Assessing metabarcoding methods for re-analysis of pond DNA samples

Towards a new metric for assessment of pond condition

November 2023

Natural England Commissioned Report NECR496

www.gov.uk/natural-england

About Natural England

Natural England is here to secure a healthy natural environment for people to enjoy, where wildlife is protected and England's traditional landscapes are safeguarded for future generations.

Further Information

This report can be downloaded from the Natural England Access to Evidence Catalogue. For information on Natural England publications or if you require an alternative format, please contact the Natural England Enquiry Service on 0300 060 3900 or email enquiries@naturalengland.org.uk.

Copyright

Catalogue code: NECR496

Report details

Authors

Dr Georgia Ward (georgia.ward@cefas.gov.uk), Claire Robertson (clarob@ceh.ac.uk), David Ryder (david.ryder@cefas.gov.uk), Prof. David Bass (david.bass@cefas.gov.uk)

Natural England Project Manager

Harriet Knafler (Harriet.Knafler@naturalengland.org.uk)

Contractor

Centre for the Environment, Fisheries and Aquaculture Sciences (Cefas)

Keywords

Biodiversity, eDNA, metabarcoding, ponds, nanopore

Acknowledgements

We wish to thank the District Level Licensing project for collecting these DNA samples, and to Dr Daniel Read for storing and providing these samples to us.

Citation

Ward, G., Robertson, C., Ryder, D and Bass, D. 2023. Assessing metabarcoding methods for re-analysis of pond DNA samples: towards a new metric for assessment of pond condition. NECR496. Natural England.

Foreword

DNA-based methods offer a significant opportunity to change how we monitor and assess biodiversity. Natural England has been exploring the potential of these methods for environmental monitoring for several years, delivering a series of reports which focus on the development of DNA-based methods with potential in a particular area.

DNA techniques hold particular promise in freshwater monitoring and several national scale pond monitoring schemes are now collecting DNA samples. Notably, the district level licensing (DLL) scheme collects DNA samples to test for the presence of the protected *Triturus cristatus* (great crested newt). We now wish to re-visit samples collected from large scale monitoring projects and determine whether these can be re-analysed to provide more information about the condition and function of ponds across the country. Our ultimate aim is to develop novel metrics for assessing pond ecological condition, which are based on DNA data.

This project was commissioned to determine whether relatively new DNA sequencing methods can be used to provide more information from archived DLL samples. The sequencing techniques employed here were unsuccessful in generating wider biodiversity information from archived DLL samples. This is likely due to the low DNA concentrations and high levels of impurities in these samples. The results of this project will be used to inform the approach taken and methods used for a broader, collaborative project which will use DNA data to derive a new metric for pond condition. This report also offers recommendations for the collection of freshwater DNA samples to increase their suitability for further analysis in future.

Natural England commission a range of reports from external contractors to provide evidence and advice to assist us in delivering our duties. The views in this report are those of the authors and do not necessarily represent those of Natural England.

Executive summary

Background

Ponds are among the most diverse and ecologically significant freshwater habitats; however, they remain relatively understudied compared to larger freshwater ecosystems such as lakes and rivers. This is partly due to the challenges in accessing them, and the large number of individual ponds which need to be sampled in order to capture the physicochemical and biotic diversity within these environments.

DNA-based studies using targeting environmental matrices (such as water, sediment, soil and air) are increasingly being applied to aquatic environments. This includes ponds, where such studies have more frequently employed quantitative polymerase chain reaction (qPCR) assays targeted towards understanding the distribution of single species, including invasive non-native species (such as crayfish), declining fish species, and protected species such as the great crested newt *Triturus cristatus*. Other studies of aquatic environmental samples utilise amplicon-based metabarcoding methods, in which taxonomically-informative “barcoding” regions of nuclear or mitochondrial DNA are amplified by polymerase chain reaction(PCR) and sequenced using high-throughput sequencing (HTS) platforms such as Illumina, generating millions of sequences per sequencing run, which can then be bioinformatically processed and taxonomically identified against curated databases. Due to the differences in downstream approaches utilised in targeted and broader metabarcoding surveys, samples are often collected and processed using different methodologies which may limit how samples can be utilised and analysed.

One of the most successful eDNA surveys of pond habitats is the district level licensing (DLL) scheme surveys led by Natural England, which use a qPCR assay to determine the presence of DNA from the great crested newt in pond samples collected across England. Under these schemes, a large archive of eDNA samples from across England has been created, and the suitability of their use for answering wider research questions, particularly related to the biodiversity present in ponds, remains to be determined.

The aims of this study were to determine whether the quality and quantity of DNA present in the DLL samples is sufficient to allow the re-analysis of the samples using amplicon-based metabarcoding using novel sequencing methodologies suitable for generating longer, more taxonomically informative sequence data. The most significant constraint to metabarcoding using HTS platforms is the relatively short maximum read length of 500 basepairs imposed by the chemistry of Illumina sequencing platforms. However, pore-based sequencing platforms, in particular those offered by Oxford Nanopore Technologies (ONT), have no constraints on amplicon length and so are increasingly being explored as the platform of choice for metabarcoding studies in order to generate longer, more taxonomically informative sequence data from community samples. Data generated was used to assess the suitability of these samples for community analysis and involvement in future work aiming to answer wider research questions relating to pond biodiversity and condition.

Methods

The quality and quantity of DNA present in samples collected under the DLL scheme was assessed using three different assays: capillary electrophoresis to determine DNA integrity, fluorimetry to quantify double-stranded DNA, and UV-VIS spectrophotometry to determine the purity of DNA extracts. This was benchmarked against the quality of a subset of community DNA samples collected from an artificial pondscape using methodologies optimised for community DNA metabarcoding.

Next, amplification of metabarcoding regions using primer sets of varying amplicon length (700, 1100 and 4500 basepairs) was attempted, in order to determine the performance of the DLL samples using amplicon-based methods for ONT metabarcoding. This was followed by the preparation of sequencing libraries from successfully amplified samples, sequencing on the ONT MinION platform, and bioinformatic analysis of sequence data output in order to determine the suitability of the DLL samples for metabarcoding analyses.

Results

Quality control assessment of the DLL samples showed that while intact, high molecular-weight DNA was present in many samples, DNA concentration was very low, and the concentration of contaminant compounds and non-target nucleic acids was high, especially in contrast to the community DNA samples which have high DNA concentrations and few impurities present.

Attempts at amplification from the DLL samples using all three primer sets were unsuccessful using best practice PCR conditions and was only possible using a universal eukaryote primer set when using a very high number of PCR cycles. Library preparation and sequencing using an ONT MinION sequencer was attempted, however carryover of inhibitory compounds affected library preparation efficiency and greatly reduced sequencing success. Analysis of data produced showed very low diversity to be recovered from the samples. This could be as a result of the small volume sampled, the nature of sample processing and required amplification conditions, and was likely impacted by the limited success of sequencing using these methods.

Conclusions and recommendations

Due to the challenging nature of the DLL eDNA samples, amplification of metabarcoding regions using the methods described here proved incredibly difficult, most likely as a result of low DNA concentrations coupled with high concentrations of impurities and inhibitory substances. As such, attempts at metabarcoding using primers targeting longer, more taxonomically-informative gene regions and Nanopore technology showed very limited success, and very limited diversity was recovered following bioinformatic analysis. This is in part due to the nature of the extracts themselves, but consideration should be given to the influence of the collection, processing and extraction methodologies employed, and whether the dataset is therefore likely to provide a reliable representation of biodiversity present in the ponds sampled. Metabarcoding using short-read platforms (i.e. Illumina) may prove more successful in generating larger number of reads, however the phylogenetic resolution of such data is incredibly limited, and so a large number of metabarcoding regions may need to be successfully amplified and sequenced in order to gain sufficient insight into the diversity contained within the samples.

Contents

List of abbreviations

DLL: District level licensing

GCN : Great crested newt

ONT: Oxford Nanopore Technologies

PCR: Polymerase chain reaction

qPCR: Quantitative polymerase chain reaction

Background and project aims

The importance of pond ecosystems

An estimated 500 million ponds and small lakes are estimated to exist worldwide, and pondscapes – networks of interconnected ponds linked by a surrounding terrestrial matrix – are among the most diverse and ecologically significant freshwater habitats, contributing more to aquatic biodiversity than larger and more studied freshwater bodies such as rivers (Hill and others, 2018; Hill and others, 2021). Ponds are also known to sustain many rare and endangered aquatic taxa, particularly amphibians, providing important refuges, particularly in heavily modified landscapes such as agricultural, suburban or urban settings (Davies and others, 2008; Hill and others, 2021). Pond habitats support a large number of terrestrial taxa, including birds, insect pollinators, bats, and other mammals which are reliant on ponds as habitats or as a source of food or water. Additionally, ponds provide many benefits to humans (ecosystem services), such as flood alleviation, pollinator support, mitigation of the effects of climate change and reducing the effects of urban heat islands, as well as having educational and amenity value (Hill and others 2018; Hill and others 2021).

There is no globally accepted definition of a pond, however the definition most widely used in the United Kingdom and Europe states a pond is “a body of standing water 0.0025 hectares to 2 hectares in area, which usually holds water for at least four months of the year” (Williams and others, 2010). This lack of universal definition is in part due to the many different types of pond, the relative ease with which they can be created, and the wide range of environmental gradients and geological conditions in which ponds can be found, which means that ponds can show much larger variations in their abiotic and biotic conditions than larger freshwater aquatic ecosystems, even across relatively small geographic areas (Davies and others 2008).

Despite the importance of ponds as ecosystems, their value is often overlooked, and freshwater research is biased towards larger water bodies such as lakes and rivers. There have been very few studies surveying ponds across large geographic areas, with most focusing on ponds within the same pondscape or within a localised region. Similarly very few studies have been able to look at how pond ecosystems change over time, with the majority of studies focusing on a single season or year, and almost none covering timespans up to or over a decade (Hill and others, 2021). As a result, significant research gaps remain, particularly in relation to biodiversity associated with ponds and how this changes over time, and across different pond environments. Our limited understanding of pond ecosystem dynamics has wide-reaching implications, for example limiting our understanding of the effectiveness of conservation and habitat restoration strategies, as well as limiting our understanding of the impacts of anthropogenic stressors such as pollution, agricultural intensification, and urbanisation on pond environments. Ponds may also be uniquely exposed to anthropogenic stressors relating to human perception of their value, such as the removal of “pest” plant species deemed to be visually unappealing, and the introduction of other, often non-native, plant species which may alter the functioning of pond ecosystems. The introduction – both intentionally and unintentionally – of invasive species (including plants, small invertebrates and vertebrates including fish) can have detrimental effects on pond ecosystems, but in many cases are still very poorly understood (Hill and others, 2021). Importantly the extent and effect of interactions between the multiple different stressors which may act upon pond ecosystems are largely unknown and unstudied.

Environmental DNA for the assessment of pond condition

Ponds can be challenging to monitor effectively for a number of reasons. Pondscapes are often inaccessible if they are in remote or challenging landscapes, or if permissions are required because they fall on private land. In addition to this, the environmentally heterogeneous nature of ponds means that a large number of individual ponds – ideally sampled repeatedly over time - may need to be sampled in order to capture all the physicochemical and biotic diversity within a pondscape. As such, the monitoring of pond ecosystems using newly-developing technologies such as environmental DNA (eDNA) and remote sensing tools has been recognised as a key research theme for pond ecology and conservation by Hill and others (2021). Used on their own, or in combination with traditional survey methods, these tools may be used to answer important questions relating to pond ecosystem structure, ecology and condition. The use of DNA to answer questions relating to the biodiversity and distribution of species within aquatic environments is not new (Creer and others, 2016; Deiner and others, 2017), and eDNA-based surveys have been used in pond environments for the detection of invasive crayfish species (Mauvisseau and others, 2017), as well as declining fish species such as the crucian carp *Carassius carassius* (Harper and others, 2019b) and the protected great crested newt *Triturus cristatus* (Biggs and others, 2014).

eDNA is a contentious and often misapplied term, and in its strictest sense refers only to DNA extracted from environmental matrices (such as water, sediment, soil, or air) of non-organismal origin, that is, originating from shed skin cells, organelles, gametes, faeces and mucus, and decaying matter (Deiner and others, 2017; Bohmann and others, 2014; Bruce and others, 2021). In wider usage, the term also includes DNA present in an environmental matrix from organismal sources, including microbial (eukaryote and prokaryote) cells (Bruce and others, 2021; Pawlowski and others, 2020). This definition may alternatively be referred to as community DNA (Deiner and others, 2017) or total eDNA (Pawlowski and others, 2018). There are important distinctions between the organismal and non-organismal fractions of total eDNA samples, and basic considerations relating to the experimental design, sample collection and processing in eDNA surveys may have significant consequences for downstream applications and the appropriate and effective use (and re-use) of samples (Deiner and others, 2017; Bohmann and others, 2014), particularly when generating and interpreting molecular sequence data relating to biodiversity and community structure.

