initiative, University of Notre Dame, Notre Dame, Indiana, 46556 USA

Keywords: Macro-organism, eDNA, species richness, bioinformatic pipeline, conservation, ecology, invasive species, biomonitoring, citizen science

Corresponding author: Kristy Deiner, Cornell University, Department of Ecology and Evolutionary Biology, Ithaca, NY 14850 USA, alpinedna@gmail.com

Running title: Macro-organism eDNA metabarcoding

Abstract

The genomic revolution has fundamentally changed how we survey biodiversity on earth. High-throughput sequencing ('HTS') platforms now enable the rapid sequencing of DNA from diverse kinds of environmental samples (termed 'environmental DNA' or 'eDNA'). Coupling HTS with our ability to associate sequences from eDNA with a taxonomic name is called 'eDNA metabarcoding' and offers a powerful molecular tool capable of non-invasively surveying species richness from many ecosystems. Here, we review the use of eDNA metabarcoding for surveying animal and plant richness, and the challenges in using eDNA approaches to estimate relative abundance. We highlight eDNA applications in freshwater, marine, and terrestrial environments, and in this broad context, we distill what is known about the ability of different eDNA sample types to approximate richness in space and across time. We provide guiding questions for study design and discuss the eDNA metabarcoding workflow with a focus on primers and library preparation methods. We additionally discuss important criteria for consideration of bioinformatic filtering of data sets, with recommendations for increasing transparency. Finally, looking to the future, we discuss emerging applications of eDNA metabarcoding in ecology, conservation, invasion biology, biomonitoring, and how eDNA metabarcoding can empower citizen science and biodiversity education.

Introduction

Anthropogenic influences are causing unprecedented changes to the rate of biodiversity loss and, consequently, ecosystem function (Cardinale *et al.* 2012). Accordingly, we need rapid biodiversity survey tools for measuring fluctuations in species richness to inform conservation and management strategies (Kelly *et al.* 2014). Multi-species detection using DNA derived from environmental samples (termed 'environmental DNA' or 'eDNA') using high-throughput sequencing ('HTS') (Box 1), is a fast and efficient method to survey species richness in natural communities (Creer *et al.* 2016). Bacterial and fungal taxonomic richness (i.e., richness of microorganisms) is routinely surveyed using eDNA metabarcoding and is a powerful complement to conventional culture-based methods (e.g., Caporaso *et al.* 2011; Tedersoo *et al.* 2014). Over the last decade, it has been recognized that animal and plant communities can be surveyed in a similar fashion (Taberlet *et al.* 2012b; Valentini *et al.* 2009).

Many literature reviews summarize how environmental DNA (eDNA) can be used to detect biodiversity, but they focus on single species detections, richness estimates from community DNA (see Box 1 for definition for how this differs and can be confused with eDNA), or general aspects of using eDNA for detection of biodiversity in a specific field of study (Table S1). To compliment these many recent reviews, here we concentrate on four aspects: a summary of eDNA metabarcoding studies on animals and plants to date, knowns and unknowns surrounding the spatial and temporal scale of eDNA information, guidelines and challenges for eDNA study design (with a specific focus on primers and library preparation), and emerging applications of eDNA metabarcoding in the basic and applied sciences.

Surveying species richness and relative abundance with eDNA metabarcoding

Conventional physical, acoustic, and visual-based methods for surveying species richness and relative abundance have been the major ways we observe biodiversity, yet they are not without limitations. For instance, despite highly specialized identification by experts, in some taxonomic groups identification errors are common (Bortolus 2008; Stribling *et al.* 2008). Conventional physical methods can also cause destructive impacts on the environment and to biological communities (Wheeler *et al.* 2004), making them difficult to apply in a conservation context. Furthermore, when a species' behavior or size makes it difficult to survey them (e.g. small bodied or elusive species), conventional methods can require specialized equipment or species-specific observation times, thus making species richness and relative abundance estimates for entire communities intractable (e.g., many amphibians and reptiles, Erb *et al.* 2015; Price *et al.* 2012). These reasons highlight the continued need to develop improved ways to survey global biodiversity, and the unique ways eDNA metabarcoding can complement conventional methods.

Species richness: eDNA metabarcoding compared with conventional methods

Environmental DNA metabarcoding can complement (and overcome the limitations of) conventional methods by targeting different species, sampling greater diversity, and increasing the resolution of taxonomic identifications (Table 1). For example, Valentini *et al.* (2016) demonstrated that, for many different aquatic systems, the number of amphibian species detected using eDNA metabarcoding was equal to or greater than the number detected using conventional methods. When terrestrial hematophagous leeches were used as collectors of eDNA (blood of hosts), endangered and elusive vertebrate species were detected using eDNA metabarcoding and served as a valuable complement to camera trap surveys in a remote geographic region (Schnell *et al.* 2015b). In plants, Kraaijeveld *et al.* (2015) demonstrated that eDNA metabarcoding of filtered air samples allowed pollen to be identified with greater taxonomic resolution relative to visual methods.

The ways that eDNA can complement and extend conventional surveys are promising, but the spatial and temporal scale of inference is likely to differ between conventional and molecular methods. For example, in a river Deiner *et al.* (2016) showed on a site by site basis that the eDNA metabarcoding method resulted in higher species detection compared to a conventional physical-capture method (i.e., kicknet sampling) (Table 1). However, eDNA in this case may have detected greater species richness at a site not because the species themselves are present, but rather because their DNA has been transported from another location upstream, creating an inference challenge in space and time for eDNA species detections. Therefore, research is needed to understand the complex spatiotemporal dynamics of the various eDNA sample types (Fig 1), which at present we know very little about. In addition, all sampling methods have inherent biases caused by their detection probabilities. Detection probabilities often vary by species, habitat, and detection method (e.g., the mesh size of a net or a primer's match to a target DNA sequence) and use of bias-corrected species richness estimators will be important to account for these biases when conducting statistical comparisons between the outcomes in measured richness (Gotelli & Colwell 2011; Olds *et al.* 2016).

Future methodological comparisons could also benefit from a quantitative ecological approach in the design of sampling by matching sample effort and scope of sampling between eDNA and conventional methods. Multimethod species distribution modeling or site occupancy modeling is one example for how this can be achieved and has been demonstrated in cases comparing qPCR for a single species and conventional methods (Hunter *et al.* 2015; Rees *et al.* 2014a; Schmelzle & Kinziger 2016; Schmidt *et al.* 2013), but rarely for eDNA metabarcoding

(Ficetola *et al.* 2015). Thus, we expect the robustness of eDNA metabarcoding to reveal species richness estimates for animals and plants will be improved by coupling distribution or occupancy modeling with studies to determine the scale of inference in space and time for an eDNA sample (Fig. 1).

Species relative abundance: eDNA metabarcoding compared with conventional methods

Estimating a species' relative abundance using eDNA metabarcoding is an intriguing possibility. Here we focus on the evidence from animals in aquatic systems. Controlled studies based on detection of a single animal species in small ecosystems, such as in aquaria and mesocosms (e.g., Minamoto *et al.* 2012; Moyer *et al.* 2014; Pilliod *et al.* 2013; Thomsen *et al.* 2012a), in natural freshwater systems (e.g., Doi *et al.* 2017; Lacoursière-Roussel *et al.* 2016a) and marine environments (Jo *et al.* 2017; Yamamoto *et al.* 2016) demonstrate that eDNA can be used to measure relative population abundance with a species specific primer set and qPCR. While many more controlled experiments are needed in all ecosystems to determine the relationship of abundance to copy number observed in qPCR, evidence thus far from water samples signifies that eDNA contains information about a species' relative abundance.

Overall, ascertaining abundance information using metabarcoding of eDNA for whole communities still lacks substantial evidence, but some studies in aquatic environments have shown positive relationships with between the relative number of reads and relative or rank abundance estimated with conventional methods. Evans *et al.* (2016) showed in a mesocosm setting that relative abundance of individuals and biomass was correlated with relative read abundance in mesocosms containing fishes and an amphibian. In a natural lake, Hänfling *et al.* (2016) found that the rank abundance derived from long-term monitoring was correlated with read abundance for fish species, and positively correlated with gillnet surveys conducted at the same time as eDNA sampling. In deep sea habitats, Thomsen *et al.* (2016) found that when reads for fish were pooled to the taxonomic rank of families, there was a correlation with relative abundance of individuals and biomass captured in trawls. While these examples are promising, not all studies support such findings (e.g., Lim *et al.* 2016).

Challenges to accurate abundance estimation through eDNA metabarcoding stem from multiple factors in the field and the lab (Kelly 2016). In the field, the copy number of DNA arising from an individual in an environmental sample is influenced by the characteristics of the ‘ecology of eDNA’ (e.g., its origin, state, fate, and transport) (Barnes & Turner 2016). Because different animal and plant species are likely to have different rates of eDNA production or ‘origin’ (Klymus *et al.* 2015), exhibit different ‘transport’ rates from other locations (Civade *et al.* 2016; Deiner & Altermatt 2014), or stability or ‘fate’ of eDNA in time (Bista *et al.* 2017; Yoccoz *et al.* 2012), eDNA in an environmental sample could be inconsistent relative to a species’ true local and current abundance. Therefore, continued research on how the origin, state, fate, and transport of eDNA influences estimates of relative abundance is needed before we can understand the error this may generate in our ability to estimate abundance.

In the lab, primer bias driven by mismatches with their target have been shown to skew the relative abundance of amplified DNA from mock communities (Elbrecht & Leese 2015; Piñol *et al.* 2015). Similarly, the same mechanism could alter the relative abundance of a species’ DNA amplified from eDNA (Fig. 2). Primer bias results in an increased variance in abundance of reads observed relative to their true abundance in an environmental sample (Fig.2). Another source of error is related to library preparation methods. Analysis of mock communities has shown that amount of subsampling during processing steps can drive the loss of rare reads (Leray & Knowlton 2017) and likely occurs for eDNA samples as well (Shelton *et al.* 2016). The combination of primer bias and library preparation procedures alone could cause a large variance in reads observed for any given species and could prevent rare species detection altogether (Fig 2). Technical approaches and potential solutions to alleviate primer bias and alternative library preparation methods are discussed in the “Challenges in the field, in the laboratory, and at the keyboard” section. While in the end, it may be that eDNA metabarcoding is not the most accurate method for simultaneously measuring the relative abundance for multiple species from eDNA, researchers should consider whether the eDNA metabarcoding method is accurate enough for application in a particular study or an applied setting. Other methods such as capture enrichment are being examined and are promising because they avoid PCR and hence the bias this may cause, but they do require extensive knowledge of the biodiversity to design targeted gene capture probes and they come with a greater costs for analysis (Dowle *et al.* 2016). Future studies comparing qPCR, eDNA metabarcoding, and capture enrichment will be beneficial to determine which method yields accurate estimates of relative abundance from eDNA.

Before ruling out the plausibility entirely, in the short term, simulations could certainly be used to test the effects of technical laboratory issues and account for the ecology of eDNA to decipher under what conditions reliable estimates for abundance can be achieved from eDNA metabarcoding. Promising steps in this direction have been investigated through simulation to learn the nature of how datasets deliberately “noised” conform to neutral theory parameters in estimation of rank abundance curves (Sommeria-Klein *et al.* 2016). Results from simulations studies such as this could then be used to inform mock community experiments and test hypotheses (e.g., type of error distribution expected) under realistic semi-natural environments.

Ecosystems, their sample types and known scales of inference in space and time

Freshwater ecosystems

Environmental DNA metabarcoding of different sample types has been highly successful in obtaining species richness estimates for animals in aquatic systems (Fig.1, Table 1). In one of the first seminal studies, Thomsen and colleagues (2012a) used surface water from lakes, ponds, and streams in Denmark to demonstrate that eDNA contained information about aquatic vertebrate and invertebrate species known from the region. However, there are a notable lack of eDNA metabarcoding studies assessing living aquatic plant communities, and this remains an open area for further research.

Mounting evidence suggests that the spatial and temporal scale of inference for eDNA sampled from surface water differs for rivers and lakes (Fig. 1). Specifically, river waters measure species richness present at a larger spatial scale (Deiner *et al.* 2016) compared to eDNA in lake surface waters (Hänfling *et al.* 2016). Differences between lake and river eDNA signals may be due to the transport of eDNA over larger distances in rivers compared to longer retention times of water in lake systems (Turner *et al.* 2015). However, lakes and ponds with river and surface runoff inputs, combined with lake mixing or stratification, may serve as eDNA sources for catchment level terrestrial and aquatic diversity estimates similar to rivers (Deiner *et al.* 2016). No studies to date have estimated the sources of eDNA in surface water from a lake's catchment and related it to the diversity locally occurring in the lake. However, ancient DNA from sediment cores in lakes (sedaDNA) has been used to determine historical plant (e.g., Pansu *et al.* 2015b; Parducci *et al.* 2013) and livestock communities (Giguet-Covex *et al.* 2014), thus indicating that lakes do receive DNA from species in their catchments which can be incorporated into their sediments. For a more extensive review of sedaDNA being used to reconstruct past ecosystems see Pederson *et al.* (2015) and Brown and Blois (2016).

Most often, species richness estimates generated from eDNA in surface waters of lakes and rivers reflects recent site biodiversity, while those from eDNA found in surface sediments may reflect a temporally extended accumulation of eDNA. For example, Shaw *et al.* (2016) compared estimates of fish species richness from water and surface sediment samples. Generally they found species were detected in both samples, but estimates of species richness from water samples were in better agreement with the species physically present at the time of sampling. The temporal scale of inference in surface sediments is largely unknown and needs further examination (Fig. 1).

In addition to surface freshwater (~1%), groundwater (~30%) and ice (~69%) comprise much of Earth's freshwater (Gleick 1993). While the other freshwater habitats far surpass the amount of surface water, their extant biodiversity is rather poorly described (Danielopol *et al.* 2000). Groundwater is known to harbor a wide range of specialist taxa which are difficult to assess using conventional survey methods due to the inaccessibility of these habitats (Danielopol *et al.* 2000). Groundwater micro-organism metabarcoding studies have shown high fungal (Sohlberg *et al.* 2015) and bacterial (Kao *et al.* 2016) diversity, and there are examples of species-specific studies on the cave-dwelling amphibian *Proteus anguinus* (e.g., Gorički *et al.* 2017; Vörös *et al.* 2017). However, there is a clear lack of eDNA metabarcoding studies that could shed light on the diversity of a wide range of macro-organisms known to inhabit groundwater; including turbellarians, gastropods, isopods, amphipods, decapods, fishes and salamanders. The spatiotemporal scale of inference of eDNA samples from groundwater is currently unknown. Surveying eDNA in systems with knowledge of the complex hydrology and interactions between surface and ground water will be interesting places to start to reveal the scale of inference for eDNA surveys for these environments.

