Testing the validity of using eDNA for carbon origin analysis from sediment cores

September 2024

Natural England Commissioned Report NECR527

www.gov.uk/natural-england

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: NECR527

Report details

Author(s)

Eulyn Pagaling, Victoria G. Buswell, Amy Cooper, Pete Hedley, Peter J. A. Cock, Lucinda Robinson, Jennifer Morris

Natural England Project Manager

Josh Parker, Marine Ecology Specialist, Natural England

Joshua.Parker@naturalengland.org.uk

Contractor

The James Hutton Institute

Craigiebuckler

Aberdeen

AB15 8QH

Keywords

Blue carbon, intertidal, marine, terrestrial, metabarcoding, carbon origin analysis, DNA

Citation

Pagaling, E., Buswell, V. G., Cooper. A., Hedley, P., Cock, P. J. A., Robinson, L. & Morris, J. 2024. Testing the validity of using eDNA for carbon origin analysis from sediment cores. NECR527. Natural England.

Foreword

Natural England commissioned this project to pilot the approach of using environmental DNA metabarcoding for carbon origin analysis. This was a proof-of-concept project which aims to understand how DNA metabarcoding can be used to assess the origin and flows of carbon within intertidal sediments, complementing conventional stable isotope analysis and carbon content analysis.

Certain habitats, such as saltmarsh and seagrass, have been extensively studied as carbon stocks. However, intertidal and subtidal sediments remain understudied with significant evidence gaps relating to carbon origins and flows.

Understanding the origin of carbon within marine sediments could inform management of habitats, both on land and at sea, as well as climate change and Net Zero reporting.

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

Blue carbon is the carbon captured by living organisms in marine ecosystems and stored in biomass and sediments. There is growing interest in identifying and characterising blue carbon habitats because they are important for understanding how they could contribute towards Net Zero targets in the future. Currently, blue carbon habitats are not included in the UK Greenhouse Gas Iinventory (GHGi), however the Intergovernmental Panel on Climate Change (IPCC) Wetland Supplement (IPCC, 2014) includes guidelines for the quantification and accounting of GHG emissions and removals associated with the management of different wetland types, which could provide a mechanism for including saltmarsh and seagrass in the UK GHGi. Other blue carbon habitats, such as subtidal and intertidal marine sediments, do not currently have a mechanism for inclusion to the UK GHGi.

Working towards the potential inclusion of saltmarsh and seagrass in the UK GHGi is listed as the first Objective in the UK Blue Carbon Evidence Partnership (UKBCEP) Evidence Needs Statement (UKBCEP, 2023), and the focus of blue carbon research within the UK is on filling the evidence gaps associated with the carbon stocks, characteristics and accumulation within these habitats. However, little is known about the carbon stocks associated with marine sediments. In addition to quantifying the carbon stocks and accumulation rates to understand the potential importance of these habitats as carbon sinks, characterisation of the sources and transport pathways of carbon within marine sediments is highly important to inform conservation and management of marine, intertidal and terrestrial habitats.

Metabarcoding of environmental DNA (eDNA) is emerging as a method for characterising origins of carbon stocks, giving insights into the relative contribution of marine, coastal and terrestrial organisms. While there have been numerous biodiversity studies using eDNA on temperate marine sediments, all current studies using eDNA methods on sediments to investigate the origins of blue carbon stocks have primarily been in tropical or sub-tropical environments.

The present pilot study tested the validity of metabarcoding to characterise blue carbon stocks in the UK using sediments from the Solway Firth. This was achieved through three main objectives:

1. Develop a publicly available and open access reference library for DNA metabarcoding of eDNA samples to investigate the origin of organic carbon from intertidal sediment samples (including both marine and terrestrial species).
2. Test the validity and suitability of using metabarcoding of eDNA samples as a method for carbon origin analysis from sediment cores in order to demonstrate its use for UK carbon calculations.
3. Create a spreadsheet listing the organisms identified (to the lowest taxonomic resolution, including scientific and common names), the fragment of DNA used to

identify that species, the reference library and sequence ID used to identify the species, the confidence value of the match and the number of reads.