There has historically been very little consensus and few universally applied standards for the collection of eDNA samples. This is true of all steps of eDNA sampling, from sample collection and processing to DNA extraction (Bruce and others, 2021). eDNA samples from water are typically processed using one of two broad principal methods: either ethanol precipitation, or filtration. Ethanol precipitation methods mix water samples, typically of volumes 30 ml or less, with an ethanol solution (sometimes also including sodium or potassium acetate) to precipitate the DNA, which is then pelleted at the bottom of the collection tube to be used in downstream DNA extraction (Bruce and others, 2021; Xing and others, 2022). Precipitation-based methods have been largely superseded by filtration, in which the water sample is passed through a porous membrane, upon which organismal and non-organismal material – including DNA – is captured and preserved, ready for DNA extraction (Deiner and others, 2015; Bruce and others, 2021). Filtration-based methods are increasingly favoured as they allow for the processing of significantly larger volumes of water, therefore increasing detection rates for rarer targets compared to small-volume precipitated samples (Muha and others, 2019), and in some cases may offer improved yields and species detection even when comparable water volumes are processed (Deiner and others, 2015).

A wide range of DNA extraction methods are also used, the choice of which will have important implications for downstream analyses, and limit sample use and re-use. An important consideration is the manner in which material contained within the processed sample is lysed. For eDNA surveys largely concerned with non-organismal or animal cell targets, enzymatic (proteinase) and/or chemical lysis is often favoured, usually as part of commercially-available DNA extraction kits such as the Qiagen DNEasy Blood and Tissue kit. However, where samples are to be used for studies investigating total biodiversity and community composition (including microbes and DNA integrated into more robust matrices), more robust mechanical lysis (usually using commercial homogeniser bead tubes and homogenising instruments) is required to disrupt cell walls and structures which are not effectively disrupted by enzymatic or chemical incubation steps. In addition to selecting an appropriate lysis method, it is equally important to ensure the chosen methodology limits the carryover of inhibitory substances, both from the original sample and the reagents used in the extraction process itself (Bruce and others, 2021). This is a particularly significant consideration for pond water samples, as pond water is often likely to be stagnant and turbid, particularly compared to samples from flowing freshwater or marine sites, and therefore contain higher concentrations of inhibitory humic substances (Harper and others, 2019a).

In order for eDNA samples to be suitable for re-analysis to address new research questions, the original sample must have been collected in such a manner that as few biases as possible have been introduced into the sample by the chosen methodologies. This is not straightforward, as it can be difficult to anticipate future requirements, particularly in fields such as molecular ecology, which can evolve rapidly as new technologies develop.

Metabarcoding technologies

Metabarcoding technologies, in which a standardised “barcode” region of DNA (usually a relatively short section of the ribosomal RNA (rRNA) operon or mitochondrial genome of eukaryotes) is amplified by PCR using broadly targeted primers and sequenced using massively parallel, high-throughput sequencing (HTS) technologies, is widely used to explore the diversity associated with a range of sample types. This includes community DNA (total eDNA) samples, plant and animal tissues, and biological samples including blood and stool (Compson and others, 2020). The most widely used sequencing platforms for metabarcoding are those developed by Illumina, whose MiSeq platform allows the sequencing of millions of reads across multiple samples on a single run, for amplicons up to 500 basepairs in length. Many protocols exist for metabarcoding using Illumina platforms – including MiSeq, HiSeq and NextSeq – and many attempts have been made, with limited success, to standardise the primers used for amplification of metabarcoding regions across different taxonomic groups. There are a large number of sequencing library preparation options compatible with Illumina platforms, including those offered by third-party companies. Additionally a number of custom library preparation workflows, in which amplification and individual indexing of samples (to allow assignation of sequenced reads to samples on multiplexed sequencing runs) is performed in a single step, have been developed, including the widely-adopted prokaryote 16S metabarcoding protocol of Kozich and others (2013), later adapted by Minardi and others (2022) to target the eukaryote small subunit rRNA V4 hypervariable region in two companion primer sets, one targeting eukaryotes broadly and the other biased against metazoans and therefore suitable for the amplification of microbial taxa from animal tissues.

The greatest constraint on the use of Illumina platforms for metabarcoding studies remains the limitations on sequencing read length (and therefore barcoding amplicon length). In turn this can limit primer choice, as ideally the primers used should have as broad a taxonomic coverage as required, as well producing an amplicon of a length compatible with the chosen sequencing platform, which is currently limited to 500 basepairs.

Third generation, such as those developed by Pacific Biosciences (PacBio; now part of Illumina) and the pore-based sequencing platforms of Oxford Nanopore Technologies (ONT) – effectively have no constraints on read length. In contrast to the sequence-by-synthesis technology utilised by Illumina platforms, where fluorescence-labelled nucleotides are incorporated into newly synthesised DNA strands matching the library template and imaged to generate sequence data, pore-based technologies generate sequence data by passing DNA strands through membrane-embedded pores, generating a characteristic electric current dependent on the nucleotide occupying the pore. Pore-based technologies to not rely on a finite pool of reagents for sequencing, and so are capable of rapidly sequencing much longer DNA molecules, up to tens of kilobases in length. Of the two pore-based sequencing platforms, PacBio offers greatest read accuracy (a measure of the number of errors introduced during sequencing) – comparable to that offered by Illumina platforms – however both the sequencers and reagents required for PacBio are considerably more expensive than other platforms, particularly ONT. ONT read accuracy is lower than both Illumina and PacBio (but rapidly improving as new reagent and flow cell kits are released), but ONT sequencers are affordable and readily available to many molecular biology labs, and are small and adaptable to a wide range of applications, particularly metabarcoding.

It is therefore unsurprising that pore-based ONT platforms are increasingly utilised for metabarcoding projects. The lack of read length constraint allows much longer amplicons to be sequenced. This not only gives end users more freedom over choice of primer, but also gives the option of producing longer, more taxonomically informative sequence data, allowing phylogenetic placement of reads generated from metabarcoding studies.

The potential value of large, archived DNA datasets

Large-scale environmental sampling is expensive, particularly when collecting material for eDNA studies. Standardisation across datasets is also desirable, as is cost-effectiveness, and so decisions are often made on what is the most appropriate methodology for the immediate research question. “Sample once, test many” is a relatively new concept, in which samples are collected in a way which enables them to be used repeatedly to answer multiple research questions through the application of various methodologies, and archived in a format suitable for re-analysis with as-yet undeveloped technologies. Unfortunately the concept is difficult and often impossible to apply retrospectively, especially given the speed at which technologies – particularly new molecular sample collection, extraction and sequencing technologies – develop.

Despite this, the existence of large, archived DNA datasets available for re-analysis to answer new research questions is a prospect worthy of exploration. In particular such datasets may prove invaluable for the generation of data which cannot be collected from new samples – such as data relating to the past distribution of species – or samples collected in locations which are difficult to access, especially if the samples themselves are accompanied by well-curated metadata.

One such dataset is the large collection of eDNA samples collected and archived under Natural England’s district level licensing (DLL) schemes for the detection of the great crested newt *Triturus cristatus* using a targeted quantitative polymerase chain reaction (qPCR) assay (see Appendix 1 for links to published data files, up to March 2023). The samples were collected and processed using a standardised protocol, developed and optimised by Biggs and others (2014), and includes samples collected from ponds across England, Scotland and Wales. Each sample is of a standardised water volume (90 ml), and DNA extracted using a reproducible, commercially available kit, using appropriate controls to monitor for contamination. Metadata were also collected at each site, relating to habitat suitability, pond permanence, water quality, and the presence of other taxa including fish, birds, and aquatic plants. Once used for their intended purpose, the eDNA samples are archived, and as such may represent an important dataset for reanalysis. Of particular interest to this study is establishing whether this archive is an appropriate dataset for reanalysis using modern and developing metabarcoding technologies, aimed at exploring biodiversity and community structure of pond sites across the UK.

Aims and objectives

This project aims to determine the suitability of archived eDNA samples collected for the District Level Licensing (DLL) scheme to monitor for the presence of the great crested newt (GCN) for re-analysis using novel metabarcoding technologies, to generate high-resolution taxonomic data using longer reads than is currently possible using Illumina short-read HTS platforms. The quality and quantity of DNA present in a subset of 120 samples from the wider dataset is determined, and amplification attempted using three broadly-targeted primer sets suitable for sequencing using Oxford Nanopore Technology's MinION platform. The quality and amplification success of these samples was benchmarked against a set of community DNA samples – including water and sediment samples – collected specifically for the purpose of metabarcoding, to establish how well small-volume eDNA samples perform in comparison to larger-volume community DNA extracts.

Successfully amplified DLL samples were prepared for sequencing, and sequencing attempted using a MinION platform. The resulting sequence data was processed and genotypes generated taxonomically assigned against curated eukaryote sequence databases to determine the diversity present in the sample. Based on the quality and performance of the samples in amplification and sequencing attempts, recommendations are made on the suitability of these samples, and by implication the wider sample set, for inclusion in future metabarcoding and community analysis projects.

Materials and methods

Sample background

A total of 120 DLL samples collected across England were included in this study. Of these, 90 have previously been successful in attempts to amplify at least one short (approximately 300 basepair) metabarcoding region for Illumina sequencing. A further 30 samples could not be successfully amplified using Illumina-targeted primer sets. The generation of high-quality metabarcoding data is reliant on good quality DNA extractions which are free of inhibitors, and which have been collected and processed in such a matter as to reduce (as far as is practicable) the introduction of biases into downstream datasets. Therefore, for the assessment of quality of the DNA samples collected as part of the Great Crested Newt eDNA District Level Licensing (DLL) survey, a further dataset of 30 water or sediment samples collected from a freshwater pondscape (Pinkhill Meadows) using methods specifically optimised for community metabarcoding were also included in the quality assessment and PCR optimisation stages, to allow a benchmark of quality and suitability of the DLL samples for metabarcoding-based methodologies.

DLL sample processing and DNA extraction

The DLL DNA samples were collected in accordance with the protocol developed by Biggs and others (2014), briefly summarised here. 20 water samples of 30 mL volume were collected from around the margin of a single pond and pooled into a single sample in a sterile plastic bag. The bag was then sealed and the sample homogenised by gentle shaking to ensure eDNA is evenly dispersed throughout the sample. Next, 6 subsamples of 15 mL were taken from the bag and mixed in sterile tubes with 35 mL of a molecular-grade ethanol solution to preserve eDNA. Samples were then kept refrigerated as far as possible until further processing.

For each sample, all 6 replicate subsamples were centrifuged at 14,000 g for 30 minutes at 6 °C and the supernatant (liquid comprising water from the pond and preservative ethanol) was discarded, leaving behind pelleted precipitated DNA and other substances. 360 µl of sample lysis buffer ATL (Qiagen, USA) was added to the first replicate tube and the tube vortexed (mechanically shaken vigorously) to resuspend pelleted material. This liquid was then transferred to the second tube and vortexing repeated. This process was repeated until the liquid had been transferred through all six sub-samples. After this, the lysis buffer containing the resuspended pellet from all 6 samples was transferred to a 2 mL tube, and DNA extracted using a DNEasy Blood and Tissue kit (Qiagen, USA) using the standard protocol for mammalian tissue. The DNEasy Blood and Tissue kit uses enzymatic lysis of samples using Proteinase K to degrade protein matrices and cell structures, followed by treatment with a guanidine hydrochloride buffer to deactivate proteins, before passing the samples through silica-based columns and ethanol precipitation to isolate DNA from cell debris. Samples were eluted into a final volume of 200 µl elution buffer (Tris-EDTA buffer, pH 7.4).

Community DNA sample processing and DNA extraction

For the community DNA samples collected at Pinkhill Meadows, water samples of approximately 1000 ml volume were collected and transported to the lab for immediate processing by filtering. All samples were pre-filtered under pressure using a vacuum pump through 12 µm cellulose nitrate membranes (AE100; Whatman, USA), before further filtering under pressure through a 0.45 µm cellulose nitrate membrane (7141-114). Filter papers were removed from the filter housing and placed in 5 mL centrifuge tubes and stored at 4 °C until extraction. The filter housings were sterilised by acid washing or soaking in 10% hypochlorite bleach between samples. Whole sediment samples were collected in sterile centrifuge tubes and stored frozen until DNA extraction.DNA was extracted from filter papers using the DNEasy PowerWater DNA extraction kit (Qiagen, USA), using the standard kit protocol and eluted into 100 µl elution buffer EB.Sediment samples were processed by defrosting the whole sediment sample and subsampling 0.25 g material, using a microbalance, sampling boat and disinfected forceps. DNA was extracted from subsamples using a DNEasy PowerSoil DNA extraction kit (Qiagen USA) using the standard kit protocol and eluted into 100 µl elution buffer EB.

