Ancient and modern environmental DNA

Mikkel Winther Pedersen1, Søren Overballe-Petersen1, Luca Ermini1, Clio Der Sarkissian1, James Hailé1,2, Micaela Hellstrom1, Johan Spens1,3, Philip Francis Thomsen1, Kristine Bohmann1,4, Enrico Cappellini1, Ida Baehrel Schnell1,5, Nathan A. Wales1, Christian Carøe1, Paula F. Campos1, Astrid M. Z. Schmidt1, M. Thomas P. Gilbert1, Anders J. Hansen1, Ludovic Orlando1 and Eske Willerslev1

1Centre for GeoGenetics, The Natural History Museum of Denmark, Oester Voldgade 5 – 7, Copenhagen C 1350, Denmark
2Trace and Environmental DNA Laboratory, Curtin University, Kent Street, Bentley, Perth, Western Australia 6102, Australia
3Department of Wildlife, Fish and Environmental Studies, SLU, Umeå S-901 83, Sweden
4School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK
5Center for Zoo and Wild Animal Health, Copenhagen Zoo, Frederiksberg, Denmark

DNA obtained from environmental samples such as sediments, ice or water (environmental DNA, eDNA), represents an important source of information on past and present biodiversity. It has revealed an ancient forest in Greenland, extended by several thousand years the survival dates for mainland woolly mammoth in Alaska, and pushed back the dates for spruce survival in Scandinavian ice-free refugia during the last glaciation. More recently, eDNA was used to uncover the past 50 000 years of vegetation history in the Arctic, revealing massive vegetation turnover at the Pleistocene/Holocene transition, with implications for the extinction of megafauna. Furthermore, eDNA can reflect the biodiversity of extant flora and fauna, both qualitatively and quantitatively, allowing detection of rare species. As such, trace studies of plant and vertebrate DNA in the environment have revolutionized our knowledge of biogeography. However, the approach remains marred by biases related to DNA behaviour in environmental settings, incomplete reference databases and false positive results due to contamination. We provide a review of the field.

1. Introduction

For over a decade, researchers have exploited the fact that environmental DNA (eDNA) derives not just from microbes, but from a wide range of organisms, including plants and vertebrates. A large proportion of the ancient flora and fauna do not fossilize, but leave extracellular DNA traces in the sediments. In a pioneering 2003 study, sediments from Siberia and New Zealand were found to contain traces of DNA from extinct animals, such as the woolly mammoth and moa birds [1]. The study showed that modern plant DNA could also be recovered from surface soil. The same year, another team reported the retrieval of DNA from the extinct giant ground sloth and other Pleistocene animals from a dry cave in the southwest US [2]. Since then, several studies of both past and present biodiversity have been published using eukaryotic eDNA recovered from a variety of settings including basal ice [3–5] and lake cores [6–10], surface soils [11], cave sediments [12,13], and water from lakes, streams [14–16] and oceans [17,18] (figures 1 and 2). Importantly, studies have revealed that eDNA data and other proxies such as pollen, macrofossils, living mammals and plants seem to complement each other showing wider diversity of species than using the methods separately [9–11,20–22]. Therefore, eDNA should be viewed as a complementary, rather than alternative, approach to assays of more traditional environmental proxies. Here, we discuss the experimental and bioinformatics challenges facing eDNA and provide examples of its uses for addressing biological questions.
2. Origins and behaviour of environmental DNA

The origins and behaviour of eDNA are still poorly understood. It appears that eDNA can be deposited through skin flakes [23], urine [24], faeces [25,26], eggshells [27], hair [28,29], saliva [30], insect exuviae [31], regurgitation pellets [32], feathers [33], leaves [34,35], root cap cells, in rare cases pollen [9,36], or in living prokaryotes through the secretion of plasmid and chromosomal DNA [37] (figure 1). From bacterial and plant studies, evidence exists that dead cells entering the environment may quickly be lysed with their DNA immediately being released [38]. Upon release into the environment, the DNA molecule has three possible fates.

(a) Metabolism by bacterial and fungal exonucleases

Following its release into the environment, DNA becomes vulnerable to bacterial and fungal DNases, with the former commonly believed to be the primary mechanism for extracellular DNA degradation in the environment [39].