Objective 1 summary

The study focussed on characterisation of macrophytes and molluscs. Two bespoke reference libraries were created for the rbcL gene (for the characterisation of macrophytes) and the 18S rRNA gene (for the characterisation of molluscs) by downloading publicly available sequence and taxonomic data associated with the target genes. The downloaded data was curated and used to construct a classifier in order to assign taxonomy to the sequence data generated from the Solway Firth samples.

Objective 2 summary

Collected sediments from the Solway Firth Special Area of Conservation (SAC), Cumbria, England were processed for eDNA. Issues with low DNA yield required a change in extraction methods. The custom databases were used to select primer sets for amplification of the rbcL gene and 18S rRNA gene, respectively, to ensure good coverage and amplicon length. These genes were PCR amplified from eDNA and sequenced using an Illumina MiSeq sequencer, generating paired-end 250 bp reads. A total of 8.2M high quality reads were obtained. However, distribution of read counts were uneven amongst samples, possibly due to low amplicon concentration. The sequences were all processed in the QIIME2 environment, which included quality checks and denoising before undergoing taxonomic assignment with the custom classifier. A lower threshold of 10,000 reads was used to filter out samples with low reads prior to taxonomic assignment.

Objective 3 summary

The final results were presented as read abundance tables, provided in a separate spreadsheet. Most of the rbcL dataset could be resolved to a lower taxonomic resolution (genus) and consisted of diatoms, while the 18S rRNA gene dataset was difficult to resolve to a low taxonomic resolution (order) and consisted of worms and arthropods. The community compositions of the rbcL dataset were consistent across the samples but were highly variable in the 18S rRNA gene dataset. The results suggested issues with primer bias (preference for marine and freshwater organisms with the rbcL primer set and preference for worms with the 18S rRNA gene primer set), but this could also be due to issues with low input DNA being carried through the process.

Recommendations

Metabarcoding of eDNA is a potential tool for characterisation of carbon origins in UK intertidal sediments. However, optimisation of the different steps in this process would be required for future studies to fully characterise the origins of UK carbon stocks. These are:

1. Review of sampling protocols to capture samples with high organic matter and provide sufficient material to allow replication, especially for low biomass samples.
2. Exploration of DNA preservation in deeper sediments containing locked-in carbon.
3. Consider higher sample replication for statistical analysis (10 sampling stations and 10 replicate samples per station).
4. Exploration of DNA extraction methods to reduce extraction bias.
5. Perform technical replicates at DNA extraction and PCR amplification stages to account for sediment heterogeneity and PCR stochasticity, respectively.
6. Review target genes or consider a 'toolbox' approach (selection of multiple target genes) to provide better coverage.
7. Test new selected primer sets *in silico* and *in vitro* (using mock communities) to determine primer bias. Test different PCR reagents, conditions, and volumes *in vitro* to determine species coverage.
8. Conduct temperature gradient PCR with new selected primer sets to obtain optimal PCR conditions. Use 40 PCR cycles for low biomass samples.
9. Include mock communities and/or DNA from single species in the sequencing to allow a direct test of the taxonomic assignment.
10. Ensure sufficient amplicon concentration for sequencing to ensure even distribution of reads amongst samples.
11. If the data allows, use the recommended threshold of >50,000 quality reads for further analysis during sequence processing. Otherwise, a threshold of >10,000 reads is reasonable for low biomass samples.

Contents

1 Introduction

1.1 Context

Blue carbon is carbon captured by living organisms in marine ecosystems and stored in biomass and sediments. There is growing interest in identifying and characterising blue carbon habitats because they are important for understanding how they could contribute towards Net Zero targets in the future (Ortega et al., 2020).

Working towards the potential inclusion of saltmarsh and seagrass in the UK GHGi is listed as the first Objective in the UK Blue Carbon Evidence Partnership (UKBCEP) Evidence Needs Statement (UKBCEP, 2023), and the focus of blue carbon research within the UK is on filling the evidence gaps associated with the carbon stocks, characteristics and accumulation within these habitats. Currently, blue carbon habitats are not included in the UK Greenhouse Gas Iinventory (GHGi), however the Intergovernmental Panel on Climate Change (IPCC) Wetland Supplement (IPCC, 2014) includes guidelines for the quantification and accounting of GHG emissions and removals associated with the management of different wetland types, which could provide a mechanism for including saltmarsh and seagrass in the UK GHGi. Characterisation of the sources and transport pathways of carbon is important for informing conservation and management of marine and intertidal habitats (Ortega et al., 2020) and to ensure the contribution of a carbon stock to net zero is not double counted (for example, counted once as peatland carbon stock, and again as a saltmarsh carbon stock due to carbon runoff from the peatland habitat).