Blank extractions (which followed the same extraction protocol but included no sample) were carried out alongside water and sediment samples to act as a control for contamination.

Assessment of sample quality

Of the DLL samples, a representative subset of 30 which were amplified successfully using at least one short-read (approximately 250 basepair) metabarcoding primer set were assessed for quality, with a further 30 samples which could not be successfully amplified using the same approaches. Alongside these, 30 community DNA (total eDNA) samples from Pinkhill Meadows, all of which were successfully amplified and sequenced using Illumina technologies, were also assessed using the same tests. Each of these samples (90 in total) was subject to three targeted assays to determine the integrity, quantity and purity of DNA present in the sample.

DNA integrity (the average fragment size of double-stranded DNA present in the extracts, measured in basepairs) was determined using an Agilent TapeStation 4500 with genomic DNA reagents (Agilent Technologies, USA). The TapeStation uses capillary electrophoresis against molecular marker standards of known weight to assign molecular weight to double stranded DNA present within a small volume of sample.

Next, the quantity of double-stranded DNA in 1 µl of each extract was determined using a QuantiFlour ONE dsDNA fluorometric assay kit on a GloMax Explorer plate reader (Promega, USA).

Finally, UV-Vis spectrophotometry was carried out on 1 µl each sample on a NanoDrop ONE microvolume UV spectrophotometer (ThermoFisher Scientific, USA), using the

Qiagen EB elution buffer (used by the DNEasy Blood and Tissue, DNEasy PowerWater and DNEasy PowerSoil extraction kits used for the DLL, Pinkhill Meadows water, and sediment samples respectively) as an assay blank. This assay was conducted to determine the presence and extend of non-target nucleic acids and other inhibitive molecules within the DNA extracts.

Primer choice and PCR optimisation

All samples were tested using three different primer sets of differing amplicon length and target region in order to determine the feasibility of their use in metabarcoding studies using long-read, pore-based technologies. Generation of amplicons for metabarcoding studies typically uses normalised DNA templates of known concentration, in order to produce a standardised dataset and reproducible methodologies (Bruce and others, 2021; Tedersoo and others, 2022). However, the concentration of DNA present in the DLL samples is too low in most cases to allow for this, so 5 µl extracted DNA was used as a template for each PCR reaction. A subset of 30 DLL samples (the same 30 used for assessment of sample quality) were also diluted 1:10 in Tris-EDTA (pH 7.5) for use as a PCR template, to determine whether inhibitory substances could be effectively diluted out to improve amplification, as suggested by Wilson (1997). For the Pinkhill Meadows samples, DNAs were normalised to a final concentration of 5 ng/µl in Tris-EDTA (pH 7.5) for use as PCR templates. For each primer set, reactions were carried out in duplicate and products for each sample pooled prior to gel electrophoresis. All primer sequences are shown in Table 1. All primers were used as constructs comprising universal sequence tails to facilitate downstream Nanopore barcoding without the need for prior adapter ligation, without significantly affecting the specificity or efficiency of the PCR reaction. For each primer in Table 1, the sequence of this tail is shown underlined.

Optimisation of each primer set was carried out using the Pinkhill Meadows samples, and was attempted using various high-fidelity, proof-reading polymerases designed for amplification of longer target regions, including DreamTaq DNA polymerase (ThermoFisher Scientific, USA), Q5 high-fidelity DNA polymerase, LongAmp Taq DNA polymerase (both New England Biolabs, USA), and PrimeStar GXL Premix (Takara Biosciences, Japan), in order to identify the most suitable polymerase for robust and specific amplification of the target amplicon with minimal non-specific amplification. Proofreading polymerases are enzymes used to synthesise DNA strands in PCR which include domains which allow the enzyme to check the correct nucleotide has incorporated into the strand being synthesised and remove incorrectly incorporated nucleotides. This function is important for metabarcoding applications as it greatly improves the accuracy of amplicon generation during PCR amplification. For each primer set, the polymerase which showed the most consistent and robust amplification was used to attempt amplification of the ribosomal RNA from all Pinkhill Meadows and DLL samples tested.

Table 1. Sequences of primers with Nanopore-specific tails used for amplification of metabarcoding target regions

Assay target	Primer name and sequence [note 1]	Primer orientation	Reference
Eukaryote ribosomal RNA array	3NDf: TTTCTGTTGGTGCTGATATTGCGGCAAGTCTGGTGCCAG	Forward	Jamy and others (2021)
Eukaryote ribosomal RNA array	21R: ACTTGCCTGTCGCTCTATCTTCGACGAGGCATTTGGCTACCTT	Reverse	Jamy and others (2021)
Plant internal transcribed spacers	ITSp5: TTTCTGTTGGTGCTGATATTGCCCTTATCAYTTAGAGGAAGGAG	Forward	Cheng and others (2016)
Plant internal transcribed spacers	ITSu4: ACTTGCCTGTCGCTCTATCTTCRGTTTCTTTTCCTCCGCTTA	Reverse	Cheng and others (2016)
Eukaryote small subunit V4 hypervariable region	Hug574*f: TTTCTGTTGGTGCTGATATTGCCGGTAAYTCCAGCTCYV	Forward	Hugerth and others (2015)
Eukaryote small subunit V4 hypervariable region	1132r: ACTTGCCTGTCGCTCTATCTTCCCGTCAATTHCTTYAART	Reverse	Hugerth and others (2015)

Note 1: Underlined sections indicate the Nanopore-specific tails used to facilitate PCR barcoding using the PCR Barcoding Kit (Oxford Nanopore Technologies, UK)

Eukaryote ribosomal RNA array primers

The first primer set tested amplifies an approximately 4500bp fragment of the ribosomal RNA array of eukaryotes, covering the small subunit (18S) gene, the internal transcribed spacer (ITS) 1 and 2 regions, and part of the large subunit (28S) gene. Primers 3NDf (Cavalier-Smith and others, 2009) and 21R (Schwelm and others, 2016), used in this study, were shown by Jamy and others (2021) to have an *in silico* taxon coverage matching against 91.5% and 88.1% of eukaryote sequences in the SILVA database (release 132) (Quast and others, 2013) respectively.

PCRs were carried out in 25µl volumes comprising 1X Takara PrimeStar GXL Premix (a preformulated mastermix containing reaction buffer, dNTPs and high-fidelity polymerase, designed to be used in difficult reactions and for templates with GC-rich regions), final concentration of 0.4µM each primer, and either 1µl normalised template (Pinkhill Meadows samples) or 5 µl genomic DNA or 5µl diluted DNA (DLL samples). The reaction volume was made up to 25 µl using molecular grade water. PCR reactions were conducted in a VeritiPro thermal cycler (Applied Biosystems, USA) with the heated lid set to 110 °C. Reaction conditions consisted of an initial denaturation at 98 °C for 10 seconds, followed by 30 cycles of denaturation at 98 °C for 10 seconds, annealing at 55 °C for 15 seconds, and extension at 68 °C for 4 minutes. This was followed by a final extension step at 68 °C for ten minutes, and sample storage at 4 °C until electrophoresis on 1% agarose-TBE gels stained with GreenSafe (NZYTech, Portugal) and visualised using a UV transilluminator.

Plant internal transcribed spacer primers

The second primer set trialled targeted the ITS regions of vascular plants, using forward primer ITSp5 and reverse primer ITSu4 (both from Cheng and others, 2016). *In silico* and *in vitro* analyses presented by Cheng and others (2016) demonstrate these primers are suitable for the amplification of 95% of plants in most groups. This primer set produces an amplicon between 700 and 1100 basepairs in length.

PCRs were carried out in 25µl volumes comprising 1X Q5 High-Fidelity Mastermix, final concentration of 0.5 µM each primer, and either 1µl normalised template (Pinkhill Meadows samples), or 5 µl genomic DNA or 5µl diluted DNA (DLL samples). The reaction volume was made up to 25 µl using molecular grade water. PCR reactions were conducted in a VeritiPro thermal cycler (Applied Biosystems, USA) with the heated lid set to 110 °C. Reaction conditions consisted of an initial denaturation at 98 °C for 30 seconds, followed by 30 cycles of denaturation at 98 °C for 10 seconds, annealing at 55 °C for 30 seconds, and extension at 72 °C for 30 seconds. This was followed by a final extension step at 72 °C for two minutes, and sample storage at 4 °C until electrophoresis on 2% agarose-TBE gels stained with GreenSafe (NZYTech, Portugal) and visualised using a UV transilluminator.

Eukaryote small subunit V4 primers

The final primer set used was a pair of universal eukaryote primers targeting an approximately 700 basepair fragment of the small subunit gene, spanning the V4 hypervariable region frequently used for microeukaryote metabarcoding. This approach, using primers Hug574*f and 1132r (Hugerth and others, 2014) was shown *in silico* by

Hugerth and others (2014) and Minardi and others (2022) to amplify a broad range of microbial eukaryotes and metazoan taxa PCRs were carried out using the same reaction composition and cycling conditions as for the plant ITS primer set above.

As very little amplification was observed for the DLL samples using any primer set tested, PCRs using the eukaryote V4-targeted primer set were repeated using the undiluted DLL samples only as above, but using a total of 40 amplification cycles. The volume of template DNA added for the DLL samples was not increased as quality assessment showed the presence of large amounts of potentially inhibitory substances which could interfere with PCR amplification. Total cycle number for amplicon generation was not raised above a maximum of 40 in order to reduce the amplification of artefacts and chimeras.

Library preparation and sequencing

48 DLL samples which successfully amplified using the universal eukaryote V4-targeted primer set were taken forward to attempt Nanopore sequencing. Amplicons were cleaned using AMPure XP sample purification beads (Beckman Coulter, USA) using 0.8X reaction volumes using the manufacturer-recommended protocol, and eluted into 25 µl molecular-grade water.

1µl volumes of each cleaned amplicon were quantified using a QuantiFlour ONE dsDNA fluorometric assay kit on a GloMax Explorer plate reader (Promega, USA), and 200 fmol of amplicon per sample normalised to a volume of 24 µl in PCR water ready for individual barcoding using the PCR Barcoding Expansion Pack 1 – 96 (Oxford Nanopore Technologies, UK). Because amplification primers included the barcoding adapter, the adapter ligation step was omitted from the library preparation protocol. Barcoding reactions consisted of 24 µl normalised amplicon DNA, 25 µl LongAmp Taq 2X mastermix (New England Biolabs, USA), and 1 µl PCR Barcode (Oxford Nanopore Technologies, UK). Barcoding reactions were conducted in a VeritiPro thermal cycler using the following cycling conditions: initial denaturation at 95 °C for 3 minutes, followed by 15 cycles of denaturation at 95 °C for 15 seconds, annealing at 62 °C for 15 seconds, and extension at 65 °C for 1 minute. This was followed by a final extension at 65 °C for 2 minutes.

Barcoded amplicons were then purified using 40 µl AMPure XP sample preparation beads (0.8X reaction volume) as above, before elution into 21 µl molecular-grade water. 1 µl of each amplicon was then quantified using a QuantiFlour ONE dsDNA fluorometric assay kit on a GloMax Explorer plate reader, and 21 ng of DNA from each sample pooled into a single tube to prepare a 1 µg pool of barcoded libraries for adapter ligation and sequencing.

The barcoded amplicon pool was then subject to end repair using the NEBNext Ultra II End Repair/dA-tailing Module (New England Biolabs, USA). Reaction mixtures comprised 49µl pooled, barcoded DNA, 1 µl control DNA sample, 7 µl Ultra II End-prep Reaction Buffer, and 3 µl Ultra II End-prep Enzyme Mix. The reaction mixture was thoroughly mixed by pipetting the whole volume multiple times, before incubation at 20 °C for 5 minutes and then 65 °C for 5 minutes in a thermal cycler. Next, the reaction mixture was purified using 60 µl Ampure XP sample purification beads as above, and eluted into 61µl molecular-grade water into an Eppendorf Lo-Bind tube (Eppendorf, Germany). Adapters were then ligated onto amplicons using Adapter Mix F supplied with the Ligation Sequencing Kit (Oxford Nanopore Technologies, UK) and NEBNext Quick T4 Ligase (New England Biolabs). The ligation mixture comprised 60 µl end-repaired DNA from the previous step, 25 µl ligation buffer LNB, 10 µl NEBNext Quick T4 DNA Ligase, and 5 µl Adapter Mix F. After incubation, 40 µl AMPure XP sample preparation beads were added to the reaction and incubated at room temperature for 5 minutes before being placed onto a magnetic rack. After beads were pelleted, the supernatant was removed, and 250 µl short fragment buffer added. The tube was removed from the magnet and flicked to resuspend the beads, before being returned to the magnet to re-pellet the beads. The supernatant was then removed by pipette and discarded, and this wash step repeated once more. The pellet was then dried for approximately 30 seconds before resuspension into 15 µl EB (Oxford Nanopore Technologies, UK). The resuspended DNA was then incubated for ten minutes at room temperature away from the magnet rack, before being returned to the magnet and 15 µl resuspended DNA transferred to a clean 1.5 mL Eppendorf DNA Lo-Bind tube. The final barcoded, adapted library pool was quantified using the QuantiFlour ONE dsDNA kit on a Quantus fluorometer (Promega, USA), and run on an Agilent Tapestation 4500 using D5000 ScreenTape and reagents (for analysing DNA fragments between 100 and 500 basepairs in length) to determine library size and ensure no unligated adapters were carried over into the final pool.