(b) Persistence in the environment

DNA survival can be helped through the binding to environmental compounds such as clay minerals, larger organic molecules and other charged particles, which shields the adsorbed DNA from nuclease activity [40] (figure 1). Binding of nucleases also inhibits their ability to hydrolyse extracellular DNA [39]. For example, clay minerals such as Montmorillonite can absorb more than their own weight in DNA, because of their relatively large negatively charged surface area [41–44]. Furthermore, humic acids, of which some are resistant to decay, also bind DNA molecules due to a negative surface charge, and therefore prolong DNA survival. Similarly, DNA in preserved animal guts and faeces is protected from degradation by adsorption to humic acids and other organic molecules. Compared with clays, sand has been found less effective in binding DNA, the primary explanation being its small surface area. However, adsorption to sand is possible and increases with cation concentrations—particularly of divalent cations such as Ca$^{2+}$ and Mg$^{2+}$, which are most effective at forming sand–DNA bridges [45].

(c) Natural transformation

Natural transformation is a process through which cells take up extracellular DNA from the surroundings and integrate it into their own genomes [46,47]. Many bacteria are known to be agents for natural transformation, as are some archaea.

Figure 1. Environments where eDNA of plants and/or animals have been reported: basal glacier ice, terrestrial sediments, lake, rivers and lake sediments, and ocean water. The eDNA comes mainly from plant fine rootlets, faeces, urine and skin cells. The eDNA can remain in the cells, or be released from the cells in which case it may bind to inorganic particles that protect the DNA from microbial and spontaneous chemical degradation. Extracellular DNA may also be incorporated into the genomes of bacteria (bacterial natural transformation of short and degraded DNA). (a) The last may happen when extracellular DNA meets a bacterium’s surface and crosses the outer cell wall via protruding structures named pili. At the inner membrane, one strand of DNA is transported into the cell while the opposite DNA strand is degraded. (b) Once inside the cell, the DNA fragment may encounter the bacterial genome and binding at a single-stranded region during genome replication. (c) When the two new genomes segregate, one of the daughter-cells carries the inserted environmental DNA sequence.
and even a eukaryotic group of micro-invertebrates, the bdelloid rotifers [48–51]. The majority of DNA that microbes take up is quickly degraded and re-metabolized in the cell, but some DNA persists for long enough to recombine with the host genome [52]. Classical natural transformation is efficient with kilobase-long DNA, but recently it has been shown that very short DNA fragments, down to 20 bp long, remain available for integration into the bacterial genome, even when severely damaged (figure 1) [52]. Although the integration depends on similarity between bacteria and source DNA, the authors succeeded in incorporating woolly mammoth mtDNA fragments, albeit after genetically modifying the bacteria to resemble mammoth mtDNA.

In general eDNA, in particular that from ancient samples, is extremely fragmented and chemically modified with abasic sites, deaminated cytosines and cross-links [52–58]. DNA half-life is a complex function of the interplay between the physical, chemical and biological properties of the microenvironment. Turnover time of eDNA in both sea and freshwater was originally thought to be very rapid, just 6.5–25 h [59,60], but more sensitive approaches have shown survival for up to several weeks [16,17,35,61]. By contrast, in soils and sediments, moa DNA from 3000 years (kyr) old dry temperate sediments has been recovered [12], mammoth DNA dating to 30 kyr BP from permafrost sediments has been amplified, as well as 400–600 kyr old plant DNA [1] and approximately 0.5 million year old DNA from glacial basal ice [3] (figure 2).

Most eDNA studies rely on the assumption that the age of the DNA molecule recovered is the same as the age of the sediments in which it is found, but in certain conditions DNA molecules can leach through the strata and contaminate lower layers [12]. With regard to this point, DNA leaching in permanently frozen soil (permafrost) or in sediments recently frozen has not been observed [62,63]. However, in sediments from both temperate and desert environments, leaching has been reported [12,20,64] and must be taken into account as a possible concern [12,64]. In our view, DNA leaching is not the most challenging issue for proper dating of eDNA, rather it is re-deposition of sediments carrying eDNA molecules with them. Therefore, it is crucial for ancient eDNA studies to be supported by good geological profiling, providing evidence of a site’s geological stratigraphy and depositional history [65].