Previously, stable isotope analysis was used to differentiate between terrestrial and marine organic matter in sediments (Ortega et al., 2020), but this method is expensive, has poor taxonomic resolution and may underestimate the contribution of marine organisms (Reef et al., 2017). Instead, metabarcoding of environmental DNA (eDNA) is emerging as a method for characterising origins of carbon stocks, giving insights into the relative contribution of marine, coastal and terrestrial organisms. However, all current studies using eDNA methods on sediments to investigate the origins of blue carbon have primarily been in tropical or sub-tropical environments (Reef et al., 2017, Ortega et al., 2020, Miyajima et al., 2022) or focus on specific macrophytes (Anglès d'Auriac et al., 2021). The present study was a pilot to test the validity of metabarcoding methods on marine sediments for potential use to characterise blue carbon stocks in the UK.

1.2 Scope

As defined by the UK Parliament, blue carbon systems include saltmarsh, seagrass, mangroves, and less well understood systems such as kelp, shelly reefs (bivalves) and maerl beds (calcifying seaweed) (Parliament, 2021). In this study, we investigated the

origin of organic carbon within intertidal sediments of the Solway Firth Special Area of Conservation (SAC), Cumbria, England.

We investigated the contribution of macrophytes (including seagrass, marine algae and land angiosperms) and molluscs (due to the carbon stored in their shells) to blue carbon stocks. Other organisms were out of scope of this small proof of concept study.

1.3 Aims and objectives

The overarching aim of this pilot study was to test the validity of using eDNA for carbon origin analysis from sediment cores. This was achieved through the following objectives:

1.3.1 Objective 1

Develop a publicly available and open access reference library for DNA metabarcoding of eDNA samples to investigate the origin of organic carbon from intertidal sediment samples (including both marine and terrestrial species).

1.3.2 Objective 2

Test the validity and suitability of using metabarcoding of eDNA samples as a method for carbon origin analysis from sediment cores in order to demonstrate its use for UK carbon calculations.

1.3.3 Objective 3

Create a spreadsheet listing the organisms identified (to the lowest taxonomic resolution, including scientific and common names), the fragment of DNA used to identify that species, the reference library and sequence ID used to identify the species, the confidence value of the match and the number of reads.

1.4 Approach

Sediments collected from the Solway Firth SAC, Cumbria, England were processed for eDNA. Macrophytes were characterised by the amplification of the plastid gene rbcL because it is an agreed plant DNA barcode by the Consortium for the Barcode of Life, which offers high coverage of plant and algal groups, with an extensive GenBank reference sequence library (Reef et al., 2017). Molluscs were characterised by amplification of the 18S rRNA gene because it shows good coverage of bivalves on a global scale (Espineria et al., 2009), including *Spisula* spp., which are prevalent in the Solway Firth (https://sac.jncc.gov.uk/site/UK0013025). Two bespoke reference databases were developed by downloading publicly available sequence and taxonomic data

associated with the target genes and taxa of interest using RESCRIPt (Robeson *et al.,* 2021). The downloaded data was then curated and used to construct a classifier in order to assign taxonomy to the sequence data generated from the Solway Firth samples. The custom databases were used to select primer sets for amplification of the rbcL gene and 18S rRNA gene, respectively. These genes were PCR amplified from eDNA and sequenced using an Illumina MiSeq sequencer, generating paired-end 250 bp reads. The sequences were all processed in the QIIME2 environment (version 2023.7). QIIME2 analysis includes quality checks and denoising before undergoing taxonomic assignment with the custom classifier. Final results were presented as read abundance tables. This allowed identification of organisms contributing to carbon in the intertidal sediments but does not directly measure organism abundance.

2 Methodology

2.1 Field site and sampling

Samples were collected by Natural England in February 2023 from intertidal sediments within the Solway Firth SAC, Cumbria, England at six sampling stations (B1;2;3 and C1;2;3; Table 1). Cores were taken for the analysis of carbon stocks from intertidal muddy sand using a Russian corer at each sampling station.