Environmental DNA found in sediment cores and ice core sediments generally reflects a historical biodiversity sample (Fig. 1) and is more commonly used as a source of ancient DNA (Willerslev *et al.* 2007). To date animal and plants surveyed from lake sediment cores suggest that information about terrestrial and aquatic communities can be estimated as far back as 6 to 12.6 thousand years before present (Giguet-Covex *et al.* 2014; Pedersen *et al.* 2016), whereas eDNA from sediments in ice cores have successfully been used to reconstruct communities as far back as 2000 years before present (Willerslev *et al.* 1999). The spatial scale of inference for sediment samples types has not been tested, but when samples from multiple locations are combined, large areas can be surveyed for the past presence of species (Anderson-Carpenter *et al.* 2011). For modern communities, snow has served as a viable sample type and enabled a local survey of wild canids in France (Valiere & Taberlet 2000). Environmental DNA metabarcoding of water from glacial runoff will also likely be a valuable tool to survey animal and plant richness living in glacial and subglacial habitats, which are undergoing dramatic change because of climate warming (Giersch *et al.* 2017).

Marine ecosystems

The use of eDNA metabarcoding is often described as challenging in marine ecosystems, due to the potential dilution of eDNA in large volumes of water and additional abiotic factors (salinity, tides, currents) that likely impact eDNA transport and degradation (Foote *et al.* 2012; Port *et al.* 2016; Thomsen *et al.* 2012b), not to mention the logistics involved in undertaking such surveys. Nonetheless, eDNA metabarcoding surveys of marine fish from coastal water samples have demonstrated that eDNA can detect a greater taxonomic diversity compared to conventional survey techniques (Table 1), while simultaneously improving detection of rare and vagrant fish species, and revealing cryptic species otherwise overlooked by visual assessments

(O'Donnell *et al.* 2017; Port *et al.* 2016; Thomsen *et al.* 2012b; Thomsen *et al.* 2016). Marine mammals have been surveyed with acoustic surveys and eDNA metabarcoding, and here the conventional acoustic methods detected a greater species richness (Foote *et al.* 2012). Nevertheless, this study used low sample volumes compared to other marine studies (15 – 45 mL vs. 1.5 – 3.0 L) and the authors concluded that larger sample volumes would likely lead to greater similarity between eDNA and conventional methods. In Monterey Bay, California, water sampled from depths less than 200 m or greater 200 m were used to detect marine mammals such as seals, dolphins, and whales in addition to many fishes and sharks (Andruszkiewicz *et al.* 2017). The taxonomic groups detected were spatially explicit and were found more or less in water associated with their expected habitat.

Longitudinal transport of animal and plant eDNA in marine environments is not well studied. But, similar to freshwater sediment cores from lakes, vertical transport into marine sediments is likely to preserve a large proportion of eDNA from particulate organic matter or eDNA that has become directly adsorbed onto sediment particles. This absorption shields nucleotides from degradation (particularly oxidation and hydrolysis) and facilitates long-term preservation of genetic signals over potentially large spatiotemporal scales (Fig. 1). Marine sediment eDNA concentrations have been shown to be three orders of magnitude higher than in seawater eDNA (Torti *et al.* 2015) and eDNA from both ancient and extant communities is typically recovered (Lejzerowicz *et al.* 2013). Similar to lake sediments, marine sediments can accumulate genetic information from both terrestrial and pelagic sources (Torti *et al.* 2015).

Marine sediments are difficult to sample because of the logistical effort involved in obtaining samples, which often requires ship time and specialized coring equipment. Even though much work remains to be done to understand the spatiotemporal scale of inference for marine sediment cores, comparisons between eDNA and environmental RNA (eRNA) metabarcoding are hypothesized to allow inference between present and past diversity. Environmental RNA is thought to be only available from live organisms in the community, thus the comparison between eDNA and eRNA has been investigated. In applied settings, eDNA metabarcoding of surface sediments has revealed benthic impacts of aquaculture for Atlantic salmon farming on short spatial scales using both eDNA and eRNA (Pawlowski *et al.* 2014). Guardiola *et al.* (2016) showed through a comparison of eDNA and eRNA that spatial trends in species richness from these two sources were similar, but that eDNA detected higher diversity. Overall, the fate, transport, and decomposition of animal and plant eDNA in marine environments is poorly known compared to other environments, and there is pressing need for further studies.

Terrestrial and aerial ecosystems

Environmental DNA from terrestrial sediment cores is a valuable tool for investigating past environments and reconstructing animal and plant communities (Fig. 1, Haouchar *et al.* 2014; Jørgensen *et al.* 2012; Willerslev *et al.* 2003). Animal remains also provide opportunities to reconstruct past trophic relationships. For example, eDNA metabarcoding of pellets in herbivore middens have been used to identify species in ancient animal and plant communities (Fig. 2, Murray *et al.* 2012) and DNA traces from microplant fossils within coprolites were used to reconstruct former feeding relationships in rare and extinct birds (Wood *et al.* 2012). Again here, the recent reviews of Brown & Blois (2016) and Pedersen *et al.* (Pedersen *et al.* 2015) provide a more extensive overview for how ancient DNA is used to uncover past animal and plant communities.

In modern environments, eDNA isolated from top soils has been used to characterize biodiversity in earthworms (Bienert *et al.* 2012; Pansu *et al.* 2015a), invertebrates (McGee & Eaton 2015), plants (Taberlet *et al.* 2012c; Yoccoz *et al.* 2012) and vertebrate species (Andersen *et al.* 2012). In what is perhaps the most comprehensive analysis using eDNA metabarcoding for any environment, Drummond *et al.* (2015) simultaneously surveyed all three domains of life in top soil using PCR primers that amplified five different metabarcoding regions, thus demonstrating the power of this method for assessing total richness for an area. However, the spatial scale of inference for many terrestrial eDNA samples is an open question (Fig. 1). Research on the time scale of inference for DNA in top soil suggests that long fragments of DNA break down quickly, but short fragments remain detectable for days to years after the presence of the species (Taberlet *et al.* 2012c; Yoccoz *et al.* 2012). Thus, the fragment length amplified can change the temporal resolution of a soil sample.

There are many additional sources for eDNA sampling besides soil in terrestrial ecosystems. For animals, blood meals from leeches (Schnell *et al.* 2012) and carrion flies (Calvignac-Spencer *et al.* 2013) have been used to survey mammal diversity. Saliva on browsed twigs was tested as a source of eDNA to survey ungulates (Nichols *et al.* 2012) and on predated eggs and carcasses of ground-nesting birds to discover predators or scavengers (Hopken *et al.* 2016). DNA extracted from spider webs has also been used to detect spiders and their prey (Xu *et al.* 2015). For plants, pollen within honey has revealed honey bee foraging preferences (De Vere *et al.* 2017; Hawkins *et al.* 2015). Craine *et al.* (2017) surveyed dust from indoor and outdoor environments throughout the United States and found that plant DNA from known allergens was almost twice as high outdoor compared with indoor environments. In addition to allergen detection from pollen, there remain many potential applications of dust eDNA to assess animal species richness. Fecal DNA has also been used as a source of eDNA to assess diet composition, but most studies utilizing this source of eDNA are focused on single species detections and population genetic inferences (see review from Rodgers & Janečka 2013) and are not necessarily using eDNA sources from fecal DNA to estimate species richness of terrestrial communities. Boyer *et al.* (2015) proposed that surveys of feces from generalist predators can act as ‘biodiversity capsules’ and analysis of this eDNA source should give rise to biodiversity surveys for prey communities in landscapes. While all of these sources are available, most of these sample types (e.g., leaves from a tree, fecal pellets, spider webs, and dust) do not have a known scale of inference in space and time. A single sample of eDNA from these sources is not likely to confirm species richness for more than a local scale, but combination of multiple sample sources (e.g., leaves, fecal pellets, and spider webs throughout a park) and sampled over time may allow for spatial and temporal estimates of terrestrial species richness.

Surveys of airborne eDNA have placed greater emphasis on the detection of bioaerosols that cause infection or allergic responses in animals and plants (West *et al.* 2008). For example, Kraaijeveld (2015) investigated airborne pollen that can cause hay fever and asthma in humans and showed that the source of allergenic plant pollen could be identified more accurately using eDNA from plant pollen filtered from the air compared to microscopic identification. A particularly interesting area for further research is to gain an understanding of the scale of inference for air samples in space and time (Fig. 1). While plant eDNA can be ascertained, surveying other species such as birds and insects from aerial eDNA sources has not been tested to our knowledge.

Challenges in the field, in the laboratory, and at the keyboard

Despite the obvious power of the approach, eDNA metabarcoding is affected by a host of precision and accuracy challenges distributed throughout the workflow in the field, in the laboratory, and at the keyboard (Thomsen & Willerslev 2015). Following study design (e.g., hypothesis/question, targeted taxonomic group, *etc.* Fig. 3), the current eDNA workflow consists of three components: field, laboratory, and bioinformatics. The field component consists of sample collection (e.g., water, sediment, air) that is preserved or frozen prior to DNA extraction. The laboratory component has four basic steps: 1) DNA is concentrated (if not done in the field) and purified, 2) PCR is used to amplify a target gene or region, 3) unique nucleotide sequences called 'indexes' (also referred to as 'barcodes') are incorporated using PCR or are ligated onto different PCR products, creating a 'library' whereby multiple samples can be pooled together, 4) pooled libraries are then sequenced on a high-throughput machine (most often the Illumina HiSeq or MiSeq platform). The final step after laboratory processing of samples is to computationally process the output files from the sequencer using a robust bioinformatics pipeline (Fig. 3, Box 2). Below we emphasize the important and rapidly evolving aspects of the eDNA metabarcoding workflow and give recommendations for ways to reduce error.

In the field

As for any field study, the study design is of paramount importance (Fig. 3, Box 2), since it will impact the downstream statistical power and analytical interpretation of any eDNA metabarcoding dataset. For example, sampling effort and replication (especially biological), are positively correlated with the probability of detecting the target taxa (Furlan *et al.* 2016; Willoughby *et al.* 2016). Despite the extensive evidence of the occurrence of macro-organism DNA in the environment, our fundamental understanding of what 'eDNA' is from any environmental sample is still lacking. For an illustration of this challenge, we summarize what is known about eDNA in freshwater environments. The current state-of-the-art relies on the fact that we can access eDNA by precipitating DNA from small volumes of water samples (e.g., 15 mL, Ficetola *et al.* 2008), or filter eDNA from the water column using a variety of filter sizes (0.22 µm and upwards) (Rees *et al.* 2014b). Filtration protocols lead to a working hypothesis that aqueous eDNA is either derived from cellular or organellar sources (e.g., mitochondria, Lacoursière-Roussel *et al.* 2016b; Turner *et al.* 2014; Wilcox *et al.* 2015), and precipitation protocols suggest extracellular sources (Torti *et al.* 2015). It is clear that at least some freshwater eDNA comes from intact cellular or organellar sources because it has recently been demonstrated to be available in the genomic state (Deiner *et al.* 2017b). Thus, eDNA in water exists in both un-degraded and degraded forms (Deiner *et al.* 2017b). However, continued research on the origin, state, and fate of eDNA will greatly inform numerous strategies regarding its acquisition (filtering, replication, sample volumes and spatial sampling strategies) (Barnes & Turner 2016). Many methods for solving current challenges of false negatives (e.g., use of biological replicate sampling, improved laboratory methods) and false positives (e.g., use of negative controls) in the field are explored in a recent review (Goldberg *et al.* 2016), we therefore refer readers to this review rather than treat those topics in-depth here.

In the laboratory

There are a number of recent studies that focus on the capture, preservation, and extraction of eDNA and the literature reviewed therein summarizes the important considerations and trade-offs that should be tested before a large scale study is conducted (e.g., Deiner *et al.* 2015; Renshaw *et al.* 2015; Spens *et al.* 2017). Rather than reiterate those aspects here, we focus on primer choice and library preparation. For animal and plant studies, PCR primers most often target mitochondrial or plastid loci or nuclear ribosomal RNA genes (Table S2). The standard barcoding markers defined by the Consortium for the Barcode of Life (CBOL) are Cytochrome c oxidase subunit I (COI or cox1), for taxonomical identification of animals (Hebert *et al.* 2003), and a 2-loci combination of *rbcL* and *matK* as the plant barcode (Hollingsworth *et al.* 2009) with ITS2 also suggested as valid plant barcode marker (Chen *et al.* 2010). However, there are limitations for using the standard barcoding markers in macro-organism eDNA metabarcoding. Specific to COI, other DNA regions are commonly used because not all taxonomic groups can be differentiated to species equally well (Deagle et al. 2014) and because it is challenging to design primers in this gene for a length that is suitable for short amplicon analysis, but some regions have been identified (Leray *et al.* 2013). The most common alternative markers used are mitochondrial ribosomal genes such as 12S and 16S or protein coding genes such as Cytochrome B (Table S2). Specific to the plant barcoding loci, the 2-loci primarily used for barcoding plants can be independently generated, but is not always possible to recover which fragment from each gene is associated with each other in an eDNA sample; rendering species identification using the standard plant barcode challenging. Bioinformatic methods can help resolve these situations to some extent, and may work when diversity is low in a sample (Bell *et al.* 2016). Therefore, often one or different markers are used (e.g., P6 loop of the trnL intron (Sønstebø *et al.* 2010; Taberlet *et al.* 2007)) (Table S2).

Additionally, some highly-evolving non-coding loci, such as ITS rRNA, are used (Table S2), but these markers do not always allow for the construction of alignments to determine MOTUs during data analysis because they have intragenomic variation that complicates their use in biodiversity studies (plant ITS rRNA may be an exception (Bell *et al.* 2016)). For these loci, an unknown environmental sequence is often discarded unless it has an exact database match reducing a dataset to only known and sequenced biodiversity. Due to these factors, other metabarcoding loci such as 18S rRNA genes may be more appropriate (e.g., in studies of marine invertebrates, Bik *et al.* 2012), especially if phylogenetic analysis is needed to narrow down taxonomic assignments and circumvent database limitations (Box 3).