3. Experimental design

(a) Sampling and handling of samples for environmental DNA studies

Given the relatively low number of endogenous molecules of DNA from higher organisms in most environmental samples, contamination remains among the greatest experimental challenges to the field. Currently, several strategies for taking
eDNA samples exist for aquatic systems [16,61,66,67], lake sediments [10,13,68], permafrost soil [69,70] and ice [1,53,71–74]. The use of trace substances, such as unique plastid DNA, smeared on exposed surfaces and equipment, represents an efficient means of determining whether contaminants have penetrated inside the sample during sampling, transport, storage or subsequent subsampling [69].

For downstream analyses, samples with ancient DNA must be handled in appropriately designed laboratories divided into a pre- and post-PCR environment to reduce carry-over contamination. For ancient eDNA studies, these should be physically separated, and the former equipped with nightly UV irradiation of surfaces and positive air pressure [75,76]. Bleach and CoPA solution (a copper-bis-(phenanthroline)-sulfate/H$_2$O$_2$ solution, US patent number 5858650) is most efficient when decontaminating surfaces, gloves and equipment [77]. Other DNA decontaminating products such as RNAse away (Molecular Bioproducts) and other detergents are less effective, but in combination with UV-irradiation serve may as a non-corrosive alternative for equipment sensitive to bleach. Carry-over contamination can be limited by wearing gloves, masks and full-body suits [77]. Blank controls are crucial for identifying laboratory contamination, but are not 100% reliable, due to low levels of sporadic contamination and carrier effects [76,78]. Blank controls are likely sufficient for controlling contamination from certain species that are likely showing up only by previously produced amplicons. For other taxa, contaminants can be difficult to distinguish from endogenous DNA. For example, DNA contaminants from various sources are found in reagents [10,21,77–82]. Although most of these are from readily identifiable domesticated animals or cultivated plants, others such as Salix [83] are not and can be mistaken for genuine environmental diversity. We stress the importance of controls for each new reagent stock and systematically keeping track of these, especially now that the massive throughput of next generation sequencing (NGS) platforms makes it possible to sequence even traces of contamination. For example, commercial PCR primers were recently found to be contaminated with plant DNA (K. Andersen 2013, personal communication). Studies on eDNA using NGS technology have probably overlooked the magnitude of this problem (including our own group). Therefore, recent attempts to compile contamination databases of control sequences are extremely welcome [84].

**b) DNA extraction of environmental samples**

The high level of biological complexity in environmental samples makes unbiased extractions a major challenge. The ability to extract the DNA from samples with equal efficiencies seems unlikely, considering the wide range of sample types. Currently, no generic extraction method performs equally well across all environments or taxonomic groups [85–91]. However, numerous commercial and custom extraction protocols have been adjusted to handle different combinations of sample types and organisms. Some of these are generic and have successfully been used for eDNA studies in lakes, ancient sediments and ice [1,4,12,16,17,71,92,93], although a better understanding of extraction bias will benefit the field tremendously.

Inhibition of proteinase, DNA polymerase and DNA ligase activities can preclude eDNA analyses [94]. Several strategies have been developed to identify and overcome this problem:

(i) DNA spiking to gauge the presence of inhibitors [95,96],
(ii) DNA extract dilution to reduce inhibition, (iii) additional purification (phenol–chloroform, silica-based columns) to remove inhibitors, and (iv) incapacitating the inhibitors by using enzyme facilitators that bind lipids, phenols and other organic inhibitors such as BSA, RSA, Tween20, PEG 400 and Gp32 [94,97].

**c) Generic versus specific primers**

Metabarcoding uses generic (or universal) primers, which are designed to target several taxa simultaneously [98–102], in contrast to specific primers, which are designed to amplify only a few selected species. The advantage of using generic primers is the simultaneous amplification of a multitude of taxa and detection of new unexpected taxa. The biggest caveat when using generic primers is that the results might be skewed towards preferential amplification of certain taxa, while others (in particular rare taxa) remain undetected [9,99,101,103]. This problem results from (i) interspecific differences in decaying processes of tissue and DNA, (ii) primer-binding biases due to target sequences not matching equally well to primers [102], (iii) PCR stochasticity, and (iv) inhibition. One disadvantage of specific primers in multi-species surveys is the need for larger volumes of DNA templates, which are often in limited supply in eDNA settings. Therefore, in some cases, generic- and species-specific primers may be used in combination to maximize diversity resolution, as the two approaches may detect non-overlapping taxa [101,104]. Enrichment approaches for specific loci, possibly targeting a range of taxonomic groups simultaneously, might in the future provide a solution to such problems [105,106].