Sequencing was carried out using a R9.4.1 flow cell on a MinION Mk1C (Oxford Nanopore Technologies, UK). Prior to sequencing, a flow cell check was performed on the flow cell to ensure active pores were present. The flow cell was then primed for sequencing as per the standard Oxford Nanopore Technologies. 40 fmol of library was prepared for sequencing using the Ligation Sequencing Kit (Oxford Nanopore Technologies) and immediately loaded onto the flow cell. Sequencing was programmed to continue for 36 hours with barcode balancing switched off, and both basecalling and barcoding switched on. Data was output in FAST5 and FASTQ formats.

The first sequencing run underperformed, with a much lower proportion of flow cell pores active during sequencing than expected. The pooled library was requantified using the QuantiFlour ONE dsDNA kit, and a Qubit dsDNA BR kit (Invitrogen, USA), a similar fluorometric-based assay for double-stranded DNA quantification, using a Qubit 4 fluorometer (Invitrogen, USA). The same value was obtained from both assays, and was very close to the original quantification value (small fluctuations in DNA quantity would be expected due to degradation of the library pool in cold storage). Two further sequencing runs were attempted, using the same run preparation protocol described above, using 20 fmol library and 80 fmol library (both normalised to a volume of 12.5 µl for loading library preparation), however the performance of these runs was roughly equivalent to the first and so could not be improved. The data output from all three runs was combined for downstream data analysis.

Data processing and taxonomic assignment

Sequencing data was basecalled using Super High Accuracy algorithms using the Oxford Nanopore Technologies Guppy software (version 6.4.6) with the dna r9.4.1 450bps model. Next, sequences containing both the forward and reverse amplification primers were identified, and primer sequences trimmed from each read using Cutadapt version 3.4 (Martin, 2011), before consensus sequences were constructed from the data using NGSpeciesID (Sahlin and others, 2021) version 0.1.3, with minimum quality set to 15, profile for clustering transcript reads enabled, consensus polishing set to spora and detect reverse compliments enabled. The maximum number of sequences to form a consensus with spoa was set to 10,000, and the minimum to 20, with the minimum to identify a reverse complemented consensus set to 0.9. The aligned threshold and mapped threshold were both set to 0.8.

Trimmed reads from each sample were aligned against a corresponding set of consensus sequences using kma 1.4.9 (Clauson and others, 2018), with the second version of the conclave algorithm, altered indel calling for ONT data, the preset for ‘3rd gen genefinding’ and one query to one template mapping options all enabled. The minimum support to call bases was set to 0.7, with the minimum overall and base depth both set to 20, the minimum relative alignment score set to 0.80, the minimum query coverage set to 0.9 and the minimum and maximum read trimming parameters set to 500 and 800bp, respectively.

Following initial alignment of reads against each set of consensus sequences, the polished consensus sequences output by kma for each sample were combined into a single file and renamed with the prefix ‘OTU’. The merged set of consensus sequences were then processed using the locate function from seqkit 2.3.0 (Shen and others, 2016) to identify soft masked regions at the ends of each sequence, and the coordinates of any such regions were saved in the bed format. Bedtools v2.30.0 (Quinlan and others, 2010) was then used to identify the remaining portion of each sequence, after excluding terminal, soft masked regions, and save it to a separate file. cd-hit-est version 4.8.1) (Fu and others, 2012) was used to cluster the trimmed, polished consensus sequences, with local sequence identity option enabled and cd-hit configured to cluster sequences into the most similar cluster meeting a 95% identity threshold, a word length of 10bp, and the alignment coverage required to be at least 90% of the longer sequence.

Following trimming and polishing of consensus sequences, the alignment step using kma, as described above, was repeated, with the option to output additional features enabled.

The R statistical programming language version 4.1.2 (R Core Team, 2021) was used to import alignment results from each sample and reshape the results into an OTU table for downstream data processing.

Taxonomy was assigned using the assignTaxonomy function of the DADA2 package (Callaghan and others, 2016) in RStudio version 2022.12.0-353, against the PR2 (Protist Ribosomal Reference) database version 4.14.0, a curated database of ribosomal RNA sequences from eukaryotes, including animals, fungi, plants and protozoa (Guillou and others, 2013). The assign taxonomy function, which uses a naïve Bayesian classifier (based on that of Wang and others, 2007) to assign taxonomy to sequences at taxonomic ranks between kingdom and genus, producing a confidence bootstrap value for each rank. The bootstrap cutoff was set to 80% (sequences showing less than 80% similarity to sequences in the reference database were left unassigned at the taxonomic rank at which a sufficiently high percentage match could not be found). Data were then visualised in RStudio using the tidyverse (Wickham and others, 2019) and phyloseq (McMurdie and Holmes, 2013) and packages.

Results

Sample quality and implications

90 samples (60 from the DLL sample set and a further 30 from the Pinkhill Meadows dataset) were subject to three assays to determine the integrity, quantity and purity of DNA present in the extract.

DNA integrity assessment – conducted here using an Agilent TapeStation 4500 - gives a measure of the degradation of DNA in a sample, most often by determining the molecular weight of double-stranded DNA present. Samples with high DNA integrity will contain a good quantity of high molecular weight DNA, with little to no single-stranded or degraded DNA present in the sample, while samples containing DNA which is degraded, damaged or modified will have low integrity, which can negatively impact downstream applications, particularly PCR and sequencing library construction (Gassman and McHoull, 2016). DNA integrity values are shown in summary for each sample set in Table 2 below as average DNA fragment size, and in detail for each sample in Appendix 2 and 3 for the DLL and Pinkhill Meadows sample sets respectively.

The quantity of double-stranded DNA in each sample was determined using a fluorometric assay. Fluorometric assays use targeted fluorescent probes which bind only to double-stranded DNA and excite at known wavelengths to allow highly accurate determination of double-stranded DNA concentration. Such assays target only double-stranded DNA, and do not react to single-stranded DNA, RNA or free nucleotides in solution, which can all lead to overinflation of quantitation values using other methods, and so are significantly more sensitive and specific than either capillary electrophoresis or UV spectrophotometric methods, particularly for low-concentration samples which contain contaminants. The assay used can accurately quantify double-stranded DNA present in 1 µl of sample within a sensitivity range of between 0.2 – 400 ng/µl.

UV-Vis spectrophotometry assessment of each sample was used to determine the presence and extent of contaminants in the samples, measured by calculating the absorbance at three wavelengths (230 nm, 260 nm, and 280 nm), and calculating A260/280 and A260/230 purity ratios to indicate the presence of contaminant proteins and chemicals in sample extracts which may inhibit downstream assays, particularly PCR. As noted above, though this assay also provided estimates of nucleic acid concentration, this value was not used as a measure of double-stranded DNA concentration, as it cannot discriminate between high quality double-stranded DNA and other nucleic acids such as single-stranded DNA, RNA and free nucleotides. A summary of the values from the three quality control assays carried out on the 60 DLL samples is shown in Table 2 and detailed in Appendix 2. This subset of 60 samples included 30 which were previously successfully amplified using primer sets targeting short (less than 300 basepair) metabarcoding regions, and 30 which could not be successfully amplified using these primer sets.

Table 2. A summary of values recorded for the three quality assessment assays run on the DLL and Pinkhill Meadows DNA extracts, including measure of DNA integrity (TapeStation fragment size), DNA quantity (double-stranded DNA), and sample purity (nucleic acids, A260/A280 ratio and A260/A230 ratio). Values in brackets indicate the standard deviation in each measurement.

Sample set	Average DNA fragment size (basepairs)	Double-stranded DNA (ng/µl)	Nucleic acids (ng/µl) [note 1]	A260/ A280 ratio	A260/ A230 ratio
DLL (amplifiable using short-read primers)	37861 (SD 17794)	2.63 (SD 2.21)	75.41 (SD 70.57)	1.59 (SD 0.18)	0.58 (SD 0.17)
DLL (not amplifiable using short-read primers)	39895 (SD 16889)	1.94 (SD 1.83)	27.29 (SD 20.34)	1.87 (SD 0.27)	0.34 (SD 0.18)
Pinkhill Meadows water	10329 (SD 1346)	25.56 (SD 11.98)	26.73 (SD 19.75)	1.86 (SD 0.07)	1.31 (SD 0.58)
Pinkhill Meadows sediment	10856 (SD 2598)	51.19 (SD 34.39)	46.87 (SD 34.46)	1.80 (SD 0.23)	1.47 (SD 0.65)

Note 1: the value for nucleic acids given here includes all nucleic acids present in the sample, including double-stranded DNA and contaminant single-stranded DNA, RNA and free nucleotides, and cannot be used as an accurate quantification of good quality, amplifiable double-stranded DNA.

Assessment of DNA integrity showed that the majority of DLL samples contained high molecular weight DNA. This was true of samples which had previously amplified using short-read metabarcoding primers, and those which could not be successfully amplified. For samples where DNA integrity could be measured, the average fragment size was 37861 basepairs, demonstrating that where double-stranded DNA is present is highly intact. The concentration of double-stranded DNA was generally very low, in many cases lower than 2 ng/µl, and often less than 1 ng/µl. The value for nucleic acid concentration from the UV-Vis spectrophotometry assay is much higher. This likely indicates that non-target nucleic acids are present at high quantities in the DLL extracts. Two other values are given from the UV-Vis spectrophotometric assay: the A260/A280 ratio, and the A260/A230 ratio. Both values give a measure of the purity of nucleic acids in a sample, and are the ratio of the absorbance value of a sample at a wavelength of 260nm (the wavelength at which all nucleic acids absorb) to the absorbance at wavelength 280 nm and 230nm, respectively. For pure DNA samples, the A260/A280 ratio should be approximately 1.8. Values lower than this can indicate the presence of contaminant proteins and other chemicals which absorb at a wavelength of 280nm in the sample. The pH of the sample can also influence this value: values for basic solutions can be over-represented by up to 0.2, and those for acidic solutions under-represented by the same fraction. The A260/A230 value for pure nucleic acid samples should be between 2.0 and 2.2. Values lower than this indicate the presence of contaminant carbohydrates and chemicals including guanidine hydrochloride, commonly used in commercial column-based DNA extraction kits, including the DNEasy Blood and Tissue kit used to extract the DLL samples. For all the DLL samples, the A260/A230 ratios are much lower than expected for pure nucleic acids, indicating the significant carryover of contaminant substances, potentially from both the original sample and the extraction process.

For comparison, the values for the three quality assessment assays carried out on the Pinkhill Meadows community DNA samples are also summarised in Table 2, and detailed for each sample in Appendix 3. Overall the average fragment size is lower, for both water (average 10329 basepairs) and sediment (10856 basepairs) extracts, however the values for the concentration of double-stranded DNA are considerably higher (average 25.56 ng/µl and 51.19 ng/µl for water and sediment extracts, respectively). It is also noteworthy that the values for DNA quantification from the fluorometric assay are very similar to those from the UV-Vis spectrophotometric assay, indicating that very little carryover and contamination of the samples with other nucleic acids and substances absorbing at the same wavelength as double-stranded DNA occurred. The values for the A260/A280 ratio are generally higher than for the DLL samples, and closer to the expected value of 1.8 in many cases. Most notably the A260/A230 ratio, which is a strong indicator of the carryover of contaminant carbohydrates and extraction reagents, is much higher for most – though not all – Pinkhill Meadows samples.

The most significant difference between the sample sets is the concentration of double-stranded DNA in the sample. This large disparity is most likely due to the difference in the volume of water sampled in each case – a total of 90 ml was subsampled and extracted for the DLL samples, compared to a total of 1 litre for water samples from Pinkhill Meadows. Best practice for the generation of metabarcoding amplicon data usually encourages the normalisation of DNA templates to a standardised concentration prior to use as a template for PCR (Bruce and others, 2021; Tedersoo and others, 2022), by diluting the DNA to a known concentration in a specified volume. This not only ensures more uniform amplification of samples across the same reaction composition and cycling conditions, but also serves to dilute contaminants present in DNA extracts before addition to PCR reactions, and therefore reducing the effects of these contaminants as PCR inhibitors. Due to the very low concentration of DNA present in the DLL samples, it was not possible to effectively normalise the samples or dilute out the inhibitory effect of contaminants, and so the amount of contaminant substances entering the PCR reactions could not be minimised or controlled effectively.