Once the locus or loci are chosen, primers are then designed based on the taxonomic group(s) of interest within a study, and the need for broad (multiple phyla) vs. narrow (single order) coverage to test study-specific hypotheses (Fig. 3). When choosing previously designed primers (Table S2) or when designing new primers it is important to perform rigorous testing, *in silico*, *in vitro* and *in situ* to infer their utility for metabarcoding eDNA in a new study system (Elbrecht & Leese 2017; Freeland 2016; Goldberg *et al.* 2016). Amplicon size is also an important consideration because there may be a trade-off in detection with amplicon length (e.g., short fragments are more likely to amplify). However, short fragments may persist longer in the environment and increase the inference in space or time that can be made from an environmental sample (Bista *et al.* 2017; Deagle *et al.* 2006; Jo *et al.* 2017; Yoccoz *et al.* 2012). Additionally, use of more than one locus for a target group can allow for tests of consistency between loci and increase stringency of detection for any species (Evans *et al.* 2017).

Once primers are designed and PCR products are amplified, eDNA metabarcoding relies on multiplexing large numbers of samples on HTS platforms in order to make the tool cost effective. Illumina (MiSeq and HiSeq) sequencing platforms at the moment outperform other models for accuracy (Loman *et al.* 2012) and multiplexing samples is usually achieved by the incorporation of sample-specific nucleotide indices and sequencing adapters during PCR amplification. However, multiplexing creates opportunities for errors and biases. In this facet of the workflow it is important to avoid methods that induce sample specific biases in amplification

(O'Donnell *et al.* 2016) and to reduce the potential for index crossover, or "tag jumping" (see Box 2) (Schnell *et al.* 2015a). To address these issues, Illumina has developed a two-step PCR protocol using uniformly tailed primers across samples for the first step and sample specific indexes for the second PCR, which could reduce bias related to index sequence variations (Berry *et al.* 2011; Miya *et al.* 2015; O'Donnell *et al.* 2016). Regardless of the strategy employed extreme care is needed to ensure primer quality control (e.g., both use of small aliquots from stocks as well as proper cleaning of PCR amplified products to remove indexing primers after amplification (Schnell *et al.* 2015a). When a species detection is suspected as highly unlikely in a sample, single-species quantitative PCR (qPCR) can be used to verify its presence from the same eDNA sample because qPCR does not suffer from the same technical sources of error. Additional suggestions for dealing with multiplexing artifacts are suggested in Box 2 under "abundance filtering".

In addition, both positive and negative controls must be used in the lab to ensure sample integrity (Fig. 3). Use of positive control samples (either from pooled DNA extracts derived from tissue at the PCR stage, or used at the extraction stage alongside that of eDNA samples) can help evaluate sequencing efficiency and multiplexing errors in the eDNA metabarcoding workflow (Hänfling *et al.* 2016; Olds *et al.* 2016; Port *et al.* 2016). Careful thought in the construction of the mock community is needed. Typically, species not expected in the study area are used (Olds *et al.* 2016; Thomsen *et al.* 2016) such that if there is contamination during the workflow their reads can be identified, removed and serve as a control for detecting contamination when it occurs.

Negative controls should be introduced at each stage of lab work (i.e., filtration - if done in the lab, extraction, PCR, and indexing). We recommend that an equivalent amount of technical replication should be used on negative and positive controls as that carried out on actual samples (Ficetola *et al.* 2015). Furthermore, it is becoming important that negative controls are sequenced regardless of having detectable amounts of DNA because contamination can be below detection limits of quantification and sequences found in these controls can be used to detect de-multiplexing errors or used in statistical modeling to rule out false positive detections (Olds *et al.* 2016).

Finally, an important but often neglected consideration for the eDNA metabarcoding workflow, is the identification of technical artifacts that arise independently of true biological variation. For example, recently in a study focused on bacterial biodiversity using the 16S locus it was shown that a run effect can be confounded with a sample effect if it is not accounted for (e.g., by splitting sample groups across multiple Illumina runs, Chase *et al.* 2016); however, it remains to be seen whether such technical artifacts are also prevalent for loci used for metabarcoding plant and animals from eDNA (COI, 18S, ITS, etc.) and more research is needed. Until then, careful thought into how samples are pooled and run on a sequencer seems warranted in order to not confound the hypotheses being tested.

At the keyboard

Bioinformatic processing of high throughput sequence datasets requires the use of UNIX pipelines (or graphical wrappers of such tools, Bik *et al.* 2012). Metabarcoding of animal and plant community DNA is comprehensively outlined in Coissac *et al.* (2012). Below and in Box 3 we highlight the common practices to community DNA metabarcoding and deviations for studies focusing on macro-organism eDNA metabarcoding.

Bioinformatic pipelines and parameters must be carefully considered (Box 2) and it is important to work with a knowledgeable computational researcher to understand how processing can impact the biological results and conclusions. Before computationally processing an eDNA metabarcoding dataset, perhaps the strongest message from Coissac *et al.* (2012) is to identify the differences between the analysis of data derived from microbial and macro-organismal groups. Since microbial ecologists have been inspired to use sequence-based identification of taxa over the past 40 years (Creer *et al.* 2016), the range of software solutions to analyze microbial metabarcoding datasets is unsurprisingly extensive (Bik *et al.* 2012). Perhaps more importantly, a number of established and maintained databases exist featuring many of the commonly used microbial taxonomic markers for prokaryotes (Cole *et al.* 2009), microbial eukaryotes (Guillou *et al.* 2013; Pruesse *et al.* 2007; Quast *et al.* 2012) and fungi (Abarenkov *et al.* 2010), meaning that microbial datasets can be analyzed and taxonomic affiliations established are established in a straight forward way.

For macro-organism communities, pre-processing and initial quality control of eDNA metabarcoded data sets is not different from that of microbial datasets and can be acquired using packages developed either for microbial (Caporaso *et al.* 2010), or macro-organism data (Boyer *et al.* 2016), but taxonomic assignment will require a robust dataset of locus-specific reference sequences and the associated taxonomic data from a reference database (Coissac *et al.* 2012) (Box 3). Currently the two most common reference sources for macro-organisms are NCBI's nucleotide database (Benson *et al.* 2013) and the Barcode of Life Database (Ratnasingham & Hebert 2007). The utility and taxonomic breadth of these databases can be enhanced by the creation of custom-made or hybrid databases, with the obvious additional workload and cost depending on the number of focal taxa missing from current data sources. Recently, Machida *et al.* (2017) have assembled and proposed metazoan mitochondrial gene sequence datasets that can be used for taxonomic assignment for environmental samples. While these datasets do not account for future growth, their methods could be repeated at the time of any new study to generate a custom reference dataset for taxonomic assignment.

Macro-organism eDNA metabarcoding datasets are associated with advantages compared to microbial datasets because the number of taxa in any survey will be comparatively low, reducing the computational time needed for taxonomic annotation. Moreover, the species delimitation concepts and taxonomic markers associated with macro-organisms are well-developed (de Queiroz 2005) and can even be used to analyze population genetic structure (Sigsgaard *et al.* 2016; Thomsen & Willerslev 2015), or delimit species boundaries (Coissac *et al.* 2012; Hebert *et al.* 2003; Tang *et al.* 2014). Reliance on the vast knowledge we have for animal and plant taxonomy and biogeography is a distinct advantage for eDNA metabarcoding because of the independent test that it provides to calibrate and test the tool for its precision and accuracy (Deiner *et al.* 2016).

Data archiving for transparency

As eDNA applications continue to develop, all procedures used in the field, lab, and during bioinformatic data processing require a strong commitment to transparency on the part of researchers (Nekrutenko & Taylor 2012). Here, we outline best practices for eDNA metabarcoding studies of macro-organisms, following on from well-established standards in the fields of microbiology and genomics (Yilmaz *et al.* 2011). First, raw FASTQ files from any HTS run need to be submitted to the Sequence Read Archive (SRA) of NCBI or the European Nucleotide Archive (ENA) and other such public national data bases before publication. Archiving raw data in publicly available databases is common practice in virtually all genomics and transcriptomic studies because it allows studies to be re-analyzed with new computational tools and standards. In fact, archiving raw data is becoming increasingly mandatory at many evolutionary and ecology biology journals, inclusive of Molecular Ecology. Second, researchers should adhere to minimum reporting standards defined by the broader genomics community, such as the MIMARKS (Minimum information about a marker gene sequence) and MIxS (minimum information about any "x" sequence) specifications (Yilmaz *et al.* 2011). Goldberg *et al.* (2016) have made specific recommendations for upholding these reporting standards specific to eDNA studies (see Table 1 in Goldberg *et al.* 2016).

Third, computational processing of data needs to be reproducible (Sandve *et al.* 2013). For eDNA metabarcoding studies, it is increasingly common to deposit a comprehensive sample mapping file (e.g., formatted in the QIIME tab-delimited style, containing the indexes used for creating libraries so that raw data can be de-multiplexed and properly trimmed) along with MOTU clustering or taxonomic binning of results, and documentation of all bioinformatics commands, in a complementary repository such as Dryad (http://datadryad.org/), GitHub (https://github.com/github), or FigShare (http://figshare.com). Sandve *et al.* (2013) provide 10 rules that can be followed to ensure such reproducibility, and we strongly encourage researchers using eDNA metabarcoding methods to uphold these practices and take advantage of archiving intermediate steps (Box 2) of their analysis for full transparency.

Emerging applications for eDNA metabarcoding

Applications in ecology

Quantifying the richness and abundance of species in natural communities is and will continue to be a goal in many ecological studies. Information about species richness garnered from eDNA is not necessarily different from conventional approaches (Table 1), but the scale, speed, and comprehensiveness of that information is (Fig. 4). For example, Drummond *et al.*

(2015) demonstrated the near-complete analysis of biodiversity (e.g., from bacteria to animals and plants) from top soil is possible. Collection of data on this taxonomic scale opens up new opportunities with respect to measuring community composition and turnover across space and time. In addition to estimating species richness, a major area of research in ecology is determining whether observed community changes surpasses acceptable thresholds for certain desired ecosystem functions (Jackson *et al.* 2016). Biodiversity and ecosystem functioning research requires tracking species in multiple taxonomic groups and trophic levels, along with changes in ecosystem function. Environmental DNA metabarcoding has the potential to facilitate biodiversity and ecosystem function research by improving our knowledge of predator/prey relationships, mutualisms such as plant-pollinator interactions, and food webs in highly diverse systems composed of small cryptic species (e.g., De Vere *et al.* 2017; Hawkins *et al.* 2015; Xu *et al.* 2015). Knowledge of species co-occurrences and interactions in these instances will additionally foster the study of meta ecosystems and provide data to guide management decisions at the ecosystem scale (Bohan *et al.* 2017). What will remain challenging is moving beyond richness estimates to also obtaining species abundance data (Fig. 2 & 4).

Applications in conservation biology

Given the rapid rate at which biodiversity is declining worldwide (Butchart *et al.* 2010), it is critical that we improve the effectiveness of strategies to halt or reverse this loss (Thomsen & Willerslev 2015; Valentini *et al.* 2016). Accordingly, developing tools that enable rapid, cost-effective and non-invasive biodiversity assessment such as eDNA metabarcoding, especially for rare and cryptic species, is paramount (Fig. 4). Improved estimates of the distribution of vulnerable species, and done so non-invasively, would facilitate policy development and allow for efficient targeting of management efforts across habitats (Kelly *et al.* 2014; Thomsen & Willerslev 2015). For example, documenting the presence of threatened species in a habitat can trigger a suite of actions under laws pertaining to biodiversity conservation (e.g., US Endangered Species Act). Frequently, data relevant to policy are derived from monitoring efforts mandated by environmental laws imparting a significant consequence to the data collected (Kelly *et al.* 2014).

Environmental DNA-based monitoring is likely to be a tremendous boon to often underfunded public agencies charged with compliance to data-demanding laws. Specifically, eDNA metabarcoding will be useful for monitoring communities when many species are of conservation concern. Vernal pools throughout California are a prime example because they contain 20 US federally listed endangered or threatened species of plants and animals. Monitoring species richness with soil and water samples from a habitat such as this would provide a comprehensive sampling method to ascertain needed community data for their conservation and management (Deiner *et al.* 2017a). However, while eDNA metabarcoding may be important for non-invasively gaining access to the distribution of vulnerable species, it cannot be used to differentiate between alive and dead organisms or estimate many demographic parameters important of population viability analysis (Beissinger & McCullough 2002).

Quantifying baselines of animal and plant species richness and departures from those baselines, is central to the assessment of environmental impact and conservation (Taylor & Gemmell 2016). The application of eDNA metabarcoding methods to different samples types, which taken together allow inference across time (e.g., surface water and sediment layers from a core in a lake, Fig. 1) provides a unique tool to document local extinctions and long-term changes in ecosystems. Extinction models often rely on and understanding extinction timelines (reviewed in Thomsen & Willerslev 2015). The efficiency of eDNA metabarcoding to track the timing of extinctions associated with previous glacial events has been demonstrated in mammals (Haile *et al.* 2009) and plants (Willerslev *et al.* 2014). Thus, environmental DNA metabarcoding of different sample types from the same site offers an excellent opportunity to better understand the extinction consequences of perturbations and could inform scenario modeling under climate change.

Applications in invasion biology

Because one of the first applications of eDNA to macro-organisms was the detection of North American bullfrogs in French ponds (Ficetola *et al.* 2008), the method immediately came to the attention of researchers interested in invasion biology (e.g., Egan *et al.* 2013; Goldberg *et al.* 2013; Jerde *et al.* 2011; Takahara *et al.* 2013; Tréguier *et al.* 2014). These initial studies, as well as much ongoing research, continue to be based on species-specific primers, where positive amplification provides occurrence evidence for a particular invasive species. In invasion biology with eDNA, such a targeted approach is referred to as “active” surveillance (Simmons *et al.* 2015).

On the contrary, eDNA metabarcoding makes it possible to detect the presence of many species simultaneously, including species not previously suspected of being present. This broader untargeted approach is called “passive” surveillance in management applications (Fig. 4) (Simmons *et al.* 2015). On the down side, due to a trade-off in primer specificity, we expect that eDNA metabarcoding may be less sensitive in detecting some species or that the detection rate of a species can change depending on species richness. Adopting a dual approach of passive and active surveillance could be considered in cases where the risk of a new invasion is high, and where cost effective eradication plans for undesirable species are likely to be successful (Lodge *et al.* 2016).