**d) Sequence-to-sample misidentification**

To increase overall data output during NGS-based analyses, eDNA can be PCR amplified using unique combinations of 5'-nucleotide-tagged primers, that enable subsequent pooling of amplicons originating from different samples [107]. Originally developed for the FLX platform, subsequent studies explored their use on Illumina platforms—although in this case problems were observed arising from tag recombination during the library amplification steps. This problem has also been observed in non-metabarcoding studies. Specifically, using Illumina sequencing and double-indexing, Kircher et al. [108] reported a significant fraction of sequencing reads with unused combinations of indexes. They identified two major causes: (i) cross-contamination of oligonucleotides carrying different indexes and (ii) chimera formation in which indexed templates from one library recombine with those from other libraries (‘jumping PCR’) in experiments where multiple sequencing libraries were amplified in bulk. Although unused index-combinations are easily identified, recombination that creates false, but already used index-combinations may introduce significant levels of sample misidentification.

There are several solutions to recognize and/or minimize sequence-to-sample misidentifications: (i) reducing the number of cycles during PCR indexing, (ii) generating a number of PCR replicates of the same sample using different combinations of 5'-nucleotide-tagged primers for each replicate and only keeping sequences consistent across a majority of PCRs (which also reduces sequencing errors), and (iii) using tags that are unique in both ends of the sequence to allow rapid identification of those not used in the study.
Even though studies have already looked into the causes and solution to jumping PCR, PCR stochasticity, and PCR-induced artefacts, their respective importance still needs to be tested to optimize how the sequencing output reflects the true diversity present in different environments.

(e) Processing next generation sequencing data and assigning sequences to taxa

Traditional genetic barcodes used in conventional (i.e. non-eDNA) projects exploit DNA barcodes of more than 500 bp in length. Barcodes of such length are inappropriate for eDNA analyses, as the eDNA is often fragmented into less than 150 bp pieces [109]. Therefore, sequence primers targeting short phylogenetically informative regions such as the trnL/rbcL genes [110,111], the 12S rRNA [112,113], 16S rRNA genes [114] and internal transcribed spacers [115,116] have been developed to survey ancient plant, animal, bacterial and fungal diversity. The EcoPrimers software [117] and the PrimerProspector package [118] have proved useful for achieving successful primer design [119–122].

Similar to the challenge of sequencing errors, single base substitutions introduced during PCR, and PCR-derived chimaera formation, can affect the taxonomic identification process. Thus, distinguishing these effects from true biological sequence variation is essential. Different denoising procedures have been developed to do this, initially based around 454/FLX pyrosequencing reads (Life Sciences, Roche), such as PyroNoise [123], Denoiser [124] and Amplicon Noise [125], and for chimaera detection, such as Uchime [126]. Procedures tailored to Illumina platforms, which are more cost-effective per base [127], have also emerged. Caporaso et al. [120] developed a 16S rRNA amplicon sequencing protocol for MiSeq and HiSeq platforms. Paired-end Illumina sequencing of 16S rRNA amplicons was compared to single-end sequencing and was found to increase the detected α-diversity of microbial communities, without affecting the resolution of phylogenetic clustering. A range of additional tools are available to help process NGS data, such as OBITools (http://www.grenoble.prabi.fr/trac/OBITools/) and QIIME [128], which can both handle data from multiple pooled samples.

With regard to taxonomic identification, one of the most popular tools for analysing metagenomic data is MEGAN [129], software that originally used BLAST to infer taxonomic composition. However, BLAST searching does not represent the most appropriate method for metagenomic sequence assignment. This is because alignments are local and not global, and hit similarities provide a measure of the confidence in the local sequence similarity but not of the validity of the assignment per se [130]. Input formats other than BLAST are now compatible with the latest version of the program (MEGAN 5), such as SAM files and QIIME output [131].