PCR amplification success

Three primer sets were tested in this study. The first of these targeted an approximately 4500 basepair fragment of the ribosomal RNA array of eukaryotes, from upstream of the hypervariable V4 region of the small subunit (18S) gene, spanning the internal transcribed spacer (ITS) 1 and 2 regions, and including approximately 1000 basepairs of the large subunit (28S) gene. This fragment encompasses several regions which are typically used for metabarcoding of different groups of eukaryotes: the small subunit V4 and V9 hypervariable regions, which are typically used for metabarcoding of microbial eukaryotes; both ITS regions, frequently used for plant and fungal metabarcoding, and regions of the 28S used for identification of some plant, microalgal and parasite groups. Generation of sequence data covering a large (approximately 1200bp) portion of the small subunit also allows this data to be used for the identification of metazoa, including invertebrates and vertebrates.

The second prim set trialled targeted the internal transcribed spacer (ITS) regions of vascular plants, producing an amplicon between 700 and 1100 basepairs in length. , This amplicon spans from the end of the small subunit gene to the start of the large subunit gene, therefore encompassing both taxonomically-informative internal transcribed spacer regions and the 5.8S gene.

The final primer set used was a pair of universal eukaryote primers targeting an approximately 700 basepair fragment of the small subunit gene, spanning the V4 hypervariable region frequently used for microeukaryote metabarcoding. This approach, using primers Hug574*f and 1132r (Hugerth and others, 2014) was shown *in silico* by Hugerth and others (2014) and Minardi and others (2022) to amplify a broad range of microbial eukaryotes and metazoan taxa. This primer set has also previously been successfully used for metabarcoding using the Illumina MiSeq platform by Minardi and others (2022), however the long amplicon length prevented reads from being paired during downstream analysis. These primers were also used successfully to sequence small-volume filtered water samples collected from 120 urban pond samples across England using the library preparation protocol detailed below using the Oxford Nanopore Technologies MinION platform (manuscript in preparation).

No amplification was observed in any PCR reaction across all three primer sets using the diluted DNA template from the DLL samples. This is most likely due to the very low quantity of DNA present in the template. Appendix 4 details the amplification success of each of the three assayed primer sets across all DLL samples using undiluted templates, summarised below in Table 4. PCRs targeting a 4500 basepair fragment of the ribosomal RNA gene of eukaryotes was the least successful using the DLL samples, with only a single sample of the 120 DNA extracts tested yielding an amplicon (0.83%). Success using the two shorter amplicon approaches, targeting the ITS regions of plants (approximately 1100 basepairs) and the eukaryote small subunit V4 hypervariable region (approximately 700 basepairs) were more successful, with eight samples out of 120 amplifying in each case (6.67%), though it was not the same samples amplifying across the two assays. When the PCR targeting the V4 hypervariable region was repeated increasing the number of amplification cycles to 40, 49 of 120 samples amplified (40.83%).

In contrast, eukaryote ribosomal RNA PCRs were successful on 16 of 18 normalised water samples collected at Pinkhill Meadows, and 9 of 12 normalised sediment extracts from this pondscape (88.89% and 75.0% respectively). The plant primers amplified 12 of 18 water samples (66.67%) and 9 of 12 sediment samples (75%), while the V4 hypervariable region primers amplified 15 of 18 water samples (83.33%) and 7 of 12 sediments (58.33%). Appendix 2 details amplification success across each of the Pinkhill Meadows community DNA samples included in this study.

Table 4. Summary of amplification success across the three different primer sets assayed using the undiluted DLL DNA samples and normalised Pinkhill Meadows community DNAs as a PCR template. This includes both configurations in which the eukaryote 18S V4 targeted primers were used, with amplification using 30 and 40 cycles separately tested.

Sample set	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles	Eukaryote small subunit V4 (700 basepairs), 40 cycles
DLL (undiluted)	1/120 (0.83%)	8/120 (6.67%)	8/120 (6.67%)	49/120 (40.83%)
Pinkhill Meadows water (normalised)	16/18 (88.89%)	12/18 (66.67%)	15/18 (83.33%)	Not attempted
Pinkhill Meadows sediment (normalised)	9/12 (75%)	9/12 (75%)	7/12 (58.33%)	Not attempted

The DLL sample set included in this study comprised 90 extracts which had previously amplified using alternative short-read metabarcoding primer sets, and 30 which had not successfully been amplified during previous attempts. No samples from this subset of 30 were amplifiable using the primer strategies included in this study. Of the previously amplifiable samples, 41 could not be amplified using any primer set tested here. The subset assayed for this study comprises only a fraction of the total dataset, and indeed only a subset of the total number of samples from which a previous study made attempts at amplifying short-read metabarcoding regions. That study comprised a total of 719 DNA extracts from DLL surveys, with only 90 (12.52%) proving successful at amplification. That study also trialled three primer sets – targeting the eukaryote small subunit V9 hypervariable region (approximately 200 basepairs), the second internal transcribed spacer region (ITS2) of plants (250 – 280 basepairs) and the first internal transcribed spacer region (ITS1) of fungi (approximately 230 basepairs). Further attempts at amplifying a 280 basepair region of the prokaryote 16S gene frequently used for metabarcoding were unsuccessful on this larger dataset. This is of particular concern, as bacteria are present in all environments, and in particular would be expected to be abundant in pond water samples, and so this may indicate strong taxonomic biases introduced by the sampling methods. Shorter amplicons are generally easier to amplify by PCR, even in samples with low double-stranded DNA concentration and in the presence of inhibitors, however the manner in which samples are collected and processed – particularly the omission of any mechanical lysis step to disrupt cell walls and other more robust matrices, as is the case in the DLL dataset – can influence downstream success at amplifying different taxonomic groups, particularly those which form more robust cellular structures and would not be considered to be a component of an eDNA sample in the strictest definition of the term.

It should also be noted that Biggs and others (2014) report in the development of the sample collection and great crested newt qPCR assay utilised by the DLL surveys that inhibition was observed in a high proportion of samples (approximately 11.3%). In these cases, this could sometimes be overcome by diluting the samples, however attempts presented here demonstrate this was not effective for the generation of metabarcoding amplicons. Quantitative PCR assays utilise reagents which favour sensitivity over high accuracy and proof-reading ability (which is not required for such targeted assays), and reagent formulations developed to overcome the presence of inhibitors are widely used, and indeed were utilised by Biggs and others (2014) in the development of their assay. Due to differences in downstream data interpretation, qPCR assays are also less constrained in the number of amplification cycles which can be used. For the great crested newt assay applied to this sample set, a total of 55 amplification cycles was used, which is almost double that which is considered the maximum best practice for use in metabarcoding studies (Lindahl and others, 2013; Bruce and others, 2021; Tedersoo and others, 2022).

Sequencing and data output

Three sequencing runs on separate MinION flow cells were attempted, all of which underperformed. A likely explanation for this is that inefficient barcoding and adapter ligation reactions led to a large proportion of the library pool remaining unbarcoded (and therefore unadapted and not able to be sequenced), resulting in overinflation of library quantification values when measured using fluorometric assays. Quality assessment of a subset of cleaned amplicons (Appendix 3) prior to the barcoding PCR step show significant carryover of nucleic acids and contaminants, which may inhibit the PCR reaction in which barcodes are added to amplicons, or the following ligase-mediated adapter ligation steps. The LongAmp polymerase recommended by Oxford Nanopore Technologies for the barcoding reaction was used in preliminary tests to optimise amplification of barcode regions and performed poorly compared to the two polymerases eventually used for these assays (Takara PrimeStar GXL for the eukaryote ribosomal RNA amplicons, and NEBNext Q5 Mastermix for the plant ITS and eukaryote V4 hypervariable region amplicons), using both the Pinkhill Meadows and DLL samples. It may be possible to swap the LongAmp polymerase for another proof-reading polymerase, however this would need to be empirically tested – using high-quality samples – which would require considerable time and investment. Similarly alternative methods to amplicon cleanup prior to library preparation could be explored, but are likely to be similarly costly, both in time and investment.

Usable sequence data were recovered for 25 out of 48 samples included across the sequencing runs. The average number of reads per sample was 311.8, which is much lower than is ideal for robust metabarcoding community analysis studies. 17 Operational Taxonomic Units (OTUs) – consensus sequences generated from clusters of sequences with 99% similarity – were generated across the dataset, and taxonomy assigned to each sequence against the Protist Ribosomal Reference (PR2) database. Figure 1 shows the distribution of these OTUs across the samples. Each bar represents a different sample, and each shade of grey represents a different genus (taxonomic group). The y axis shows the proportion of reads from each sample assigned to each genus. All but two of the samples are dominated by sequences classified as belonging to the Branchiopoda, an abundant class of crustaceans. Most of these reads were assigned with 100% confidence to the genera *Daphnia* and *Simocephalus* – both water fleas – but a small proportion were assigned to the genus *Chydorus*, a freshwater crustacean of similar size common in freshwater environments. Other small, zooplanktonic crustaceans are also represented in the dataset, with a small proportion of reads in a single sample (GCN000345) assigned to the class Ostracoda (seed shrimp), with 100% similarity to the genus *Heterocypris,* a widespread and abundant genus. The majority of other sequence diversity within the sample represent the ciliate families Dysteriidae, Gastrotricha, Halteriidae, Pelagostrombidiidae and Strobildiidae. Ciliates are abundant and diverse protozoa in most aquatic environments, including ponds (Andrushchysyn and others, 2003), and are also known to have extremely high ribosomal RNA copy numbers (Gong and others, 2013) and so are often over-represented in eukaryote metabarcoding datasets. Finally, a small proportion of reads in a single sample were assigned to the genus *Cryptomonas*, a common freshwater alga in temperate environments.

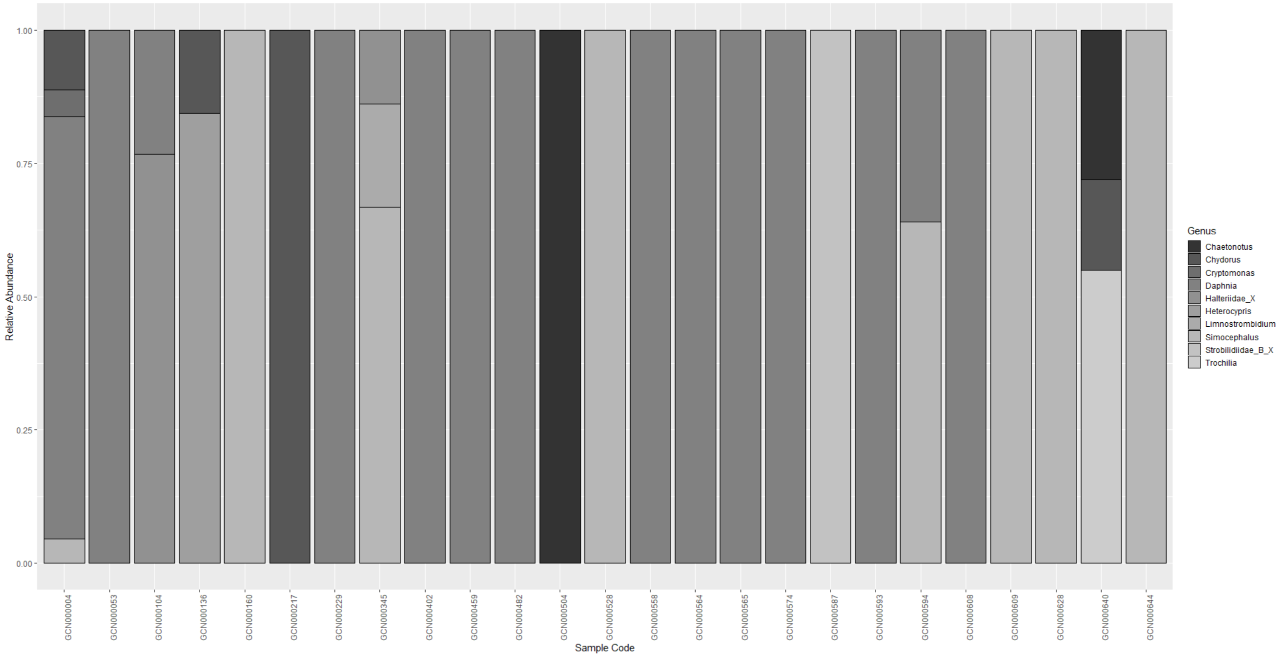

Figure 1. Taxonomic summary of reads recovered from 25 DLL samples using primers targeting the eukaryote small subunit V4 hypervariable region.