At each station, a core was taken to a depth of 60cm+ (excluding sampling stations C2 where only 20-30 cm depth was possible due to rocky ground and C3 where no at-depth sampling was possible).

From each sample, 1.5 ml sub-samples were collected from the surface layer and deepest layer of each core (typically 40-50 cm). Syringes were used to draw the wet sediment up and then transfer it to the pre-sterilised 1.5 ml Eppendorf tubes. Five replicate sub-samples were collected from each core sample at each depth where possible (54 samples in total).

During collection, new sterilised syringes and gloves were used for each sample to minimise the risk of cross contamination. Sub-samples were kept in separate labelled zip lock bags in a clean bucket at outdoor ambient temperature (<5°C). All sub-sample tubes and bags were wiped with a 10% bleach solution before storage to minimise the risk of cross contamination. Samples were frozen within 6 h of collection and stored at -20°C. No field controls were collected.

Table 1. Sample locations, sampling stations, sampling depths and broadscale habitats (BSH).

Sample station	Latitude	Longitude	Depth 1	Depth 2 (cm)	Broadscale habitat (BSH)
B1	54.90502	-3.26479	Surface	40-50	mud/sandy mud
B2	54.90398	-3.26656	Surface	40-50	mud/sandy mud
B3	54.90601	-3.26448	Surface	40-50	mud/sandy mud
C1	54.76828	-3.43618	Surface	≥43	sand/muddy sand
C2	54.76849	-3.43877	Surface	20-30	sand/muddy sand
C3	54.76858	-3.44028	Surface	NA	sand/muddy sand

2.2 DNA Extraction

Three DNA extraction methods were employed because DNA yields from the initial batches of samples were extremely low and prompted us to try alternative extraction methods. Therefore, different methods were used to extract different samples because the 3 g of sediment provided was not sufficient to use all three methods on each sample.

The first was a CTAB method, adapted from Griffiths (Griffiths et al., 2000). Briefly, sample lysis was done chemically by the addition of phenol/chloroform/isoamyl alcohol and hexadecyltrimethylammonium bromide (CTAB)/phosphate buffer, and mechanically by bead beating at 5000 rpm for 15 s using a Precellys (Bertin Technologies). Particulates and cell debris were removed by centrifugation, and organic matter was removed from the supernatant using phenol/chloroform. MaXtract tubes (Qiagen) were used to maximise aqueous phase recovery. DNA was precipitated from the aqueous phase using a PEG/NaCl solution and centrifugation. The DNA pellet was cleaned with 70% ethanol, then left to dry (to remove the ethanol) before resuspension in Tris-EDTA to prevent degradation (Lahiri and Schnabel, 1993). All sub-samples were extracted in triplicate, with

~250 mg of sediment used for each replicate. The final volume of eluate was 30 µl. One negative control was run for every set of 22 extractions, which consisted of sterile water in place of sample. Further experimental details can be found in the protocols in Appendix 1.1 and our Quality Assurance detailed in Section 2.9.

The second method used was the DNeasy PowerSoil Pro DNA extraction kit (Qiagen), which was carried out according to the manufacturer's instructions. All sub-samples were extracted in triplicate, with ~250 mg of sediment used for each replicate. The final eluate volume for this method was 50 µl. One negative extraction blank was run for every set of sample extractions, and a set varied in number between 9 and 22 extractions.

The third method used was the DNeasy PowerMax Soil DNA extraction kit (Qiagen), which was carried out according to the manufacturer's instructions. Samples were only extracted once with this kit due to insufficient available material (the kit can process up to 10 g soil in a single extraction). The entire remaining sediment sub-sample (3.3 g) was used for each extraction. The final volume of eluate was 5 ml. One negative was used for every set of 8 samples.