Alternative approaches based on phylogenetic placement have been developed, where databases are first screened for orthologues showing significant sequence similarity. Following sequence alignment, Bayesian phylogenetic trees are reconstructed and the query sequence assigned to the highest taxonomic level shared with all members of the smallest supported monophyletic clade to which it belongs. Posterior probability clade support is used as a direct measure of assignment significance [132]. For COI insect and trnL plant sequences, this approach was found to outperform BLAST both in sensitivity and specificity [132]. As the Bayesian framework is computationally intensive and incompatible with the size of NGS datasets, a heuristic approach has been introduced with no apparent loss in sensitivity. This approach is based on neighbour-joining trees and non-parametric bootstrapping for an evaluation of node robustness [132]. We acknowledge the fact that species absent from the database represent an important drawback of this method, as large portions of the biodiversity remain uncharacterized. Using a promising approach based on fuzzy theory and COI sequence data, Zhang et al. [133] have shown that this problem could potentially be addressed during the analyses. Despite this, building a good-quality reference sequence database, properly curated and even including taxonomically validated samples, still represents an essential component of all metabarcoding projects [102].

An important bottleneck observed in previous analyses is the necessity to align query sequences that often number in the millions, against orthologues. Aligning query sequences against a predefined template has provided an efficient solution to this problem. Fast methods based on a diversity of approaches, such as hidden Markov model profiles from the reference alignment, or phylogenetically aware strategies [134], have been proposed [135,136]. The nearest alignment space termination (NAST) procedure [137] is another such approach where the template sequence most similar to the query sequence is first identified using BLAST [138] and then pairwise realigned to the query sequence. Gap spacing originally present in the template alignment is then reintroduced in the pairwise alignment, generating a full global multi-alignment. The NAST procedure is provided with the QIIME software [120,128], which is compatible with Sanger, 454 and Illumina data and performs a full range of analyses for metabarcoding DNA sequences, including operational taxonomic unit (OTU) identification [139,140], α- and β-diversity measurements and clustering methods and UniFrac distances [141]. UniFrac distances are based on the fraction of the total branch length that is shared among samples and reflect how much environments/samples are taxonomically similar. This approach has shown promising results in assessing the microbial taxonomic proximity across environments [142–163] and also in monitoring changes in the human oral microflora following the Neolithic revolution and industrial revolution, in response to major changes in carbohydrate consumption [164]. With the growing availability of environmental metagenomic datasets, SourceTracker [165] appears to be a useful tool that can authenticate DNA profiles, for example, by showing different sources for the samples and their respective negative controls, or by matching samples with their expected tissue source [166].

With ever-reducing sequencing costs, shotgun sequencing now provides an alternative approach to metabarcoding for determining taxonomic profiles. Reads are first aligned to annotated reference genomes or clade-specific [167]/universal [168] markers, and taxon relative abundances can be estimated with appropriate normalization by genome size [169–172]. Such taxonomic profiles are not affected by biases typical of amplicon-based profiles, such as copy-number variation across taxonomic groups [173], target amplification efficiency variability [174] and single marker reliability [175].

The specificity of reference markers for shotgun profiling also limits biases related to evolutionary uninformative conserved regions and horizontal gene transfer [167,172,176,177]. Shotgun profiling is, however, hindered by computational constraints associated with the size of the datasets analysed. With the program MetaPhlAn [167], the speed of read assignment
was increased 50-fold compared with commonly employed methods such as PhymmBL [178], BLAST [138], RITA [179] and NBC [180]. The large fraction of taxa present in the environment, but not represented in databases is still problematic, as shown by analyses performed with mOTU [168], which estimated that current databases are only able to detect 43% of species abundance and 58% of richness present in clinical samples of faeces [168].

Shotgun datasets also contain comprehensive and useful information relating to the biological functions used in environmental communities [181]. By using alignment tools such as BLASTX [138], metagenomic reads are aligned to databases of proteins such as NCBI nr, KEGG [182], EGNgog [183] or SEED [184], and functional profiles can be analysed in MEGAN [129]. Finally, reference-free alternative approaches based on k-mer counts [185] have also proved to be 860 times faster than BLASTX, with comparable sensitivity and precision, but without loss of accuracy [186].

4. Environmental DNA case studies

(a) Soil, terrestrial sediments and basal ice

Soil and terrestrial sediments represent the most studied eDNA source (figure 2), and recent studies on surface sediments demonstrate that eDNA mirrors the diversity of terrestrial plants [11] and mammals [20] both qualitatively, and to some extent, quantitatively [11,20]. Ancient sediment has revealed the persistence of Late Quaternary megafauna for much longer timespans than their commonly surmised extinction times [19]. This demonstrated the power of eDNA approaches that target short molecular signatures in contrast to palaeontological analyses that require preservation of macrofossils to firmly establish the presence of a given species at a given time period.