Diversity across the sample set is very low, and many samples did not produce any useable reads. To an extent this is due to the limitations imposed by the poor performance of the sequencing runs, itself likely an artefact of poor sample quality. The taxonomic groups recovered are all known to be very abundant in freshwater samples, and are also groups which are unlikely to require robust mechanical lysis to disrupt cell structures. As the methodologies used for the collection, processing and extraction of the DLL samples were focused on the recovery of eDNA and not community DNA (=total eDNA), the majority of the community DNA fraction of the samples would have been selectively removed during sample processing, resulting in extracts biased towards extracellular eDNA and easily lysed cells and tissues.

Conclusions and recommendations

The work presented here demonstrates that while high molecular weight DNA is present in the majority of the assayed DNA samples collected under the district level licensing schemes great crested newt surveys, the quantity of good quality, amplifiable double-stranded DNA is limited, and the use of these extracts as template material for metabarcoding studies is further hindered by the presence of significant amounts of inhibitory substances present in the extracts. Taken together with the findings of other attempts, aiming to amplify very short metabarcoding regions from a much larger pool of samples, it is evident that success is much lower than would be desired, with only approximately 12% of over 700 samples screened amplifying with any primer set.

Improved amplification was observed using an increased number of PCR cycles with one of the primer sets tested here, targeting the eukaryote small subunit V4 hypervariable region, though success was still limited (only 41% of 120 samples could be amplified). It may be that a larger proportion of samples from the full dataset could be amplified if a higher number of cycles is used, however caution must be taken as increasing cycle number beyond the generally-recognised maximum advisable number of 30 (Bruce and others, 2021; Tedersoo and others, 2022) is likely to result in over-representation of certain taxa, as well as increasing the formation of chimeric reads.

Further optimisation of the library preparation procedure for Oxford Nanopore Technologies, or further post-amplification cleanup to remove inhibitors, may improve the performance of MinION sequencing. It may also be the case that sequencing on Illumina platforms – where pooled sequencing libraries are typically diluted up to one thousand-fold prior to loading onto the sequencer – may be more successful, however it remains to be determined whether sequencing library preparation using Illumina protocols prove more successful.

Finally, it should be considered whether the nature of the samples themselves makes them suitable for metabarcoding approaches, regardless of the sequencing platform used. The methodologies used to collect and process the samples have been shown by Biggs and others (2014) to be adequate for the original intended purpose – that is, amplification by quantitative PCR of DNA from the great crested newt – however the volumes of water sampled are very small at only 90 ml, and the processing and extraction methodologies used may be suitable only for targeted eDNA analysis, and not for answering wider questions relating to biodiversity and community fine-scale, species-level structure (Bruce and others, 2021). For such research questions, DNA extracts collected and processed using methods suitable for extracting community DNA – i.e. DNA contained in living, unlysed cells present in the sample, DNA contained within complex matrices, as well as extracellular DNA – are required, and these require vastly different processing methodologies to eliminate bias as far as is possible in the resultant extracts. While the large archive of DNA extracts amassed under the DLL great crested newt surveys has been shown to likely be of very limited use for reanalysis using Nanopore metabarcoding methodologies, without extensive troubleshooting, the dataset may still represent a hugely valuable resource for other targeted eDNA or conventional PCR assays, particularly those targeting single species, as new research priorities and species of interest emerge.

Recommendations

- The use of a number of short (c. 100 - 200 basepair) metabarcoding markers targeting different taxonomic groups, sequenced using Illumina platforms, may be considered for use on archived DLL samples. However, the difficulties encountered in attempting to amplify prokaryotes – among the most abundant taxa in most water samples – suggest that significant taxonomic biases exist in these samples, and so data should be interpreted with this in mind

- The archived samples have proven suitable for targeted detection of the great crested newt using specific qPCR assays, and so it may be that other targeted quantitative or conventional PCR assays of the sample set may offer insight into the presence or absence of other taxa, including invasive amphibians and invertebrates. However, as above, caution must be taken when interpreting results – particularly negative results – due to the high levels of inhibitors in many DLL samples
- In order to expand the suitability of samples collected under the DLL scheme for wider analyses, including metabarcoding, the methodologies utilised should be reviewed in light of recent advances. Water filtration has emerged as preferable to ethanol precipitation for capturing extracellular and community (total) eDNA in a single sample (Bruce and others, 2021), and the adoption of DNA extraction methods using optimised, commercially available kits (such as the kits used for the Pinkhill Meadows samples), which utilise mechanical sample disruption and therefore introduce fewer biases into the final extract.
- Filtration offers several advantages over ethanol precipitation, the most immediate of which is significantly increasing the volume of water which can be sampled. Filtering larger volumes of water allows for the detection of rarer taxa which may not be recovered from small-volume water samples. Filtration may use either manual or vacuum-assisted equipment. Manual filtration is more feasible *in situ* in the field, however vacuum-assisted filter manifolds allow for filtration of bulk (multiple litre) sample filtration
- A key consideration for water filtration is the pore size of the membrane used. Smaller pore sizes (typically 0.22µm) retain smaller cells and particles, as well as finer biofilms and matrices present in the sample which are not captured with large pore sizes. However, filters with smaller pore sizes become clogged more quickly, limiting water volume. More frequently, pore sizes of 0.45 µm or 0.8 µm are used for eDNA sampling, unless small particles such as viruses are the target of sampling.
- Once eDNA samples have been collected, the DNA extraction methodology used has important implications for the communities recovered. Commercial kits, specifically developed to minimise carryover of inhibitors from environmental samples, are now widely available, in many cases in high-throughput formats to facilitate consistent, simple extraction from large numbers of samples simultaneously.
- Collection of large sample sets and high-throughput processing for DNA extraction and downstream analyses (such as metabarcoding and qPCR) demands the careful consideration of the controls required at each stage. Ideally this should include control filters (through which sterile water is filtered) control DNA extractions (to control for contaminants present in reagents or arising during the DNA extraction process), and the generation of control sequencing libraries, including separately indexed no template control reactions, and libraries from mock communities of known composition to demonstrate the efficacy and suitability of extraction methods and primer choice.

References

Andrushchyshyn, O., Magnusson, A. K., and Williams, D. D. 2003. Ciliate populations in temporary freshwater ponds: seasonal dynamics and influential factors. *Freshwater Biology* 48, 548 – 564.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Griffiths, R. A., Foster, J., Wilkinson, J., Arnett, A., Williams, P., and Dunn, F. 2014. Analytical and methodological development for improved surveillance of the Great Crested Newt. Defra Project WC1067. Freshwater Habitats Trust: Oxford.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R. A., Foster, J., Wilkinson, J. W., Arnell, A., Brotherton, P., Williams, P., and Dunn, F. 2015. Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus*). *Biological conservation* 183, 19 – 28.

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., and de Bruyn, M. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution* 29(6), 358 – 367.

Bruce, K., Blackman, R., Bourlat, S. J., Hellström, A. M., Bakker, J., Bista, I., Bohmann, K., Bouchez, A., Brys, R., Clark, K., Elbrecht, V., Fazi, S., Fonseca, V., Hänfling, B., Leese, F., Mächler, E., Mahon, A. R., Meissner, K., Panskep, K., Pawlowski, J., Schmidt Yáñez, P., Seymour, M., Thalinger, B., Valentini, A., Woodcock, P., Traugott, M., Vasselon, V., and Deiner, K. 2021. A practical guide to DNA-based methods for biodiversity assessment. Advanced Books. https://doi.org/10.3897/ab.e68634.

Callaghan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. H., Johnson, A. J. A., and Holmes, S. 2016. DADA2: High resolution sample inference from Illumina amplicon data. *Nature Methods* 13(7), 581 – 583.

Cavalier-Smith, T., Lewis, R., Chao, E. E., Oates, B., and Bass, D. 2009. *Helkesimastix marina* n. sp. (Cercozoa: Sainouroidea superfam. N.) a gliding zooflagellate of novel ultrastructure and unusual ciliary behaviour. *Protist* 160, 452 – 479.

Cheng, T., Xu, C., Lei, L., Li, C., Zhang, Y., and Zhou, S. 2016. Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. *Molecular Ecology Resources* 16, 138 – 149.

Clausen, P. T. L. C.,Aarestrup, F. M., and Lund, O. 2018. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* 19(1), 307.

Compson, Z. G., McClenaghan B., Singer, G. A. C., Fahner, N. A., and Hajibabaei, M. 2020. Metabarcoding from microbes to mammals: comprehensive bioassessment on a global scale. *Frontiers in Ecology and Evolution* 8, 581835.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Kelley Thomas, W., Potter, C., and Bik, H. M. 2016. The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution* 7, 1008 – 1018.

Davies, B., Biggs, J., Williams, P., Whitfield, M., Nicolet, P., Sear, D., Bray, S., and Maund, S. 2008. Comparative biodiversity of aquatic habitats in the European agricultural landscape. *Agriculture Ecosystems and Environment* 125, 1 – 8.

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

Deiner, K., Bik, H. M., Mächler, E.,m Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., and Bernatchez, L. 2017. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology* 26, 5872 – 5895.

Gassman, M., and McHoull, B. (2015) DNA Integrity Number (DIN) with the Agilent 2200 TapeStation system and the Agilent Genomic DNA ScreenTape assay. Agilent Technologies, Waldbronn, Germany. Available at: www.agilent.com/c/library/applications/5991-5258EN.pdf (Accessed 10.03.2023).

Gong, J., Dong, J., Liu, X., and Massana, R. 2013. Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of Oligotrich and Peritrich ciliates. *Protist* 164(3), 369 – 379.

Guillou, L., Baar, D., Audic, S., Bass, D., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A-L., Siano, R., Stoeck, T., Vaulot D., Zimmerman, P., and Christen, R. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small subunit rRNA sequences with curated taxonomy. *Nucleic Acids Research* 41(D1), D587 – D604.

Harper, L. R., Buxton, A. S., Rees, H. C., Bruce, K., Brys, R., Halfmaerten, D., Read, D. S., Watson, H. V., Sayer, C. D., Jones, E. P., Priestley, V., Mächler, E., Múrria, C., Garcés-Pastor, S., Medupin, C., Burgess, K., Benson, G., Boonham, N., Griffiths, R. A., Handley, L. L., and Hänfling, B. 2019a. Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds. *Hydrobiologia* 826, 25 – 41.

Harper, L. R., Griffiths, N. P., Handley, L. L., Sayer, C. D., Read, D. S., Harper, K. J., Blackman, R. C., Li, J., and Hänfling, B. 2019b. Development and application of environmental DNA surveillance for the threatened crucian carp (*Carassius Carassius*). *Freshwater Biology* 61(1), 93 – 107.

Hill, M. J., Hassall, C., Oertli, B., Fahrig, L., Robson, B. J., Biggs, J., Samways, M. J., Usio, N., Takamura, N., Krishnaswamy, J., and Wood, P. J. 2017. New policy directions for global pond conservation. *Conservation Letters* 11, e12447.

Hill, M. J., Greaves, H. M., Sayer, C. D., Hassall, C., Milin, M., Milner, V. S., Marazzi, L., Hall, H., Harper, L. R., Thornhill, I., Walton, R., Biggs, J., Ewald, N., Law, A., Willby, N., White, J. C., Briers, R. A., Mathers, K. L., Jeffries, M. J., and Wood, P. J. 2021. Pond ecology and conservation: research priorities and knowledge gaps. *Ecosphere* 12(12), e03853.

Hugerth, L. W., Muller, E. E. L., Hu, Y. O. O., Lebrun, L. A. M., Roume, H., Lundin, D., Wilmes, P., and Andersson, A. F. 2015. Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoSONE* 9(4): e95567.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* 28(23), 3150 – 52.

Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., Bending, G., Hilton, S., Bass, D., and Burki, F. 2020. Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular ecology resources* 20(2), 429 – 443.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. 2013. Development of a dual-index sequencing strategy and curation pipeline for analysing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology* 79(17), 5112 – 5120.

Lindahl, B. D., Nilsson, R. H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjøller, R., Kõljalg, U., Pennanen, T., Rosendahl, S., Stenlid, J., and Kauserud, H. 2013. Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytologist* 199, 288 – 299.

Martin, M. 2011. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet.Journal* 17(1), 10 – 12.

Mauvisseau, Q., Coignet, A., Delaunay, C., Pinet, F., Bouchon, D., Souty-Grosset, C. 2017. Environmental DNA as an efficient tool for detecting invasive crayfishes in freshwater ponds. *Hydrobiologia* 805, 163 – 175.

McMurdie, P. J., and Holmes, S. 2013. Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8(4), e61217

Minardi, D., Ryder, D., del Campo, J., Fonseca, V. G., Kerr, R., Mortensen, S., Pallavicini, A., and Bass, D. 2022. Improved high throughput protocol for targeting eukaryotic symbionts in metazoan and eDNA samples. *Molecular Ecology Resources,* 22(2), 664 – 678.