Avoiding future introductions and reducing the spread of exotic species is paramount in natural resource policy (Lodge *et al.* 2016). Environmental DNA metabarcoding relevant to management includes early detection of incipient invasive populations in the environment, surveillance of invasion pathways, e.g., ballast water of ships (Egan *et al.* 2015; Zaiko *et al.* 2015), and the live bait trade (Mahon *et al.* 2014). While eDNA metabarcoding is not yet routinely used for biosecurity regulation of invasive species or enforcement in many settings, it has the potential to become valuable monitoring tool for biological invasions. An important challenge for the use of eDNA metabarcoding in invasive species detections are false positives and false negatives since both outcomes can trigger action or inaction when not required, causing a potentially large burden on entities responsible for invasive species mitigation and control (Fig. 4). Therefore, continued research to reduce or understand the nature of false positives and false negatives will reduce uncertainty in the tool and facilitate greater adoption.

Applications in biomonitoring

Pollution of air, water, and land resources generated from processes such as urbanization, food production, and mining is one of the many emerging global challenges we are facing in the 21st century (Vörösmarty *et al.* 2010). Determine the origin, transport, and effects of most pollution is challenging because it accumulates through both point sources (e.g., wastewater effluent) and diffused sources related to land-use types (e.g., agriculture or urbanization). In this context, the presence of tolerant or absence of sensitive organisms has been used to determine the consequences of pollution on ecosystem health throughout the world and is termed biological monitoring or ‘biomonitoring’ (Bonada *et al.* 2006). The extent to which animals and plants have been used in biomonitoring depends on the unique characteristics of the taxonomic group monitored and their relationship to the pollution of interest (Bonada *et al.* 2006; Stankovic *et al.* 2014). Most biomonitoring programs take community composition and often abundance of taxa into account and calculate what is known as a biotic index (Friberg *et al.* 2011). Biotic indices take many forms and are typically surrogates for the impacts of pollution (e.g., SPEAR index for toxicant exposure in water, Liess *et al.* 2008).

Applying eDNA metabarcoding in the context of biomonitoring is a major avenue of research. Metabarcoding of community DNA samples has shown greater sensitivity for detecting cryptic taxa or life stages and can alleviate the problem of identifying damaged specimens of which render morphological tools ineffective (Gibson *et al.* 2014; Hajibabaei *et al.* 2011). These two issues alone are known to create large variances in biotic index estimation (Pfrender *et al.* 2010). Application of eDNA metabarcoding to animals and plants used in biomonitoring requires in-depth testing of conventional survey methods and eDNA-based approaches (Fig. 4), to understand whether species richness estimates derived from the two methods result in a similar measure for the biotic index of interest or whether new biotic indices need to be development that can simultaneously consider both forms of information. Promising steps forward are being made through the DNA AquaNet COST Action (http://dnaqua.net/) which is a consortium of over 26 European union countries and four international partners working together to develop genetic tools for bioassessment of aquatic ecosystems in Europe (Leese *et al.* 2016).

Applications in citizen science and biodiversity education

The simplicity of the protocol used to collect environmental samples has created an avenue for citizen scientist programs to be built around surveying for biodiversity using eDNA (Biggs *et al.* 2015). With the development of sample kits from commercial companies specifically used for eDNA analysis (e.g., GENIDAQS, ID-GENE, Jonah Ventures, NatureMetrics, Spygen) there now exists a novel opportunity to engage the public in biodiversity science, which could accompany already established biodiversity events, such as BioBlitz (National Geographic Society). Use of eDNA metabarcoding in this context will likely provide an unprecedented tool for education and outreach about biodiversity, and increase awareness about its decline. Challenges that hinder integration of eDNA metabarcoding in citizen science projects and educational opportunities are the time and costs needed to process samples and user friendly data visualization tools to allow exploration of the data once provided. Thus, finding ways to cut costs and speed up data generation (a goal common for any application of the tool), as well as creation of applications for exploration of data on smart phones and desktops alike is needed to propel the use of eDNA applications in citizen science and education.

Conclusions

As the tool of eDNA metabarcoding continues to develop, our understanding regarding the analysis of eDNA from macro-organismal communities, including optimal field, laboratory, and bioinformatics workflows will continue to improve in the foreseeable future. Concurrently, we need to gain a better understanding of the spatial and temporal relationship between eDNA and living communities to improve precision, accuracy, and to enhance the ecological and policy relevance of eDNA (Barnes & Turner 2016; Kelly *et al.* 2014). Ultimately, the errors and uncertainties associated with eDNA metabarcoding studies can often be mitigated by thoughtful study design, appropriate primer choice, and robust sampling and replication: as Murray et al. (2015) emphasize, “no amount of high-end bioinformatics can compensate for poorly prepared samples, artefacts or contamination.”

Over time, a loop in which improved eDNA metabarcoding methods reduce uncertainty about the meaning of both positive and negative eDNA detections for a species will in turn generate the motivation for continued improvements and use of eDNA metabarcoding methods. Thus, resulting in the adoption of eDNA metabarcoding as a comparable method for estimating species richness. We predict that over the next decade eDNA metabarcoding of animals and plants will become a standard surveying tool that will complement conventional methods and accelerate our understanding of biodiversity across the planet.

Box 1: Community DNA versus environmental DNA metabarcoding of plants and animals

Terms:

Environmental DNA (eDNA). DNA captured from an environmental sample without first isolating any target organisms (Taberlet *et al.* 2012a). Traces of DNA can be from feces, mucus, skin cells, organelles, gametes or even extracellular DNA. Environmental DNA can be sampled from modern environments (e.g., seawater, freshwater, soil or air) or ancient environments (e.g., cores from sediment, ice or permafrost (e.g., cores from sediment, ice or permafrost, see Thomsen & Willerslev 2015).

Community DNA. DNA is isolated from bulk-extracted mixtures of organisms separated from the environmental sample (e.g., soil or water).

Macro-organism environmental DNA. Environmental DNA originating from animals and higher plants.

Barcoding. First defined by Hebert *et al.* (2003), the term refers to taxonomic identification of species based on single specimen sequencing of diagnostic barcoding markers (e.g., COI, *rbcL*).

Metabarcoding. Taxonomic identification of multiple species extracted from a mixed sample (community DNA or eDNA) which have been PCR amplified and sequenced on a high throughput platform (e.g., Illumina, Ion Torrent).

High Throughput Sequencing (HTS). Sequencing techniques which allow for simultaneous analysis of millions of sequences compared to the Sanger sequence method of processing one sequence at a time.

Community DNA metabarcoding: HTS of DNA extracted from specimens or whole organisms collected together, but first separated from the environmental sample (e.g., water or soil).

Molecular Operational Taxonomic Unit (MOTU): Group identified through use of cluster algorithms and a predefined percent sequence similarity (e.g., 97%) (Blaxter *et al.* 2005).

Since the inception of High Throughput Sequencing (HTS, Margulies *et al.* 2005), the use of metabarcoding as a biodiversity detection tool has drawn immense interest (e.g., Creer *et al.* 2016; Hajibabaei *et al.* 2011). However, there has yet to be clarity regarding what source material is used to conduct metabarcoding analyses (e.g., environmental DNA versus community DNA). Without clarity between these two source materials, differences in sampling, as well as differences in lab procedures, can impact subsequent bioinformatics pipelines used for data processing, and complicate the interpretation of spatial and temporal biodiversity patterns. Here we seek to clearly differentiate among the prevailing source materials used and their effect on downstream analysis and interpretation for environmental DNA metabarcoding of animals and plants compared to that of community DNA metabarcoding.

With community DNA metabarcoding of animals and plants, the targeted groups are most often collected in bulk (e.g., soil, malaise trap, or net), individuals are removed from other sample debris and pooled together prior to bulk DNA extraction (Creer *et al.* 2016). In contrast, macro-organism eDNA is isolated directly from an environmental material (e.g., soil or water) without prior segregation of individual organisms or plant material from the sample and implicitly assumes that the whole organism is not present in the sample. Of course, community DNA samples may contain DNA from parts of tissues, cells, and organelles of other organisms (e.g., gut contents, cutaneous intracellular or extracellular DNA, *etc.*). Likewise, macro-organism eDNA samples may inadvertently capture whole microscopic non-target organisms (e.g., protists, bacteria, *etc.*). Thus, the distinction can at least partly breaks down in practice.

Another important distinction between community DNA and macro-organism eDNA is that sequences generated from community DNA metabarcoding can be taxonomically verified when the specimens are not destroyed in the extraction process. Here sequences can then be generated from voucher specimens using Sanger sequencing. Since the samples for eDNA metabarcoding lack whole organisms, no such *in situ* comparisons can be made. Taxonomic affinities can therefore only be established by directly comparing obtained sequences (or through bioinformatically generated operational taxonomic units (MOTUs)), to sequences that are taxonomically annotated such as NCBI's GenBank nucleotide database (Benson *et al.* 2013), BOLD (Ratnasingham & Hebert 2007), or to self-generated reference databases from Sanger-sequenced DNA (Olds *et al.* 2016; Sønstebø *et al.* 2010; Willerslev *et al.* 2014). Then, to at least partially corroborate the resulting list of taxa, comparisons are made with conventional physical, acoustic, or visual-based survey methods conducted at the same time or compared with historical records from surveys for a location (see Table 1).

The difference in source material between community DNA and eDNA, therefore, has distinct ramifications for interpreting the scale of inference for time and space about the biodiversity detected. From community DNA it is clear that the individual species were found in that time and place, but for eDNA, the organism which produced the DNA may be upstream from the sampled location (Deiner & Altermatt 2014), or the DNA may have been transported in the feces of a more mobile predatory species (e.g., birds depositing fish eDNA, Merkes *et al.* 2014) or was previously present, but no longer active in the community and detection is from DNA that was shed years to decades before (Yoccoz *et al.* 2012). The latter means that the scale of inference both in space and time must be considered carefully when inferring the presence for the species in the community based on eDNA.

Box 2. Basic bioinformatic pipeline for eDNA metabarcoding for plants and animals

Bioinformatic processing of sequence data is one of the most critical aspects of eDNA metabarcoding studies, helping to substantiate research findings, following field and lab work components. Standardization of bioinformatics in a 'pipeline' can ensure quality and reproducibility of findings; however, some level of customization is required across studies. Customization is needed to compensate for advances in sequencing technology, software workflows, and the question being addressed. Therefore, taking raw read data and turning it into a list of taxa, requires multiple quality assurance steps – some necessary, others optional. Reaching an absolute consensus for the approaches and software used is not necessary as these will always be in flux, but here we advise careful consideration of the following pre-processing steps *at a minimum* for HTS data before embarking on further analyses (e.g., for biodiversity estimates and statistical significance). We focus primarily on processing Illumina generated data sets and therefore if the technology is different, many of the bioinformatic tools highlighted and advice is transferable to pre-processing of data produced on other platforms, but may be different.

Terms:

Chimeras: PCR artefacts made of two or more combined sequences during the extension step of PCR amplification.

Phred quality score: Quality scoring per nucleotide for Illumina sequencing providing the probability that a base call is incorrect.

Sequence merging: Combining forward (R1) and reverse (R2) reads from paired – end (PE) sequencing, using criteria such as minimum overlap or quality score.

Sequence trimming: The process of cutting / removing the beginning or end of sequencing reads. Can be performed either by searching for a specific sequence (removal of adaptors, indexes and primers) or based on quality score.

Singletons: MOTUs that appear only once in the data are likely to be rare taxa, false positives, low level contamination, or unremovedchimeras, and should be treated with appropriate consideration.

Primer – adaptor trimming. Preliminary steps of bioinformatics processing include de-multiplexing of the samples based on the indices used (unique nucleotide tags incorporated into raw sequence data) and trimming (i.e., removal) of the adaptor sequences. The adaptors are specific DNA fragments which are added during library preparation for ligation of the DNA strands to the flow cell during Illumina sequencing. Additionally, the index sequences themselves and the primer sequences should be trimmed (e.g. using software such as Cutadapt, Trimmomatic, QIIME), allowing either zero or a low level of mismatch between the exact sequence of the primer or index and the observed reads.

Merging or end trimming. Sequences from Illumina runs tend to drop in quality towards the 3' end of the reads, as phasing leads to increased noise (and lower signal) in later chemistry cycles. Thus, the quality score of reads should be reviewed to allow informed decisions on the appropriate length of end trimming (single – end runs), merging (paired end runs) and subsequent sequence quality filters. Visualizing the quality scores from raw reads or de-multiplexed sequences (using software like FastQC) will help with the selection of downstream quality cut-off levels.

When paired end (PE) sequencing is used for an amplicon of suitable size, the forward (R1) and reverse (R2) reads should be combined (merged) to form the complete amplicon. Using merged sequences improves accuracy since the lower quality bases at the tail ends of individual reads can be corrected based on the combined reads. Here, the minimum overlap for R1 and R2 reads should be specified and 'orphan' reads with little or no overlap between forward and reverse pairs can be discarded. Inspection of the quality scores, as mentioned above, can provide an estimate of optimal parameters for merging of R1 and R2 reads. Even though a specific consensus does not exist yet, in many cases an overlap of at least > 20bp is selected (Deiner *et al.* 2015; Gibson *et al.* 2015).

Quality filtering. For most HTS platforms, a Phred score is calculated and subsequently used to determine the maximum error probabilities (Bokulich *et al.* 2013). Selected strategies include filtering based on a lower Phred score cut-off, usually set at least above 20 or 30 (Bista *et al.* 2017; Elbrecht & Leese 2015; Hänfling *et al.* 2016). Quality filtering can also be performed based on maximum error (maxee) probability, which is also derived from Phred scores. The lower the maximum error, the stricter the cut-off. Selection of a maximum error filtering level of 1 or 0.5 is common in macro-organism studies (Bista *et al.* 2017; Pawlowski *et al.* 2014; Port *et al.* 2016). Additionally, in the case of single-end sequencing, or when long amplicons without sufficient overlap of the forward and reverse reads are used, it is advised that trimming should be performed from the appropriate end. It is often the case that reads are trimmed to a common length, which facilitates alignment downstream and minimizes miscalled bases since a merging step cannot be used.

Removing short reads. Many studies also select to remove short reads from the dataset before clustering since the presence of high length variation could influence the clustering process (see USEARCH manual, Edgar 2010). These sequences could result from sequencing of primer dimers which have not been removed (Pawlowski *et al.* 2014). Different studies select a variety of minimum length reads, from very short 20bp (Valentini *et al.* 2016), to medium 60 – 80 bp (Pawlowski *et al.* 2014; Shaw *et al.* 2016) and up to 100 bp (Bista *et al.* 2017; Gibson *et al.* 2015; Hänfling *et al.* 2016; Pawlowski *et al.* 2014). Note that some de-multiplexing or quality filtering workflows may automatically set a minimum sequence length when processing input data and it is advisable to check whether such a parameter is included by default.