Ancient eDNA analyses of permafrost samples distributed across the whole Arctic have provided the largest historical record of vegetation changes over the past 50 kyr [83]. Here, the authors found evidence for a diverse, but rather stable Arctic vegetation dominated by forbs until around the last glacial maximum (LGM), some 20 kyr ago where the diversity declined drastically. As the climate became warmer, a vegetation turnover was detected until the ecosystem was completely dominated by bushes and grass and depleted in forbs. Interestingly, the stomach content and faeces of Arctic megafaunal species revealed a large fraction of forbs in their diet, suggesting that the transition from a forb-dominated to a grass-dominated steppe might have contributed to the massive decline of megafaunal populations after the LGM.

In 1999, the first eDNA study was conducted on ice cores (but on microbial eukaryotes rather than higher organisms) and revealed algae and fungi diversity in the Hans Tausen ice core of southern Greenland [4]. Since then, DNA in basal ice from the DYE-3 ice core of southern Greenland revealed a diverse conifer forest with a full diversity of insects different from those found in Greenland today [3]. By dating this reconstructed ecosystem to beyond the last interglacial period (Eemian 130–115 kyr ago [187]), the authors questioned the common belief at the time, that southern central Greenland was ice-free during the Eemian. Pollen records from a marine sediment core off the south coast of Greenland further supported this claim [188] (figures 1 and 2).

(b) Marine and freshwater

Environmental DNA extracted from contemporary aquatic samples provides a good proxy of the biodiversity in and around the water (figure 2). This was first shown in freshwater ecosystems [14] with the molecular detection of the American bullfrog (Rana catesbeiana) in French wetlands. In subsequent studies, others successfully detected eDNA from invasive and low abundance species, including amphibians [16,67,189–191], fishes [15,16,192–194] and snails [195], but also from endangered amphibians, fishes, mammals and insects [16]. Furthermore, using a quantitative study design, species-specific eDNA concentrations have been found to reflect animal density [16]. The same study also demonstrated that coupling eDNA with high-throughput sequencing can account for entire lake faunas of amphibians and fishes [16], providing cost-effective approaches to monitor biodiversity.

Recently, two studies showed that seawater is also a source of macro-organismal eDNA for detection of whale species [18] and marine fish diversity [17] (figure 2). Importantly, eDNA from fresh and seawater appears to reflect contemporary rather than past diversity, as eDNA decays within a few days or weeks in the water column [16,17,61,196,197].

(c) Lake cores

Lake sediments have traditionally been used for pollen records, but have now been found to contain DNA from fishes [6], mammals [198] and plants [7–10]. This source of information was not only used to infer past human/environment interaction but also addressed a long-lasting controversy in biogeography: whether spruce survived in Scandinavia ice-free refugia during the last glaciation [8]. Two distinct mtDNA haplogroups were found in present-day Norwegian spruce, of which one is common both in and outside Scandinavia. The other is only known in Scandinavia and could represent the signature of survival in a refugium during the LGM. This was confirmed using eDNA from lake cores in areas shown to have remained ice-free during the LGM, with evidence of spruce DNA including the rare mitochondrial haplogroup.

5. Future of environmental DNA

Among the greatest benefits of eDNA is that it reduces costs and time associated with conventional bio-surveys, such as man-hours, field-training, equipment, permits, safety issues and handling of organisms. At the same time, it provides a means for undertaking large-scale biodiversity comparisons across both time and space. As such, the field of eDNA promises to revolutionize areas of archaeology, ecology and conservation [199]. The next step will be moving from metabarcoding approaches to true metagenomics. With increasing genome data being generated, this should soon be feasible and will allow for better species identifications and quantitative estimates of their abundances in environmental settings. Importantly, however, although the young field of eDNA appears to have a promising future, we emphasize that further basic studies are needed before its potential and limitations are fully explored.

Acknowledgements. The authors thank Prof. Kurt H. Kjær for help with the manuscript.

Funding statement. The Danish National Research Foundation supported this work.