Muha, T. P., Robinson, C. V., de Leaniz, C. G., and Consuegra, S. 2019. An optimised eDNA protocol for detecting fish in lentic and lotic freshwaters using a small water volume. *PLoS ONE* 14, e0219218.

Oldham, R. S., Keeble, J., Swan, M. J. S., and Jeffcote, M. 2000. Evaluating the suitability of habitat for the Great Crested Newt (*Triturus cristatus*). *Herpetological Journal* 10, 143 – 155.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., Joao Feio, M., Filipe, A. F., Fornaroli, R., Graf, R., Herder, J., van der Hoorn, B., Jones, J. I., Sagova-Mareckova, M., Moritz, C., Barquín, J., Piggot, J. J., Pinna, M., Rimet, F., Rinkevich, B., Sousa-Santos, C., Specchia, V., Trobajo, R., Vasselon, V., Vitecek, S., Zimmerman, J., Weigand, A., Leese, F., and Kahlert, M. 2018. The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment* 637-639, 1295 – 1310.

Pawlowski, J., Apothéloz-Perret-Gentil, L., and Altermatt, F. 2020. Environmental DNA: what behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology* 29, 4258 – 4264.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Resources* 41(D1), D590 – D596.

Quinlan, A. R., and Hall, I. M. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6), 841 – 42.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.* https://www.R-project.org/.

Sahlin, K., Lim, M. C. W., and Prost, S. 2021. NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data. *Ecology and Evolution* 11(3), 1392 – 98.Schwelm, A., Berney, C., Dixelius, C., Bass, D., and Neuhauser, S. 2016. The large subunit rDNA sequence of *Plasmodiophora brassicae* does not contain intra-species polymorphism. *Protist* 167, 544 – 554.

Shen, W., Shuai L., Li, Y., and Hu, G. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* 11(10), e0163962.Tedersoo, L., Bahram, M., Zinger, L., Henrik Nilsson, R., Kennedy, P. G., Yang, T., Anslan, S., and Mikryukov, V. 2022. Best practices in metabarcoding of fungi: from experimental design to results. *Molecular Ecology* 31, 2769 – 2795.

Wang, Q., Garrity, G. M., Tiedje, J. M., Cole, J. R. 2007. Naive Bayesian Classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 73(16), 5216 – 5267.

Wickham, H. Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43), 1686.

Williams, P., Biggs, J., Crowe, A., Murphy, J., Nicolet, P., Weatherby, A., and Dunbar, M. 2010. Countryside Survey: Ponds Report from 2007. Technical Report No. 7/07 Pond Conservation and NERC/Centre for Ecology & Hydrology, 77pp. (CEH Project Number: C03259).

Wilson, I.G. 1997. Inhibition and facilitation of nucleic acid amplification. *Applied and Environmental Microbiology* 63, 3741 – 3751.

Xing, Y., Gao, W., Shen, Z., Zhang, Y., Bai, J., Cai, X., Ouyang, J., and Zhao, Y. 2022. A review of environmental DNA field and laboratory protocols applied in fish ecology and environmental health. *Frontiers in Environmental Science* 10, 725360.

Glossary

A260/A280 ratio: The ratio of light absorbance of a DNA sample at a wavelength at 260 nm to the absorbance at 280 nm, used as an indicative measure of nucleic acid purity. For pure samples, a value of 1.8 is expected. Deviation from this can be the result of contamination of the sample with proteins and other chemicals which absorb light at a wavelength of 280nm.

A260/A230 ratio: The ratio of light absorbance of a DNA sample at a wavelength at 260 nm to the absorbance at 230 nm, used as an indicative measure of nucleic acid purity. For pure samples, a value of 2.0 is expected. Values much lower than this can indicate the presence in the sample of contaminant carbohydrates or common chemicals used in DNA extraction, such as guanidine hydrochloride.

Amplicon: a segment of DNA targeted and amplified during polymerase chain reaction (PCR).

Community DNA: the total DNA present in an environmental matrix, including DNA from live microbial cells, entire individuals or tissue fragments of higher organisms such as invertebrates, and extracellular DNA (DNA released from shed cells, organelles, faeces, and mucus of organisms in the locality of the matrix).

DNA integrity: a measure of the intactness of DNA molecules, used to determine the extent of degradation of the sample. Highly intact DNA is desirable for many molecular assays, particularly amplicon-based metabarcoding studies.

Ecosystem services: the contribution that ecosystems make to human health and wellbeing. This includes provisioning (for example water, food and materials), regulation (climate regulation and pollination), supporting processes (such as water and nutrient cycling and photosynthesis), and cultural aspects, such as recreation, tourism, and the role of nature in increasing quality of life and wellbeing.

Environmental DNA (eDNA): DNA extracted from environmental matrices, such as water, sediment, air or soil, originating from shed cells, organelles, faeces and mucus, as well as extracellular DNA present in the matrix from degraded matter.

Fluorometric DNA quantification: fluorometric assays use molecules (fluorophores) which excite and produce a light signal only when bound to their target substrate, which is usually double-stranded DNA. Flourometric assays are more accurate and targeted than other widely used methods for nucleic acid quantification such as UV-Vis spectrophotometry (see below), as fluorophores bind only to specific structures in the molecule of interest, and do not cross-react with similar molecules or contaminant compounds.

Habitat Suitability Index (HSI): Habitat suitability indices are scoring systems are numerical values assigned to a habitat based on its suitability to support a given species or group. This value is usually calculated from a number of different variables, including abiotic conditions, water quality, and the presence of other species. The HSI for the great crested newt was developed by Oldham and others (2000).

High-throughput sequencing: technologies which allow the parallel sequencing of large numbers (hundreds to millions) of DNA molecules simultaneously, from a single or multiple samples.

Metabarcoding: an approach which pairs PCR-based amplification of taxonomically-informative "barcode" regions of DNA with high-throughput sequencing to generate sequence data from complex, mixed samples. The sequences generated can then be taxonomically assigned against specialised, curated databases of known barcode sequences.

Polymerase chain reaction (PCR): PCR is a technique in which a targeted region of DNA is rapidly produced (amplified) over a controlled number of cycles. The region is targeted by the addition of short, synthetic DNA molecules called primers, which bind to DNA molecules present in the sample and provide a binding site for the enzyme that catalyses the amplification (the polymerase) to work.

Pond: While there is no globally accepted definition of a pond, the most widely used definition in the UK and Europe comes from Williams and others (2010) and states that a pond is "a body of standing water 0.0025 hectares to 2 hectares in area, which usually holds water for at least four months of the year".

Pondscape: A network of ponds (see above) and the terrestrial matrix which surrounds them. Within a pondscape, there is typically wide variation in environmental conditions and biodiversity.

Quantitative polymerase chain reaction (qPCR): qPCR is a technique based on PCR (see above) which uses a fluorescent probe to quantify the change in the amount of target DNA present in the sample after each cycle of amplification. This allows you to accurately and sensitively quantify the number of copies of your target present in the original sample.

UV-Vis spectrophotometry: a technique which quantifies molecules within a sample based on the absorbance of light by the sample at different wavelengths within the ultraviolet and visible range of the electromagnetic spectrum. Spectrophotometric techniques are less accurate than fluorometric assays for quantifying high-quality DNA within a sample (see above), they are widely used to determine the presence of contaminant substances in DNA extracts, which absorb light at different wavelengths to nucleic acids.

Appendices

Appendix 1

Links to publicly available, published district level licensing (DLL) data through the Defra data portals, complete as of 28.03.2023.

Great Crested Newt (GCN) environmental DNA (eDNA) and Habitat Suitability Index (HSI) data from ponds surveyed for strategic licensing can be accessed and explored interactively through the **Natural England Open Data Geoportal**. The raw datasets can also be accessed through the **Defra data portal**, including GCN eDNA and HSI results, some water pH and temperature data, and observations of other GCN life stages for the years 2017, 2018 and 2019. This link also includes copies of advice notes on HSI assessment from the Amphibian and Reptile Groups of the United Kingdom, attribute and spatial data files for the GCN DLL pond surveys carried out in the time period between 2017 and 2019, and the final report from the Freshwater Habitats Trust, Spygen, Amphibian Reptile Conservation and the Durrell Institute of Conservation Ecology.

Appendix 2

Values recorded for the three quality assessment assays run on the DLL DNA extracts, including measure of DNA integrity (TapeStation fragment size), DNA quantity (Double-stranded DNA), and sample purity (Nucleic acids, A260/A280 ratio and A260/A230 ratio). Short amplicon success indicates whether previous attempts at amplifying short metabarcoding regions were successful. NA values indicate that the sample was of insufficient quality to produce a value for the given assay.

Sample Code	Average fragment size (basepairs)	Double-stranded DNA (ng/µl)	Nucleic acids (ng/µl) [note 1]	A260/ A280 ratio	A260/ A230 ratio	Short amplicon success
GCN000482	NA	0.43	146.57	1.60	0.42	Yes
GCN000084	37935	2.96	36.37	1.49	0.59	Yes
GCN000104	22316	2.22	25.20	1.40	0.31	Yes
GCN000160	25577	1.70	84.82	1.43	0.64	Yes
GCN000594	51456	4.05	57.72	1.82	0.88	Yes
GCN000216	53266	6.02	190.71	1.48	0.70	Yes
GCN000628	18870	2.09	63.13	1.59	0.57	Yes
GCN000112	>60000	3.19	24.20	1.69	0.56	Yes
GCN000528	29776	2.36	110.89	1.53	0.72	Yes
GCN000136	8697	1.2	56.83	1.86	0.75	Yes
GCN000181	NA	0.73	50.99	1.54	0.50	Yes
GCN000053	25364	2.07	77.31	1.49	0.67	Yes
GCN000519	27817	1.61	322.13	1.51	0.49	Yes
GCN000390	NA	0.97	15.45	1.42	0.42	Yes

Sample Code	Average fragment size (basepairs)	Double-stranded DNA (ng/µl)	Nucleic acids (ng/µl) [note 1]	A260/ A280 ratio	A260/ A230 ratio	Short amplicon success
GCN000457	56244	1.01	11.88	1.87	0.40	Yes
GCN000382	57391	2.94	30.92	1.77	0.64	Yes
GCN000217	27776	10	70.95	1.55	0.78	Yes
GCN000386	NA	0.35	32.76	1.44	0.50	Yes
GCN000542	NA	0.38	30.62	1.30	0.40	Yes
GCN000459	50072	0.94	35.53	1.54	0.51	Yes
GCN000248	12409	3.88	17.55	1.67	0.55	Yes
GCN000574	11840	6.6	114.21	1.63	0.81	Yes
GCN000402	NA	0.24	40.84	1.54	0.40	Yes
GCN000450	>60000	4.72	71.88	1.51	0.78	Yes
GCN000380	52848	0.59	34.85	1.65	0.37	Yes
GCN000571	57894	4.72	38.27	1.97	0.52	Yes
GCN000505	24194	NA	236.89	1.45	0.76	Yes
GCN000481	51798	3.09	116.94	1.45	0.85	Yes
GCN000580	>60000	2.34	17.82	2.06	0.3	Yes
GCN000593	25135	3.03	97.98	1.58	0.7	Yes
GCN000167	28929	2.03	20.52	1.73	0.33	No
GCN000003	NA	0.49	84.30	1.55	0.43	No

Sample Code	Average fragment size (basepairs)	Double-stranded DNA (ng/µl)	Nucleic acids (ng/µl) [note 1]	A260/ A280 ratio	A260/ A230 ratio	Short amplicon success
GCN000027	37187	1.65	30.35	1.46	0.47	No
GCN000064	29121	2.59	37.07	2.16	0.42	No
GCN000080	NA	0.51	31.57	2.23	0.15	No
GCN000111	26689	2.14	23.15	1.87	0.46	No
GCN000180	1282	0.51	13.74	2.31	0.15	No
GCN000645	NA	0.54	18.61	2.62	0.11	No
GCN000648	31271	1.45	12.95	2.02	0.25	No
GCN000654	53485	1.15	12.90	2.20	0.21	No
GCN000227	>60000	6.18	44.32	2.04	0.79	No
GCN000171	50733	4.91	35.28	1.66	0.56	No
GCN000186	NA	0.49	54.68	1.63	0.33	No
GCN000235	28398	0.71	6.98	2.08	0.18	No
GCN000264	56886	1.98	10.61	1.85	0.21	No
GCN000295	56305	2.32	27.33	1.95	0.54	No
GCN000438	>60000	3.89	19.71	1.87	0.38	No
GCN000440	NA	2.32	29.14	1.83	0.20	No
GCN000033	42588	7.47	48.95	1.98	0.74	No
GCN000042	NA	0.27	4.05	1.76	0.13	No

Sample Code	Average fragment size (basepairs)	Double-stranded DNA (ng/µl)	Nucleic acids (ng/µl) [note 1]	A260/ A280 ratio	A260/ A230 ratio	Short amplicon success
GCN000656	>60000	0.69	6.12	2.20	0.16	No
GCN000120	52938	0.60	13.93	1.58	0.31	No
GCN000309	52976	1.06	51.77	1.97	0.64	No
GCN000389	54097	0.92	10.38	1.70	0.23	No
GCN000249	NA	2.32	30.87	1.57	0.47	No
GCN000232	23876	1.13	79.40	1.57	0.29	No
GCN000238	26220	2.75	17.37	1.61	0.28	No
GCN000347	32176	0.40	22.31	1.54	0.35	No
GCN000245	NA	4.72	3.73	1.87	0.10	No
GCN000296	12535	0.15	16.71	1.88	0.38	No

Note 1: the value for nucleic acids given here includes all nucleic acids present in the sample, including double-stranded DNA and contaminant single-stranded DNA, RNA and free nucleotides, and cannot be used as an accurate quantification of good quality, amplifiable double-stranded DNA.