Removing singletons and chimeras. Important steps after MOTU clustering involve removal of singletons and chimeras. Chimeras are by-products of the PCR amplification process from two or more parental sequences (chimeric), most commonly produced through an incomplete extension step (Edgar *et al.* 2011). It has been shown that when unique reads, such as chimeras and singletons, are withheld in analysis, the estimation of diversity can be severely inflated (Kunin *et al.* 2010). The nature of the chimeric sequences, which can be present as high quality reads, does not enable their removal directly through quality based end-trimming (Coissac *et al.* 2012). Removal of chimeras can be performed either *de novo* or based on a reference database. Most common practice to date is the *de novo* method since a sufficient reference database may not be available. Despite the variation in software used such as UCHIME (Edgar *et al.* 2011), obitools (Boyer *et al.* 2016), or ChimeraSlayer (Haas *et al.* 2011), there is a consensus regarding the importance of removing chimeras and singletons as a minimum quality control for bioinformatics pipeline.

Abundance filtering. In addition to quality filtering based on quality scores and removal of chimeras and singletons, many studies also employ further filtering for removal of low abundance sequences (Murray *et al.* 2015). This step arises from the need to control for laboratory contamination or because of cluster contamination on the flow cell (unique to Illumina platforms) (Olds *et al.* 2016).

The process of applying abundance filtering requires setting an MOTU abundance threshold by which MOTUs are only retained in analysis if their relative abundance is higher than the selected threshold (Bokulich *et al.* 2013). Selection of a threshold varies between studies and there is no generally accepted definition of what constitutes an insufficiently abundant read (Murray *et al.* 2015), perhaps with the exception of singletons. Abundance filtering may be applied minimally or avoided entirely, especially if stringent quality trimming parameters are applied to raw reads and detection of "rare" MOTUs is an important aspect of a study (Bokulich *et al.* 2013). Another option that could be used involves selection of a threshold based on availability of empirical data as was done in Valentini *et al.* (2016). An increasing number of studies have employed the sequencing of positive controls to establish a threshold level (Hänfling *et al.* 2016; Port *et al.* 2016; Stoeckle *et al.* 2017). Technical replicates can also be used to assess consistency as was shown to be effective with assessing omnivore diets (De Barba *et al.* 2014).

Using a positive control defined error level works by identifying the abundance of sequences in the control sample that belong to non-target taxa and can be the result of errors such as contamination. Furthermore, the distribution of *phiX* reads assigned to target samples has been used to investigate the presence of "tag-jumps" (Schnell *et al.* 2015a) and mis-assigned reads during de-multiplexing (Hänfling *et al.* 2016; Olds *et al.* 2016). The exact mechanisms for mis-assignment of reads remain unknown, but increasingly many studies are reporting this error to be between 0.01 and 0.03 % of reads (Hänfling *et al.* 2016; Olds *et al.* 2016; Stoeckle *et al.* 2017). Adjustments for this include use of a threshold approach based negative and/or positive controls and removes a low number of reads from any given sample. The issue of abundance filtering most significantly causes uncertainty in low abundance MOTUs and will continue to be a problem for detection of rare species. Therefore, to avoid negative impacts to scientific insights or management decisions, careful consideration and transparency regarding how technical artifacts are dealt with during bioinformatic data analysis is needed until these artifacts are well understood.

Recording removed data. For all quality control steps the data removal should be transparent. Often studies report the total number of sequences obtained, but then rarely show how each quality filtering step affects the number of sequences used in testing ecological hypothesis nor do researchers provide the subset of sequences that were retained or omitted. Deleting data without a clear justification does not allow transparency. Therefore, including a supplemental table in eDNA metabarcoding studies showing the number of sequences remaining after each filtration step is advised and archiving the subset of reads retained after each filtering step on a platform such as Dryad (http://datadryad.org/) or archiving the exact pipeline with version control information on a platform such as GitHub (https://github.com/) will allow for greater transparency and reproducibility of quality filtering.

Box 3: How to transform reads from HTS platforms into measures of richness

MOTU clustering. While this step is not always necessary and depends on the target set of taxa (Lacoursière-Roussel *et al.* 2016), the amplicon length sequenced (Deiner *et al.* 2016), and completeness of the reference database (Chain *et al.* 2016), clustering of sequencing reads into MOTUs is often performed prior to taxonomic assignment. MOTU clustering is the process whereby multiple reads are grouped according to set criteria of similarity based on an initial seed (Creer *et al.* 2016; Egan *et al.* 2013). Here, a centroid sequence is selected and depending on the set radius or similarity cut-off, closely related sequences are grouped under each centroid sequence (USEARCH, Edgar 2010). The level of similarity selected depends on the study and taxon used, based on the knowledge of intraspecific diversity of the studied taxon. Commonly used cut-offs range from 97% to 99% (Bista *et al.* 2017; Fahner *et al.* 2016; Olds *et al.* 2016). For example, the cut-off selected could depend on known levels of intraspecific diversity of the studied taxon, which could be estimated from an existing reference database. Some commonly used clustering algorithms include USEARCH (Edgar 2010), VSEARCH (Rognes *et al.* 2016), CROP (Bayesian clustering algorithm) (Hao *et al.* 2011), swarm (Mahé *et al.* 2014), and mothur (an alignment-based clustering method, Schloss *et al.* 2009).

Taxonomic assignment. Identification of HTS reads is achieved through a comparison of anonymous MOTU clusters/centroid sequences or direct comparisons of reads remaining after quality filtering against a reference database. Depending on the taxon of study and the marker used, the reference database may consist of publicly available sequences or study-generated reference sequences.

The challenges of taxonomic assignment have been the subject of a considerable literature so we only briefly discuss this important aspect of the bioinformatics pipeline (e.g., Bazinet & Cummings 2012). A number of different approaches have been suggested including assignment based on sequence similarity via alignment programs like BLAST or similarity searches using Hidden Markov Models such as jMMOTU (Jones *et al.* 2011), MG-RAST (Glass *et al.* 2010), sequence composition and machine learning approaches (e.g., RDP (Wang *et al.* 2007), TACOA (Diaz *et al.* 2009)), phylogenetic placement (e.g., pplacer Matsen *et al.* 2010), probabilistic taxonomic placement (e.g., PROTAX (Somervuo *et al.* 2016; Somervuo *et al.* 2017), minimum entropy decomposition (e.g., oligotyping Eren *et al.* 2015), MEGAN (Huson *et al.* 2007), and ecotag (Boyer *et al.* 2016). A number of widely used programs use combinations of these methods, for example, the program SAP (Munch *et al.* 2008) uses BLAST searches of the NCBI database and phylogenetic reconstruction to establish taxonomic identity of query sequences. Most of these methods and various derivatives are nicely discussed and compared by Bazinet and Cummings (2012). Two major determinants of the utility of these different approaches are the specific eDNA markers and the breadth and resolution of reference databases. Some markers have better representation in available databases and greater coverage of relevant species diversity. Taxonomic assignment using the BLAST algorithm (Camacho *et al.* 2009) is commonly used and depending on the study, different selection criteria are specified, such as e-value, maximum ID or length of matching sequence, number of top hits selected, etc. Caution is warranted in strictly relying on this approach, since errors in the curation of sequences in publicly available databases can propagate through the analysis and lead to misidentification of sequences. Ideally, a combination of approaches is used and when feasible the resultant species assignments should be vetted with independent data based on the known distribution and ecology of the species.

Diversity analysis. The goal of most eDNA metabarcoding studies is to accurately characterize the species richness of the community under study. Calculation of diversity indices using appropriate software allows modeling and ecological association of sequencing results. Important considerations when attempting ecological associations include appropriate data standardization to account for variations in sequencing depth and the careful selection of diversity indexes. The most common assessments include alpha-diversity (rarefaction, visualization of taxonomic profiles), and beta-diversity (Principal Components/Coordinates Analysis, NDMS ordination, etc.), prior to hypothesis testing via downstream statistical analysis.

References

Abarenkov K, Henrik Nilsson R, Larsson KH, *et al.* (2010) The UNITE database for molecular identification of fungi–recent updates and future perspectives. *New Phytologist* **186**, 281-285.

Andersen K, Bird KL, Rasmussen M, *et al.* (2012) Meta-barcoding of ‘dirt’DNA from soil reflects vertebrate biodiversity. *Molecular Ecology* **21**, 1966-1979.

Anderson-Carpenter LL, McLachlan JS, Jackson ST, *et al.* (2011) Ancient DNA from lake sediments: Bridging the gap between paleoecology and genetics. *BMC Evolutionary Biology* **11**, 30.

Andruszkiewicz EA, Starks HA, Chavez FP, *et al.* (2017) Biomonitoring of marine vertebrates in Monterey Bay using eDNA metabarcoding. *PloS one* **12**, e0176343.

Barnes MA, Turner CR (2016) The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics* **17**, 1-17.

Bazinet AL, Cummings MP (2012) A comparative evaluation of sequence classification programs. *BMC bioinformatics* **13**, 1.

Beissinger SR, McCullough DR (2002) *Population viability analysis* The University fo Chicago Press, Chicago.

Bell KL, de Vere N, Keller A, *et al.* (2016) Pollen DNA barcoding: current applications and future prospects. *Genome* **59**, 629-640.

Benson DA, Cavanaugh M, Clark K, *et al.* (2013) GenBank. *Nucleic acids research* **41**, D36-D42.

Berry D, Mahfoudh KB, Wagner M, Loy A (2011) Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and environmental microbiology* **77**, 7846-7849.

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

Biggs J, Ewald N, Valentini A, *et al.* (2015) Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). *Biological Conservation* **183**, 19-28.

Bik HM, Porazinska DL, Creer S, *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in ecology & evolution* **27**, 233-243.

Bista I, Carvalho G, Walsh K, *et al.* (2017) Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nature Communications* **8**.

Blaxter M, Mann J, Chapman T, *et al.* (2005) Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **360**, 1935-1943.

Bohan DA, Vacher C, Tamaddoni-Nezhad A, *et al.* (2017) Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. *Trends in ecology & evolution* **32**, 477-487.

Bokulich NA, Subramanian S, Faith JJ, *et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature methods* **10**, 57-59.

Bonada N, Prat N, Resh VH, Statzner B (2006) Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annual Review of Entomology* **51**, 495-523.

Bortolus A (2008) Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *AMBIO: A Journal of the Human Environment* **37**, 114-118.

Boyer F, Mercier C, Bonin A, *et al.* (2016) obitools: a unix-inspired software package for DNA metabarcoding. *Molecular ecology resources* **16**, 176-182.

Boyer S, Cruickshank RH, Wratten SD (2015) Faeces of generalist predators as 'biodiversity capsules': A new tool for biodiversity assessment in remote and inaccessible habitats. *Food Webs* **3**, 1-6.

Brown SK, Blois JL (2016) Ecological Insights from Ancient DNA. In: *eLS*. John Wiley & Sons, Ltd.

Butchart SH, Walpole M, Collen B, *et al.* (2010) Global biodiversity: indicators of recent declines. *Science* **328**, 1164-1168.

Calvignac-Spencer S, Merkel K, Kutzner N, *et al.* (2013) Carrion fly-derived DNA as a tool for comprehensive and cost-effective assessment of mammalian biodiversity. *Molecular Ecology* **22**, 915-924.

Camacho C, Coulouris G, Avagyan V, *et al.* (2009) BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1.

Caporaso JG, Kuczynski J, Stombaugh J, *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**, 335-336.

Caporaso JG, Lauber CL, Walters WA, *et al.* (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences* **108**, 4516-4522.

Cardinale BJ, Duffy JE, Gonzalez A, *et al.* (2012) Biodiversity loss and its impact on humanity. *Nature* **486**, 59-67.

Chain FJJ, Brown EA, MacIsaac HJ, Cristescu ME (2016) Metabarcoding reveals strong spatial structure and temporal turnover of zooplankton communities among marine and freshwater ports. *Diversity and Distributions* **22**, 493-504.

Chase J, Fouquier J, Zare M, *et al.* (2016) Geography and location are the primary drivers of office microbiome composition. *mSystems* **1**, e00022-00016.

Chen S, Yao H, Han J*, et al.* (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PloS one* **5**, e8613.

Civade R, Dejean T, Valentini A, *et al.* (2016) Spatial representativeness of environmental DNA metabarcoding signal for fish biodiversity assessment in a natural freshwater system. *PloS one* **11**, e0157366.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology* **21**, 1834-1847.

Cole JR, Wang Q, Cardenas E, *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research* **37**, D141-D145.

Craine JM, Barberán A, Lynch RC*, et al.* (2017) Molecular analysis of environmental plant DNA in house dust across the United States. *Aerobiologia* **33**, 71-86.

Creer S, Deiner K, Frey S*, et al.* (2016) The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution* **7**, 1008-1018.

Danielopol DL, Pospisil P, Rouch R (2000) Biodiversity in groundwater: a large-scale view. *Trends in ecology & evolution* **15**, 223-224.

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.

de Queiroz K (2005) Different species problems and their resolution. *BioEssays* **27**, 1263-1269.

De Vere N, Jones LE, Gilmore T, *et al.* (2017) Using DNA metabarcoding to investigate honey bee foraging reveals limited flower use despite high floral availability. *Scientific reports* **7**, 42838.

Deagle BE, Eveson JP, Jarman SN (2006) Quantification of damage in DNA recovered from highly degraded samples–a case study on DNA in faeces. *Frontiers in zoology* **3**, 11.

Deiner K, Altermatt F (2014) Transport distance of invertebrate environmental DNA in a natural river. *PloS one* **9**, e88786.

Deiner K, Fronhofer EA, Mächler E, Walser J-C, Altermatt F (2016) Environmental DNA reveals that rivers are conveyer belts of biodiversity information. *Nature Communications* **7**.

Deiner K, Hull JM, May B (2017a) Range-wide phylogeographic structure of the vernal pool fairy shrimp (Branchinecta lynchi). *PloS one* **12**, e0176266.

Deiner K, Renshaw MA, Li Y, *et al.* (2017b) Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods in Ecology and Evolution.* https://doi.org/10.1111/2041-210X.12836.

Deiner K, Walser J-C, Mächler E, Altermatt F (2015) Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation* **183**, 53-63.

Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW (2009) TACOA–Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC bioinformatics* **10**, 56.

Doi H, Inui R, Akamatsu Y, *et al.* (2017) Environmental DNA analysis for estimating the abundance and biomass of stream fish. *Freshwater Biology* **62**, 30-39.

Dowle EJ, Pochon X, C. Banks J, Shearer K, Wood SA (2016) Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. *Molecular ecology resources* **16**, 1240-1254.

Drummond AJ, Newcomb RD, Buckley TR, *et al.* (2015) Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience* **4**, 1.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-2200.

Egan SP, Barnes MA, Hwang CT, *et al.* (2013) Rapid invasive species detection by combining environmental DNA with light transmission spectroscopy. *Conservation Letters* **6**, 402-409.

Egan SP, Grey E, Olds B, *et al.* (2015) Rapid molecular detection of invasive species in ballast and harbor water by integrating environmental DNA and light transmission spectroscopy. *Environmental science & technology* **49**, 4113-4121.

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.

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

Erb LA, Willey LL, Johnson LM, Hines JE, Cook RP (2015) Detecting long-term population trends for an elusive reptile species. *The Journal of Wildlife Management* **79**, 1062-1071.

Eren AM, Morrison HG, Lescault PJ, *et al.* (2015) Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME journal* **9**, 968-979.

Evans NT, Li Y, Renshaw MA, *et al.* (2017) Fish community assessment with eDNA metabarcoding: effects of sampling design and bioinformatic filtering. *Canadian Journal of Fisheries and Aquatic Sciences* **74**, 1362-1374.

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.

Fahner NA, Shokralla S, Baird DJ, Hajibabaei M (2016) Large-Scale Monitoring of Plants through Environmental DNA Metabarcoding of Soil: Recovery, Resolution, and Annotation of Four DNA Markers. *PloS one* **11**, e0157505.

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

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

Foote AD, Thomsen PF, Sveegaard S, *et al.* (2012) Investigating the potential use of environmental DNA (eDNA) for genetic monitoring of marine mammals. *PloS one* **7**, e41781.

Freeland JR (2016) The importance of molecular markers and primer design when characterizing biodiversity from environmental DNA. *Genome* **60**, 358-374.

Friberg N, Bonada N, Bradley DC, *et al.* (2011) Biomonitoring of human impacts in freshwater ecosystems: the good, the bad and the ugly. In: *Advances in Ecological Research* (ed. Woodward G), pp. 1-68. Academic Press, London.

Furlan EM, Gleeson D, Hardy CM, Duncan RP (2016) A framework for estimating the sensitivity of eDNA surveys. *Molecular ecology resources* **16**, 641-654.

Gardham S, Hose GC, Stephenson S, Chariton AA (2014) DNA metabarcoding meets experimental ecotoxicology: advancing knowledge on the ecological effects of copper in freshwater ecosystems. *Advances in Ecological Research* **51**, 79-104.

Gibson J, Shokralla S, Porter TM, *et al.* (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences* **111**, 8007-8012.

Gibson JF, Shokralla S, Curry C, *et al.* (2015) Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PloS one* **10**, e0138432.

Giersch JJ, Hotaling S, Kovach RP, Jones LA, Muhlfeld CC (2017) Climate-induced glacier and snow loss imperils alpine stream insects. *Global Change Biology* **23**, 2577-2589.

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**.

Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols* **2010**, pdb. prot5368.

Gleick P (1993) *Water in crisis: a guide to the world's fresh water resources* Oxford University Press, New York, New York.

Goldberg CS, Sepulveda A, Ray A, Baumgardt J, Waits LP (2013) Environmental DNA as a new method for early detection of New Zealand mudsnails (Potamopyrgus antipodarum). *Freshwater Science* **32**, 792-800.

Goldberg CS, Turner CR, Deiner K, *et al.* (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution* **7**, 1299-1307.

Gorički Š, Stanković D, Snoj A, *et al.* (2017) Environmental DNA in subterranean biology: range extension and taxonomic implications for Proteus. *Scientific reports* **7**, 45054.

Gotelli NJ, Colwell RK (2011) Estimating species richness. *Biological diversity: frontiers in measurement and assessment* **12**, 39-54.

Guillou L, Bachar D, Audic S, *et al.* (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic acids research* **41**, D597-D604.

Haas BJ, Gevers D, Earl AM, *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research* **21**, 494-504.

Haile J, Froese DG, MacPhee RD, *et al.* (2009) Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proceedings of the National Academy of Sciences* **106**, 22352-22357.

Hajibabaei M, Shokralla S, Zhou X, Singer GA, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PloS one* **6**, e17497.

Hänfling B, Lawson Handley L, Read DS, *et al.* (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology* **25**, 3101-3119.

Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* **27**, 611-618.

Haouchar D, Haile J, McDowell MC, *et al.* (2014) Thorough assessment of DNA preservation from fossil bone and sediments excavated from a late Pleistocene–Holocene cave deposit on Kangaroo Island, South Australia. *Quaternary Science Reviews* **84**, 56-64.

Hawkins J, de Vere N, Griffith A, *et al.* (2015) Using DNA metabarcoding to identify the floral composition of honey: a new tool for investigating honey bee foraging preferences. *PloS one* **10**, e0134735.

Hebert PD, Ratnasingham S, de Waard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences* **270**, S96-S99.

Hollingsworth PM, Group CPW, Forrest LL, *et al.* (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* **106**, 12794-12797.

Hopken MW, Orning EK, Young JK, Piaggio AJ (2016) Molecular forensics in avian conservation: a DNA-based approach for identifying mammalian predators of ground-nesting birds and eggs. *BMC research notes* **9**, 1.

Hunter ME, Oyler-McCance SJ, Dorazio RM, *et al.* (2015) Environmental DNA (eDNA) Sampling Improves Occurrence and Detection Estimates of Invasive Burmese Pythons. *PloS one* **10**, e0121655.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome research* **17**, 377-386.

Jackson MC, Weyl OLF, Altermatt F, *et al.* (2016) Chapter Twelve - Recommendations for the Next Generation of Global Freshwater Biological Monitoring Tools. In: *Advances in Ecological Research* (eds. Alex J. Dumbrell RLK, Guy W), pp. 615-636. Academic Press.

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

Jo T, Murakami H, Masuda R, *et al.* (2017) Rapid degradation of longer DNA fragments enables the improved estimation of distribution and biomass using environmental DNA. *Molecular Ecology Resources.* https://doi.org/10.1111/1755-0998.12685.

Jones M, Ghoorah A, Blaxter M (2011) jMOTU and Taxonerator: Turning DNA barcode sequences into annotated operational taxonomic units. *PloS one* **6**, e19259.

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

Kao C-M, Liao H-Y, Chien C-C, *et al.* (2016) The change of microbial community from chlorinated solvent-contaminated groundwater after biostimulation using the metagenome analysis. *Journal of hazardous materials* **302**, 144-150.

Kelly RP (2016) Making environmental DNA count. *Molecular ecology resources* **16**, 10-12.

Kelly RP, Port JA, Yamahara KM, *et al.* (2014) Harnessing DNA to improve environmental management. *Science* **344**, 1455-1456.

Klymus KE, Richter CA, Chapman DC, Paukert C (2015) Quantification of eDNA shedding rates from invasive bighead carp Hypophthalmichthys nobilis and silver carp Hypophthalmichthys molitrix. *Biological Conservation* **183**, 77-84.

Kraaijeveld K, Weger LA, Ventayol García M, *et al.* (2015) Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Molecular ecology resources* **15**, 8-16.

Lacoursière-Roussel A, Dubois Y, Normandeau E, Bernatchez L (2016) Improving herpetological surveys in eastern North America using the environmental DNA method. *Genome* **59**, 991-1007.

Lacoursière-Roussel A, Côté G, Leclerc V, Bernatchez L (2016a) Quantifying relative fish abundance with eDNA: a promising tool for fisheries management. *Journal of Applied Ecology* **53**, 1148-1157.

Lacoursière-Roussel A, Rosabal M, Bernatchez L (2016b) Estimating fish abundance and biomass from eDNA concentrations: variability among capture methods and environmental conditions. *Molecular ecology resources* **16**, 1401-1414.

Leese F, Altermatt F, Bouchez A, *et al.* (2016) DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *RESEARCH IDEAS AND OUTCOMES (RIO)* **2**, e11321.

Lejzerowicz F, Esling P, Majewski W, *et al.* (2013) Ancient DNA complements microfossil record in deep-sea subsurface sediments. *Biology letters* **9**, 20130283.

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

Leray M, Yang JY, Meyer CP, *et al.* (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in zoology* **10**, 34.

Liess M, Schäfer RB, Schriever CA (2008) The footprint of pesticide stress in communities—species traits reveal community effects of toxicants. *Science of the total environment* **406**, 484-490.

Lim NK, Tay YC, Srivathsan A, *et al.* (2016) Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal Society Open Science* **3**, 160635.

Lodge DM, Simonin PW, Burgiel SW, *et al.* (2016) Risk analysis and bioeconomics of invasive species to inform policy and management. *Annual Review of Environment and Resources* **41**, 453-488.

Loman NJ, Misra RV, Dallman TJ, *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology* **30**, 434-439.

Machida RJ, Leray M, Ho S-L, Knowlton N (2017) Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data* **4**.

Mahé F, Rognes T, Quince C, De Vargas C, Dunthorn M (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593.

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

Margulies M, Egholm M, Altman WE*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Matsen FA, Kodner RB, Armbrust EV (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics* **11**, 1.

McGee KM, Eaton WD (2015) A comparison of the wet and dry season DNA-based soil invertebrate community characteristics in large patches of the bromeliad Bromelia pinguin in a primary forest in Costa Rica. *Applied Soil Ecology* **87**, 99-107.

Merkes CM, McCalla SG, Jensen NR, Gaikowski MP, Amberg JJ (2014) Persistence of DNA in carcasses, slime and avian feces may affect interpretation of environmental DNA data. *PloS one* **9**, e113346.

Minamoto T, Yamanaka H, Takahara T, Honjo MN, Kawabata Zi (2012) Surveillance of fish species composition using environmental DNA. *Limnology* **13**, 193-197.

Miya M, Sato Y, Fukunaga T*, et al.* (2015) MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Open Science* **2**, 150088.

Moyer GR, Díaz-Ferguson E, Hill JE, Shea C (2014) Assessing Environmental DNA Detection in Controlled Lentic Systems. *PloS one* **9**, e103767.

Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R (2008) Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology* **57**, 750-757.

Murray DC, Coghlan ML, Bunce M (2015) From benchtop to desktop: important considerations when designing amplicon sequencing workflows. *PloS one* **10**, e0124671.

Murray DC, Pearson SG, Fullagar R, *et al.* (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* **58**, 135-145.

Nekrutenko A, Taylor J (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics* **13**, 667-672.

Nichols RV, KOeNIGSSON H, Danell K, SPONG G (2012) Browsed twig environmental DNA: diagnostic PCR to identify ungulate species. *Molecular ecology resources* **12**, 983-989.

O'Donnell JL, Kelly RP, Lowell NC, Port JA (2016) Indexed PCR primers induce template-specific bias in large-scale DNA sequencing studies. *PloS one* **11**, e0148698.

O'Donnell JL, Kelly RP, Shelton AO, *et al.* (2017) Spatial distribution of environmental DNA in a nearshore marine habitat. *PeerJ* **5**, e3044.

Olds BP, Jerde CL, Renshaw MA, *et al.* (2016) Estimating species richness using environmental DNA. *Ecology and Evolution* **6**, 4214-4226.

Pansu J, De Danieli S, Puissant J, *et al.* (2015a) Landscape-scale distribution patterns of earthworms inferred from soil DNA. *Soil Biology and Biochemistry* **83**, 100-105.

Pansu J, Giguet-Covex C, Ficetola GF, *et al.* (2015b) Reconstructing long-term human impacts on plant communities: an ecological approach based on lake sediment DNA. *Molecular Ecology* **24**, 1485-1498.

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

Pawlowski J, Esling P, Lejzerowicz F, Cedhagen T, Wilding TA (2014) Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Molecular ecology resources* **14**, 1129-1140.

Pedersen MW, Overballe-Petersen S, Ermini L*, et al.* (2015) Ancient and modern environmental DNA. *Phil. Trans. R. Soc. B* **370**, 20130383.

Pedersen MW, Ruter A, Schweger C*, et al.* (2016) Postglacial viability and colonization in North America's ice-free corridor. *Nature* **537**, 45-49.

Pfrender M, Hawkins C, Bagley M*, et al.* (2010) Assessing macroinvertebrate biodiversity in freshwater ecosystems: advances and challenges in DNA-based approaches. *The Quarterly Review of Biology* **85**, 319-340.

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

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

Port JA, O'Donnell JL, Romero-Maraccini OC*, et al.* (2016) Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology* **25**, 527-541.

Price SJ, Eskew EA, Cecala KK, Browne RA, Dorcas ME (2012) Estimating survival of a streamside salamander: importance of temporary emigration, capture response, and location. *Hydrobiologia* **679**, 205-215.

Pruesse E, Quast C, Knittel K*, et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research* **35**, 7188-7196.

Quast C, Pruesse E, Yilmaz P, *et al.* (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research* **41**, D590-D596.

Ratnasingham S, Hebert PD (2007) BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular ecology notes* **7**, 355-364.

Rees HC, Bishop K, Middleditch DJ, *et al.* (2014a) The application of eDNA for monitoring of the Great Crested Newt in the UK. *Ecology and Evolution* **4**, 4023-4032.

Rees HC, Maddison BC, Middleditch DJ, Patmore JR, Gough KC (2014b) REVIEW: The detection of aquatic animal species using environmental DNA–a review of eDNA as a survey tool in ecology. *Journal of Applied Ecology* **51**, 1450-1459.

Renshaw MA, Olds BP, Jerde CL, McVeigh MM, Lodge DM (2015) The room temperature preservation of filtered environmental DNA samples and assimilation into a phenol–chloroform–isoamyl alcohol DNA extraction. *Molecular ecology resources* **15**, 168-176.

Rodgers TW, Janečka JE (2013) Applications and techniques for non-invasive faecal genetics research in felid conservation. *European Journal of Wildlife Research* **59**, 1-16.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584.

Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten simple rules for reproducible computational research. *PLoS Comput Biol* **9**, e1003285.

Schloss PD, Westcott SL, Ryabin T, *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75**, 7537-7541.