Appendix 3

Values recorded for the three quality assessment assays run on the Pinkhill Meadows DNA extracts, including measure of DNA integrity (TapeStation fragment size), DNA quantity (Double-stranded DNA), and sample purity (Nucleic acids, A260/A280 ratio and A260/A230 ratio). Amplification successful indicates whether the sample was successfully amplified using any of the three primer pairs tested in this study. NA values indicate that the sample was of insufficient quality to produce a value for the given assay

Sample code	Sample type	Average fragment size (base pairs)	Double stranded DNA (ng/µl)	Nucleic acids (ng/µl) [note 1]	A260/ A280 ratio	A260/ A230 ratio	Amplification successful
SSPMWN	Water	11573	23.94	21.12	1.84	0.99	Yes
GPaWN	Water	10848	27.23	21.08	1.77	1.85	Yes
MP11WN	Water	11431	18.30	12.62	1.91	1.46	Yes
MP7WS	Water	12512	13.84	9.62	1.87	1.43	Yes
SWPmWS	Water	8941	41.02	41.62	1.89	1.91	Yes
GPCWS	Water	10880	19.47	15.13	1.80	0.51	Yes
GPbWN	Water	9104	22.83	19.31	1.91	0.37	Yes
SRB1WS	Water	12479	48.89	48.69	1.91	1.91	Yes
MP9WS	Water	10382	33.13	30.22	1.90	1.73	Yes
EP1WS	Water	10414	41.40	22.83	1.94	1.83	Yes
MP9WN	Water	9759	16.15	11.55	1.79	0.35	No
SRb1WN	Water	7131	22.12	12.25	1.74	0.39	Yes
GPCWN	Water	11565	21.34	8.06	2.03	0.69	Yes

Sample code	Sample type	Average fragment size (base pairs)	Double stranded DNA (ng/µl)	Nucleic acids (ng/µl) [note 1]	A260/ A280 ratio	A260/ A230 ratio	Amplification successful
MP11WS	Water	9873	33.13	26.49	1.91	1.54	Yes
EP5WS	Water	8970	41.40	22.10	1.82	1.68	Yes
GWPMwS	Water	9851	7.37	90.29	1.89	1.65	Yes
EP1WN	Water	10414	21.34	44.02	1.87	1.60	Yes
MP18WN	Water	9789	7.24	24.18	1.86	1.72	Yes
MP11SN	Sediment	11444	122.29	123.74	1.91	2.13	Yes
MP18SS	Sediment	16495	65.15	69.54	1.92	1.96	Yes
GPaSS	Sediment	9407	72.68	76.51	1.88	1.81	Yes
MP18SN	Sediment	9256	67.89	51.2	1.9	1.89	Yes
FWBS	Sediment	NA	7.37	1.82	1.36	0.45	No
SPP1SS	Sediment	10020	72.84	67.81	1.88	1.69	Yes
MP16SN	Sediment	NA	7.24	0.79	1.24	0.26	Yes
GP3SN	Sediment	13470	48.89	37.93	1.94	1.91	Yes
SWP2SS	Sediment	9820	72.84	37.90	1.90	2.04	Yes
GPcSS	Sediment	10626	16.15	45.98	1.88	1.23	Yes
MP9SS	Sediment	11072	22.12	24.20	1.85	0.69	Yes
EP1SN	Sediment	6951	38.80	25.02	1.89	1.53	Yes

Note 1: the value for nucleic acids given here includes all nucleic acids present in the sample, including double-stranded DNA and contaminant single-stranded DNA, RNA and free nucleotides, and cannot be used as an accurate quantification of good quality, amplifiable double-stranded DNA.

Appendix 4

Summary of amplification success of the three different primer sets tested using the eDNA samples from the district level licensing great crested newt survey included in this study

Sample code	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles	Eukaryote small subunit V4 (700 basepairs), 40 cycles
GCN000004	No	No	No	Yes
GCN000006	No	No	No	Yes
GCN000012	No	No	No	No
GCN000015	No	No	No	No
GCN000023	No	No	No	No
GCN000033	No	No	No	No
GCN000042	No	No	No	No
GCN000045	No	No	No	No
GCN000053	No	Yes	No	Yes
GCN000073	No	No	No	No
GCN000084	No	No	No	No
GCN000096	No	No	No	No
GCN000102	No	No	No	No
GCN000104	No	No	No	Yes
GCN000112	No	No	No	Yes

Sample code	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles	Eukaryote small subunit V4 (700 basepairs), 40 cycles
GCN000115	No	No	No	No
GCN000120	No	No	No	No
GCN000122	No	No	No	Yes
GCN000134	No	No	No	Yes
GCN000136	No	No	No	Yes
GCN000160	No	No	No	No
GCN000171	No	No	No	No
GCN000179	No	No	No	No
GCN000181	No	No	No	No
GCN000186	No	No	No	No
GCN000216	No	No	No	Yes
GCN000217	No	No	No	Yes
GCN000225	No	No	No	No
GCN000229	No	No	No	Yes
GCN000232	No	No	No	No
GCN000235	No	No	No	No
GCN000238	No	No	No	No
GCN000240	No	No	No	No

Sample code	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles	Eukaryote small subunit V4 (700 basepairs), 40 cycles
GCN000245	No	No	No	No
GCN000248	No	No	Yes	Yes
GCN000249	No	No	No	No
GCN000264	No	No	No	No
GCN000265	No	No	No	No
GCN000271	No	No	No	No
GCN000283	No	No	No	No
GCN000291	No	No	No	No
GCN000295	No	No	No	No
GCN000296	No	No	No	No
GCN000301	No	No	No	No
GCN000309	No	No	No	No
GCN000316	No	No	No	No
GCN000345	No	No	No	Yes
GCN000347	No	No	No	No
GCN000352	No	No	No	Yes
GCN000378	No	No	No	No

Sample code	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles	Eukaryote small subunit V4 (700 basepairs), 40 cycles
GCN000380	No	No	Yes	Yes
GCN000382	Yes	No	No	Yes
GCN000386	No	Yes	Yes	Yes
GCN000389	No	No	No	No
GCN000390	No	No	Yes	Yes
GCN000399	No	No	No	No
GCN000402	No	Yes	No	Yes
GCN000404	No	No	No	No
GCN000431	No	No	No	No
GCN000438	No	No	No	No
GCN000440	No	No	No	No
GCN000444	No	No	No	No
GCN000450	No	Yes	Yes	Yes
GCN000451	No	No	No	Yes
GCN000452	No	No	No	Yes
GCN000453	No	No	No	No
GCN000457	No	Yes	Yes	Yes
GCN000459	No	Yes	No	No

Sample code	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles	Eukaryote small subunit V4 (700 basepairs), 40 cycles
GCN000481	No	No	Yes	Yes
GCN000482	No	No	No	Yes
GCN000485	No	No	No	No
GCN000504	No	No	No	Yes
GCN000505	No	Yes	No	No
GCN000518	No	No	No	No
GCN000519	No	No	No	No
GCN000527	No	No	No	Yes
GCN000528	No	No	No	Yes
GCN000536	No	No	No	No
GCN000541	No	No	No	Yes
GCN000542	No	Yes	No	Yes
GCN000544	No	No	No	No
GCN000545	No	No	No	No
GCN000552	No	No	No	Yes
GCN000556	No	No	No	Yes
GCN000558	No	No	No	Yes
GCN000561	No	No	No	No

Sample code	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles	Eukaryote small subunit V4 (700 basepairs), 40 cycles
GCN000564	No	No	No	Yes
GCN000565	No	No	No	Yes
GCN000566	No	No	No	No
GCN000570	No	No	No	No
GCN000571	No	No	No	No
GCN000574	No	No	No	Yes
GCN000577	No	No	No	Yes
GCN000580	No	No	No	No
GCN000581	No	No	No	Yes
GCN000587	No	No	No	Yes
GCN000588	No	No	No	No
GCN000593	No	No	No	Yes
GCN000594	No	No	Yes	No
GCN000608	No	No	No	Yes
GCN000609	No	No	No	Yes
GCN000623	No	No	No	No
GCN000625	No	No	No	No
GCN000628	No	No	No	Yes

Sample code	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles	Eukaryote small subunit V4 (700 basepairs), 40 cycles
GCN000634	No	No	No	No
GCN000640	No	No	No	Yes
GCN000644	No	No	No	Yes
GCN000645	No	No	No	No
GCN000648	No	No	No	No
GCN000654	No	No	No	No
GCN000656	No	No	No	No

Appendix 5

Summary of amplification success of the three different primer sets tested using the community DNA samples from the Pinkhill Meadows pondscape used as a quality benchmark for this study.

Sample code	Sample type	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles
SSPmWN	Water	Yes	Yes	No
GPaWN	Water	Yes	Yes	No
MP11WN	Water	Yes	Yes	Yes
MP7WS	Water	Yes	No	Yes
SWPmWS	Water	Yes	Yes	Yes
GPcWS	Water	Yes	Yes	Yes
GPbWN	Water	Yes	Yes	Yes
SRB1WS	Water	Yes	Yes	Yes
MP9WS	Water	Yes	No	Yes
EP1WS	Water	Yes	No	Yes
MP9WN	Water	No	No	No
SRB1WN	Water	No	Yes	Yes
GPcWN	Water	Yes	No	Yes
MP11WS	Water	Yes	Yes	Yes
EP5WS	Water	Yes	Yes	Yes

Sample code	Sample type	Eukaryote rRNA (4500 basepairs)	Plant ITS (1100 basepairs)	Eukaryote small subunit V4 (700 basepairs), 30 cycles
GWPmWS	Water	Yes	Yes	Yes
EP1WN	Water	Yes	Yes	Yes
MP18WN	Water	Yes	No	Yes
MP11SN	Sediment	Yes	Yes	Yes
MP18SS	Sediment	Yes	Yes	Yes
GPaSS	Sediment	Yes	No	Yes
MP18SN	Sediment	Yes	Yes	Yes
FWBS	Sediment	No	No	No
SPP1SS	Sediment	No	Yes	Yes
MP16SN	Sediment	Yes	No	No
GP3SN	Sediment	Yes	Yes	Yes
SWP2SS	Sediment	Yes	Yes	No
GPcSS	Sediment	Yes	Yes	No
MP9SS	Sediment	No	Yes	No
EP1SN	Sediment	Yes	Yes	Yes

Appendix 6

Values recorded for the quality assessment assays run on a subset of cleaned eukaryote small subunit V4 hypervariable region amplicons generated from DLL samples. Samples were checked for double-stranded DNA quantity (using a Quantus ONE dsDNA fluorometric assay) and sample purity, including carryover of non-target nucleic acids, A260/A280 ratio and A260/A230 ratio using a NanoDrop One UV-Vis spectropohotmetric assay using molecular-grade water as a blank.

Sample code	Double-stranded DNA (ng/µl)	Nucleic acids (ng/µl)	A260/A280 ratio	A260/A230 ratio
GCN000004	1.7	227.89	2.49	1.56
GCN000482	16.0	68.11	1.90	2.43
GCN000104	147.2	237.02	1.69	0.64
GCN000160	289.9	429.15	1.80	1.18
GCN000594	73.0	243.55	2.01	1.05
GCN000216	91.7	189.49	1.18	0.31
GCN000628	84.2	210.86	1.90	2.69
GCN000112	46.6	215.07	1.15	0.31
GCN000528	21.9	55.90	1.50	0.40
GCN000136	37.1	88.23	1.91	0.70
GCN000053	1.5	509.29	1.41	0.20
GCN000390	31.4	105.02	1.65	0.83
GCN000457	41.4	97.51	1.56	0.76
GCN000122	40.4	96.15	1.76	1.29
GCN000382	33.6	133.75	1.56	0.83

Sample code	Double-stranded DNA (ng/µl)	Nucleic acids (ng/µl)	A260/A280 ratio	A260/A230 ratio
GCN000217	75.0	145.98	1.72	0.91

www.gov.uk/natural-england

NATURAL
ENGLAND