Schmelzle MC, Kinziger AP (2016) Using occupancy modelling to compare environmental DNA to traditional field methods for regional-scale monitoring of an endangered aquatic species. *Molecular ecology resources* **16**, 895-908.

Schmidt BR, Kery 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.

Schnell IB, Bohmann K, Gilbert MTP (2015a) Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular ecology resources* **15**, 1289-1303.

Schnell IB, Sollmann R, Calvignac-Spencer S*, et al.* (2015b) iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool–prospects, pitfalls and avenues to be developed. *Frontiers in zoology* **12**, 1.

Schnell IB, Thomsen PF, Wilkinson N*, et al.* (2012) Screening mammal biodiversity using DNA from leeches. *Current biology* **22**, R262-R263.

Shaw JL, Clarke LJ, Wedderburn SD*, et al.* (2016) Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biological Conservation* **197**, 131-138.

Shelton AO, O'Donnell JL, Samhouri JF*, et al.* (2016) A framework for inferring biological communities from environmental DNA. *Ecological Applications* **26**, 1645-1659.

Sigsgaard EE, Nielsen IB, Bach SS*, et al.* (2016) Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution* **1**, 0004.

Simmons M, Tucker A, Chadderton WL, Jerde CL, Mahon AR (2015) Active and passive environmental DNA surveillance of aquatic invasive species. *Canadian Journal of Fisheries and Aquatic Sciences* **73**, 76-83.

Sohlberg E, Bomberg M, Miettinen H, *et al.* (2015) Revealing the unexplored fungal communities in deep groundwater of crystalline bedrock fracture zones in Olkiluoto, Finland. *Frontiers in microbiology* **6**.

Somervuo P, Koskela S, Pennanen J, Henrik Nilsson R, Ovaskainen O (2016) Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics* **32**, 2920-2927.

Somervuo P, Yu DW, Xu CC, *et al.* (2017) Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods in Ecology and Evolution* **8**, 398-407.

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.

Sønstebø J, Gielly L, Brysting A, *et al.* (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular ecology resources* **10**, 1009-1018.

Spens J, Evans AR, Halfmaerten D, *et al.* (2017) Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods in Ecology and Evolution* **8**, 635-645.

Stankovic S, Kalaba P, Stankovic AR (2014) Biota as toxic metal indicators. *Environmental Chemistry Letters* **12**, 63-84.

Stoeckle MY, Soboleva L, Charlop-Powers Z (2017) Aquatic environmental DNA detects seasonal fish abundance and habitat preference in an urban estuary. *PloS one* **12**, e0175186.

Stribling JB, Pavlik KL, Holdsworth SM, Leppo EW (2008) Data quality, performance, and uncertainty in taxonomic identification for biological assessments. *Journal of the North American Benthological Society* **27**, 906-919.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012a) Environmental DNA. *Molecular Ecology* **21**, 1789-1793.

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

Taberlet P, Coissac E, Pompanon F, *et al.* (2007) Power and limitations of the chloroplast trn L (UAA) intron for plant DNA barcoding. *Nucleic acids research* **35**, e14-e14.

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

Takahara T, Minamoto T, Doi H (2013) Using environmental DNA to estimate the distribution of an invasive fish species in ponds. *PloS one* **8**, e56584.

Tang CQ, Humphreys AM, Fontaneto D, Barraclough TG (2014) Effects of phylogenetic reconstruction method on the robustness of species delimitation using single-locus data. *Methods in Ecology and Evolution* **5**, 1086-1094.

Taylor HR, Gemmell NJ (2016) Emerging Technologies to Conserve Biodiversity: Further Opportunities via Genomics. Response to Pimm et al. *Trends in ecology & evolution* **31**, 171-172.

Tedersoo L, Bahram M, Põlme S, *et al.* (2014) Global diversity and geography of soil fungi. *Science* **346**, 1256688.

Thomsen P, Kielgast J, Iversen LL, *et al.* (2012a) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology* **21**, 2565-2573.

Thomsen PF, Kielgast J, Iversen LL, *et al.* (2012b) Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PloS one* **7**, e41732.

Thomsen PF, Møller PR, Sigsgaard EE, *et al.* (2016) Environmental DNA from Seawater Samples Correlate with Trawl Catches of Subarctic, Deepwater Fishes. *PloS one* **11**, e0165252.

Thomsen PF, Willerslev E (2015) Environmental DNA–an emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation* **183**, 4-18.

Torti A, Lever MA, Jørgensen BB (2015) Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Marine genomics* **24**, 185-196.

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

Turner CR, Barnes MA, Xu CC, *et al.* (2014) Particle size distribution and optimal capture of aqueous macrobial eDNA. *Methods in Ecology and Evolution* **5**, 676-684.

Turner CR, Uy KL, Everhart RC (2015) Fish environmental DNA is more concentrated in aquatic sediments than surface water. *Biological Conservation* **183**, 93-102.

Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. *Trends in ecology & evolution* **24**, 110-117.

Valentini A, Taberlet P, Miaud C, *et al.* (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology* **25**, 929–942.

Valiere N, Taberlet P (2000) Urine collected in the field as a source of DNA for species and individual identification. *Molecular Ecology* **9**, 2150-2152.

Vörös J, Márton O, Schmidt BR, Gál JT, Jelić D (2017) Surveying Europe's Only Cave-Dwelling Chordate Species (Proteus anguinus) Using Environmental DNA. *PloS one* **12**, e0170945.

Vörösmarty CJ, McIntyre PB, Gessner MO*, et al.* (2010) Global threats to human water security and river biodiversity. *Nature* **467**, 555-561.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**, 5261-5267.

West JS, Atkins SD, Emberlin J, Fitt BD (2008) PCR to predict risk of airborne disease. *Trends in microbiology* **16**, 380-387.

Wheeler QD, Raven PH, Wilson EO (2004) Taxonomy: impediment or expedient? *Science* **303**, 285-285.

Wilcox TM, McKelvey KS, Young MK, Lowe WH, Schwartz MK (2015) Environmental DNA particle size distribution from Brook Trout (Salvelinus fontinalis). *Conservation Genetics Resources* **7**, 639-641.

Willerslev E, Cappellini E, Boomsma W*, et al.* (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* **317**, 111-114.

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

Willerslev E, Hansen AJ, Binladen J*, et al.* (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* **300**, 791-795.

Willerslev E, Hansen AJ, Christensen B, Steffensen JP, Arctander P (1999) Diversity of Holocene life forms in fossil glacier ice. *Proceedings of the National Academy of Sciences* **96**, 8017-8021.

Willoughby JR, Wijayawardena BK, Sundaram M, Swihart RK, DeWoody JA (2016) The importance of including imperfect detection models in eDNA experimental design. *Molecular ecology resources* **16**, 837-844.

Wood JR, Wilmshurst JM, Wagstaff SJ, *et al.* (2012) High-resolution coproecology: using coprolites to reconstruct the habits and habitats of New Zealand's extinct upland moa (Megalapteryx didinus). *PloS one* **7**, e40025.

Xu CC, Yen IJ, Bowman D, Turner CR (2015) Spider web DNA: a new spin on noninvasive genetics of predator and prey. *PloS one* **10**, e0142503.

Yamamoto S, Minami K, Fukaya K, *et al.* (2016) Environmental DNA as a 'Snapshot' of Fish Distribution: A Case Study of Japanese Jack Mackerel in Maizuru Bay, Sea of Japan. *PloS one* **11**, e0149786.

Yilmaz P, Kottmann R, Field D, *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature biotechnology* **29**, 415-420.

Yoccoz N, Bråthen K, Gielly L, *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology* **21**, 3647-3655.

Zaiko A, Martinez JL, Schmidt-Petersen J, *et al.* (2015) Metabarcoding approach for the ballast water surveillance–An advantageous solution or an awkward challenge? *Marine pollution bulletin* **92**, 25-34.

Table 1: Representative studies comparing richness estimates with traditional sampling or historical data for a geographic location to that of eDNA metabarcoding.

Habitat	**Macro-organism taxonomic focus**	**eDNA sample type**	**Traditional sampling method**	**eDNA efficacy finding***	**Authors**	**Year**
Air	Plants	air pollen trap	morphological identification	Better taxonomic resolution	Kraaijeveld *et al.*	2015
Freshwater	Fish	flowing water	depletion-based electro fishing	Higher diversity	Olds *et al.*	2016
Freshwater	Invertebrates	flowing water	kicknet in stream and historical data	Higher diversity	Deiner *et al.*	2016
Freshwater	Fish	stagnant water	gill-net, trapping, hydroacoustics, analysis of recreational anglers' catches	Complementary	Hänfling *et al.*	2016
Freshwater	Reptiles, amphibians	stagnant water	species distribution model based on historical data (i.e. distribution range and habitat type)	Increase species distribution knowledge	Lacoursière-Roussel *et al.*	2016
Freshwater	Amphibians, fish	stagnant water; flowing water	amphibians: visual encounter survey, mesh hand-net; Fish: electrofishing, and/or netting protocols (fyke, seine, gill)	Greater detection probability	Valentini *et al.*	2016
Freshwater	Amphibians, fish, mammals, invertebrates	stagnant water; flowing water	active dip-netting, fresh tracks or scat, electrofishing with active dip-netting	Complementary	Thomsen *et al.*	2012
Freshwater	Fish	stagnant water; flowing water; surface sediment	fyke net	Higher diversity	Shaw *et al.*	2016
Freshwater	Invertebrates	water column; surface sediment	sediment collected using a Van Veen grab	Higher diversity	Gardham *et al.*	2014

Freshwater	Fish / Diptera	Surface and bottom water column	Long-term data, electro fishing (fish) and emerging traps (Diptera) at time of eDNA sampling	Higher diversity compared to sampling but lower diversity compared to long-term data	Lim *et al.*	2017
Marine	Fish	Surface and bottom water column	Long term observation	Complementary	Yamamoto *et al.*	2017
Marine	Fish	Bottom water column	Trawl catch data	Similar Family richness	Thomsen *et al.*	2016
Marine	Fish	water column	scuba diving	Higher diversity	Port *et al.*	2015
Terrestrial	Plants	honey	melissopalynology (i.e. pollen grains retrieved from honey are identified morphologically)	Complementary	Hawkins *et al.*	2015
Terrestrial	Mammals, plants	midden pellets	historical surveys	Higher diversity	Murray *et al.*	2012
Terrestrial	Mammals	saliva	local knowledge (i.e. physical evidence) and camera data	Complementary	Hopken *et al.*	2016
Terrestrial	Birds, invertebrates, plants	top soil	invertebrates: leaf litter samples & pitfall traps; reptiles: pitfall traps and under artificial ground covers; birds: distance sampling method; plants: above-ground surveys	Complementary for plants & invertebrates	Drummond *et al.*	2015
Terrestrial	Earthworms	top soil	irrigated quadrats with 10 L of allyl isothiocyanate solution and hand collected emerging worms	Complementary	Pansu *et al.*	2015
Terrestrial	Plants	top soil	historical surveys	Complementary	Jørgensen *et al.*	2012
Terrestrial	Plants	top soil	above-ground surveys	Complementary and better taxonomic resolution	Yoccoz *et al.*	2012

Terrestrial	Vertebrates	top soil	local knowledge from safari parks, zoological gardens and farms; visual observations; historical surveys	Complementary	Andersen *et al.*	2012

* Complementary means the two survey methods detected different diversity, but does not exclude that some of the diversity was detected by both methods. Higher diversity means the study found more diversity was detected compared to conventional, but does not exclude that some of the diversity was *not* detected by both methods. Better taxonomic resolution means that sequence based identifications could be resolved to a lower taxonomic rank compared with the conventional method.

Figure legends

Figure 1: Environmental DNA sample types have different spatial and temporal scopes of inference from different habitats. Consider each sample type as a single sample from that environment. Placement of a sample type in a quadrant is not quantitive, but represents a common scale at which it has been used. Dashed arrows indicate the potential for a sample type to confer information at multiple scales of inference, but additional research to quantify these possibilities in needed.

Figure 2: Challenges for estimating abundance from environmental DNA metabarcoding. For simplicity, assume one DNA molecule depicted in the pond is equal to one organism and colors represent different species. Additionally for this example, assume that sampling is no biased (i.e., DNA copies are sampled in their true abundance), that boxes surrounding DNA molecules represent 1 uL and one DNA molecule represents 1 ng of DNA. Thus, values illustrated show the effect of primer bias, sub-sampling and their combination on the ability to estimate abundance.

Figure 3: Important guiding questions for consideration in the design and implementation phases of an environmental DNA metabarcoding study.

Figure 4: Opportunities and challenges of using environmental DNA as a tool for assessing community structure in different fields of study. The tool is reliant on a foundation (blue half circle) of continued research to improve technological aspects and continued development of DNA-based reference libraries for the identification of sequences found in the environment.

Acknowledgements

We thank Nigel G. Yoccoz Tromsø and three additional anonymous reviewers whose feedback was valuable in revising our manuscript. We thank Kristina Davis for help in drafting figures. Manuscript collaboration was facilitated by the National Science Foundation through the Coastal SEES grant to DML (EF-1427157; KD, DML, MEP), a DoD SERDP grant to DML (W912HQ-12-C-0073 (RC-2240); KD, DML, MEP); and a NOAA CSCOR grant to DML (KD, DML). National Science Foundation Research Coordination Network award to HMB (DBI-1262480) supported a related eDNA-focused Symposium at the 2016 Annual Meeting of the Ecological Society of America organized by KD, DML and MEP. In addition this article is based upon work from COST Action DNAqua-Net (CA15219; KD, FA, SC), supported by the COST (European Cooperation in Science and Technology) program, by the Swiss National Science Foundation Grant No PP00P3_150698 (to FA) and Eawag (FA and EM); the Natural Environment Research Council (NERC): NBAF pilot project grant (NBAF824 2013-14), Standard Grant PollerGEN (NE/N003756/1) and Highlight Topic Grant LOFRESH (NE/N006216/1) and the Freshwater Biological Association (FBA) (Gilson Le Cren Memorial Award 2014). IB was funded by a Knowledge Economy Skills Scholarship (KESS) a pan-Wales higher-level skills initiative led by Bangor University on behalf of the HE sector in Wales. It is part funded by the Welsh Government's European Social Fund (ESF) convergence programme for West Wales and the Valleys.

Author contributions

K.D. outlined and edited the review. All authors contributed at least one section of primary writing and contributed to editing of the manuscript. K.D., H.M.B, and E.M. synthesized sections and drafted figures.

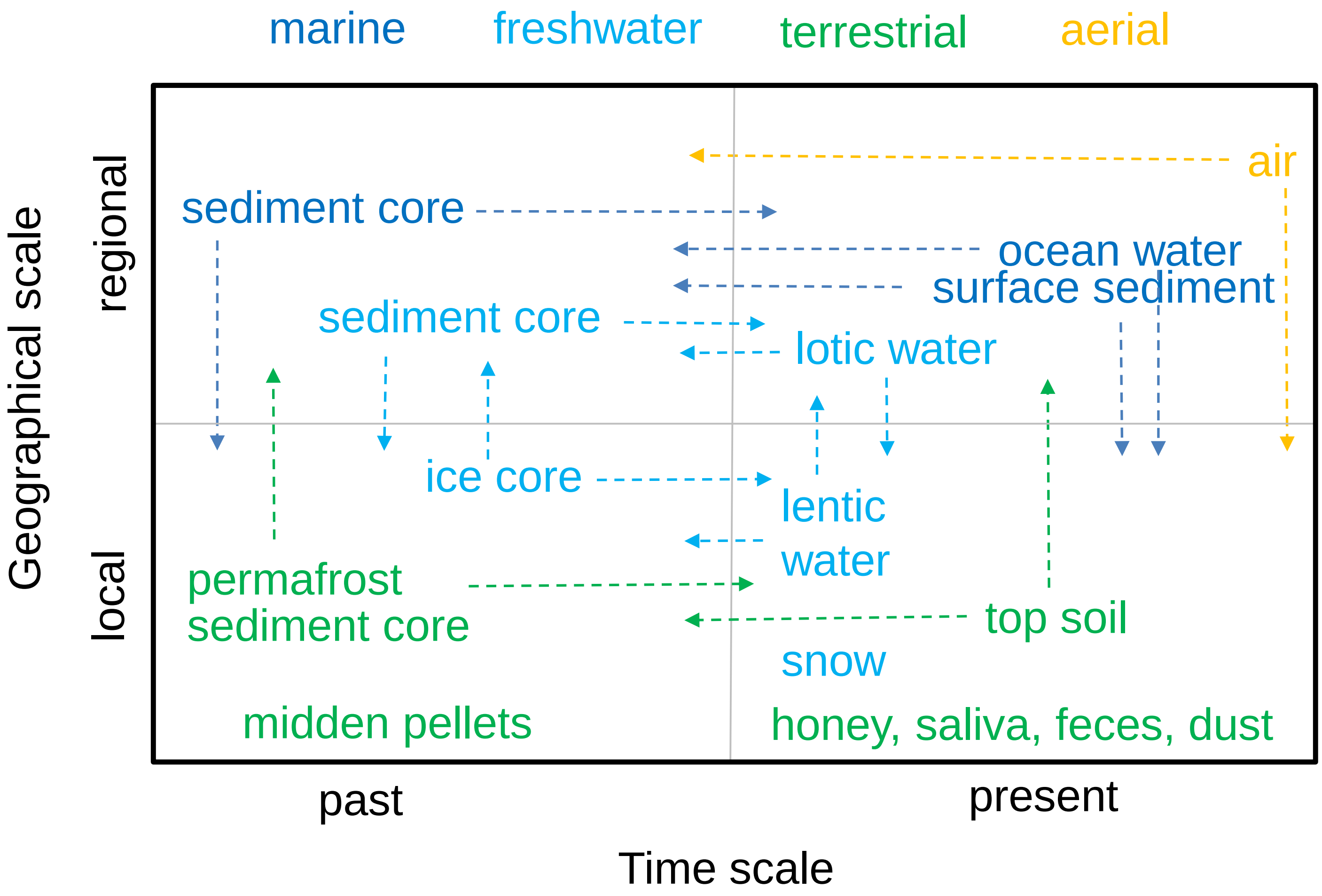
Data Accessibility

No data are associated with the manuscript

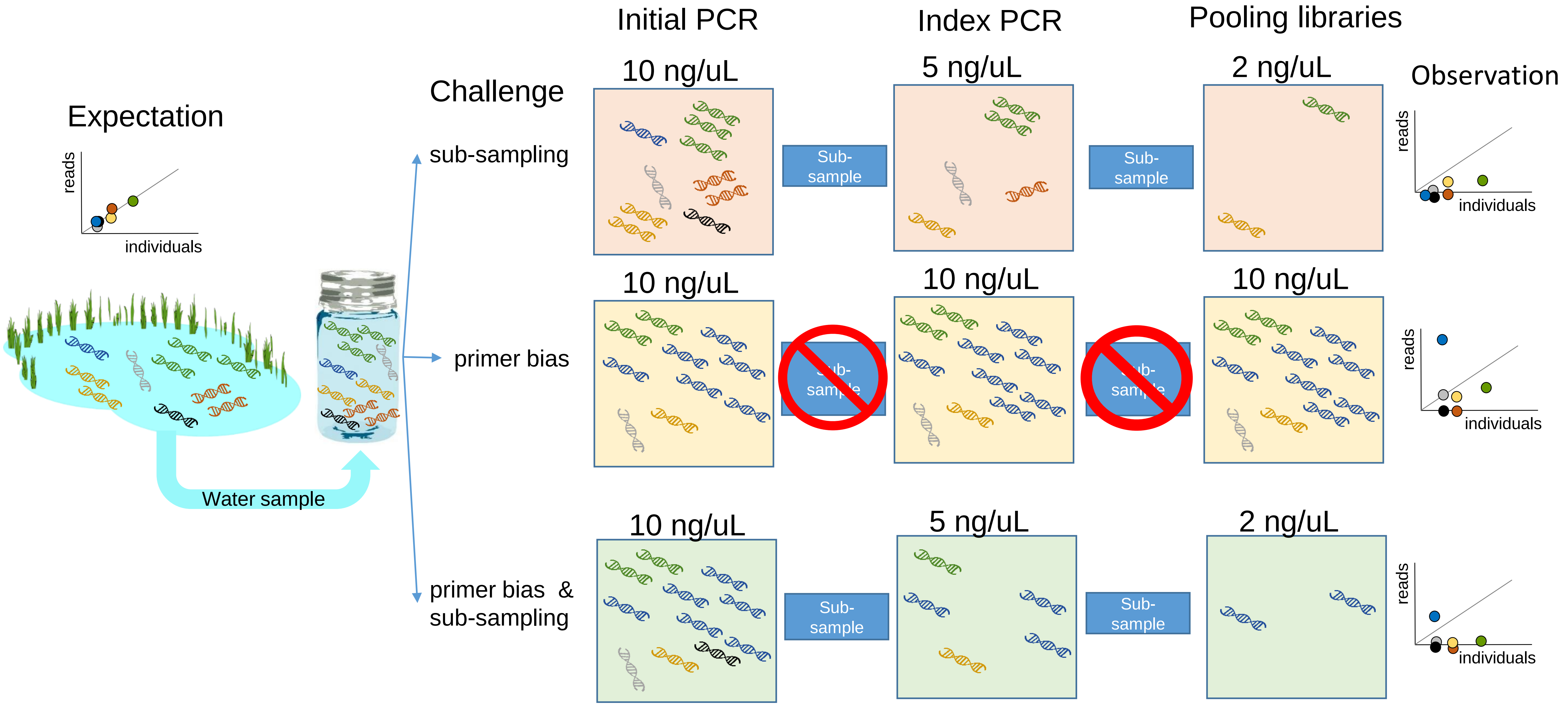
Supplemental Material

Table S1: Reviews about use of environmental DNA for species detection

Table S2: Review of primers used in eDNA metabarcoding



marine
freshwater
terrestrial
aerial
Geographical scale
regional
local
air
sediment core
ocean water
surface sediment
sediment core
lotic water
ice core
lentic
water
permafrost
sediment core
top soil
snow
midden pellets
honey, saliva, feces, dust
past
present
Time scale



Expectation
reads
individuals
Water sample
Challenge
sub-sampling
primer bias
primer bias &
sub-sampling
Initial PCR
Index PCR
Pooling libraries
10 ng/uL
Sub-sample
5 ng/uL
Sub-sample
2 ng/uL
10 ng/uL
10 ng/uL
10 ng/uL
10 ng/uL
Sub-sample
5 ng/uL
Sub-sample
2 ng/uL
Observation
reads
individuals
reads
individuals
reads
individuals

WORKFLOW

Study design

Basic science or applied? (e.g., environmental biomonitoring)

What is your study goal?

- **presence/absence**
- **diversity assessment**
- **absolute quantification**

What taxa will you target?

Is the scale of inference for your sample type appropriate to your question?

Can you compare complementary data types? (e.g. traditional vs. eDNA)

Does your sampling/ replication scheme provide good statistical power?

In the field

What type of sample is needed? (water, soil, air)

What metadata should you collect?

How many replicates will you collect?

Does your sampling protocol minimize/ control for :

- **contamination** (e.g., positive and negative controls)
- **any known biases** (e.g., inhibitors, sample volume)

In the laboratory

Sample Handling Phase

What extraction method? (physical vs. chemical)

How much sample?

What locus and primers?

Do you need to generate reference sequence data?

Are technical replicates needed?

What library preparation method will you use?

How many samples will you index and pool?

What sequence depth is needed per sample ?

What read length will you use?

DNA Processing Phase

What sequencing platform will you use?

Do you need paired end sequencing?

Have you included appropriate quality assurances? (e.g., mock community, qPCR, bioanalyzer traces)

Does your laboratory protocol minimize/ control for:

- **contamination** (e.g., positive and negative controls)
- **any known biases** (e.g., primer bias, coverage, taxonomic resolution)

At the keyboard

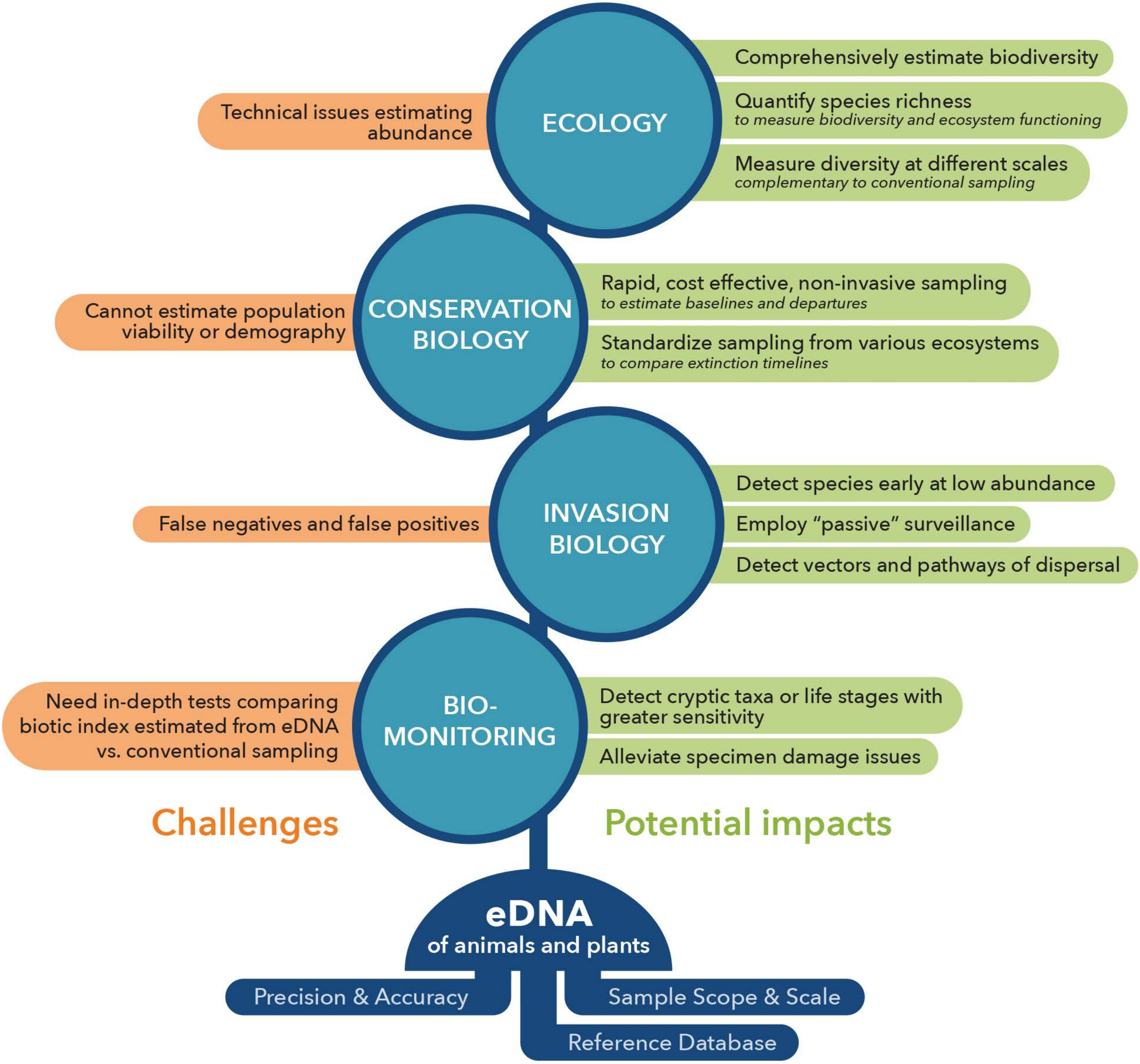
How complete is the reference database?

Do you have adequate sequencing coverage across samples?

Are you using appropri-ate choices for software tools, parameters?

Are your biological conclusions upheld using alternative parameters and workflows?

Are you including appropriate quality filtering of your data? (see Box 2)


ECOLOGY
Technical issues estimating abundance
Comprehensively estimate biodiversity
Quantify species richness
to measure biodiversity and ecosystem functioning
Measure diversity at different scales
complementary to conventional sampling
CONSERVATION BIOLOGY
Cannot estimate population viability or demography
Rapid, cost effective, non-invasive sampling
to estimate baselines and departures
Standardize sampling from various ecosystems
to compare extinction timelines
INVASION BIOLOGY
False negatives and false positives
Detect species early at low abundance
Employ "passive" surveillance
Detect vectors and pathways of dispersal
BIO-MONITORING
Need in-depth tests comparing biotic index estimated from eDNA vs. conventional sampling
Detect cryptic taxa or life stages with greater sensitivity
Alleviate specimen damage issues
Challenges
Potential impacts
eDNA
of animals and plants
Precision & Accuracy
Sample Scope & Scale
Reference Database