Positive controls for DNA extraction and PCR were also extracted. Positive control soil was extracted alongside the sediment samples as detailed above. The rbcL positive controls consisting of strawberry fruit and strawberry leaf (*Fragaria* sp.), grass (potential *Holcus sp.*), spider plant leaf (*Chlorophytum comosum*), a green leaf taken from an unknown tree, and seaweed tissue ('Organic Sea Salad' dried seaweed flakes, the Cornish Seaweed Company; contains Dulse, 'ocean greens', Nori) were extracted using the Dneasy Plant Pro Kit (Qiagen). Tropical fish food (raw, frozen fish food cubes of a compressed mixture of the following: Bloodworms (family Glyceridae), Mysis (Mysidae), Daphnia (Daphniidae), Cyclops (Cyclopidae), Brine shrimp/Artemia (Artemiidae), and algae (unknown family)) was also used as a positive control for rbcL due to the presence of algae in the mix, and was extracted using the Dneasy Blood & Tissue kit (Qiagen). The 18S rRNA gene positive controls consisting of bivalve tissue (Chilean mussel, *Mytilus chilensis*, shelled, rope-farmed in Chile, purchased as part of a supermarket raw seafood mix) was extracted with the Dneasy Blood & Tissue kit (Qiagen). Negative extraction controls consisted of molecular grade UltraPure Distilled Water, Dnase and Rnase Free. (Invitrogen) in place of sample.

DNA yield was measured using a Qubit Flex fluorometer and the dsDNA High Sensitivity assay (Thermo Fisher), and the quality of the DNA was checked by measurement of the 260/280 ratio using a NanoDrop spectrophotometer (Thermo Fisher), both according to manufacturer's instructions. A ratio of 1.8 is generally accepted to indicate pure DNA. However, due to the low DNA yields obtained from the samples, it was difficult to obtain accurate readings of the 260/280 ratio and were all outside this acceptable range (readings ranged from -39.58 to 142.61) (see Section 3.2). Visualisation of DNA integrity was done using agarose gel electrophoresis. A 1.5% gel and 1.5% TBE buffer was used

with 6X gel loading dye (New England Biolabs) and was run for 45 min at 80 V. DNA was stored at -20ºC.

2.3 Custom databases and primer selection

Two custom databases were built by obtaining sequences and taxonomic labels from the National Centre for Biotechnology Information (NCBI). Specifically, the QIIME2 plugin RESCRIPt (Bolyen et al., 2019, Robeson et al., 2021) was used to download data for the two databases. Data for macrophytes (rbcL gene) database was downloaded if it had the label ‘rbcl’ and the associated sequence was >1000 bp. For the mollusc database (18S rRNA gene), data was downloaded if it had both an ‘18S’ and ‘Mollusca’ or ‘molluscs’ labels and was >1000 bp in length.

A review of the literature for metabarcoding primers targeting marine macrophytes and molluscs identified 8 primer sets for macrophytes (rbcL) and 2 primer sets for molluscs (18S rRNA gene) (Table 2). To ascertain the most appropriate primer set, the primer sequences were used to extract the targeted region from the reference databases. Specifically, reads were extracted using the ‘feature-classifier extract-reads’ commands in QIIME2. This command performs *in silico* PCR that extracts amplicons from reference sequences that match the input primer sequences using an identity default threshold of 0.8. The reads were extracted, primers were removed and any reads shorter than 50 bp were discarded. The primer sets were assessed based on the number of sequences and taxonomic groups that the primers capture and whether the fragment lengths produced would be appropriate for the sequencing platform (250 bp paired end). The primer sets that were selected are in bold in Table 2. The *in silico* analysis confirmed that the rbcL primer set picks up both terrestrial and marine organisms and the 18S rRNA gene primer set picks up bivalves and gastropods.

Table 2. Primers identified for amplification of rbcL and 18S rRNA genes. Primers sets selected for the study are in bold.

Primer name	Primer Sequence (5'-3')	Target group	Target gene	Amplicon size (bp)	Reference
rbcL-F	GCGGGTGTTAAAGAGTACAA	Marine macrophytes	rbcL	146	(Ortega et al., 2019)
rbcL-R	AGTAGAAGATTCGGCAGCTA	Marine macrophytes	rbcL	146	(Ortega et al., 2019)
Z1aF	ATGTCACCACCAACAGAGACTAAAGC	Marine macrophytes	rbcL	600	(Reef et al., 2017)
R604	CTGRGAGTTMACGTTTTCATCATC	Marine macrophytes	rbcL	600	(Reef et al., 2017)
F52_tag	GTTGGATTCAAAGCTGGTGTTA	Marine macrophytes	rbcL	350	(Reef et al., 2017)
rcblB_tag	AACCYTCTTCAAAAAGGTC	Marine macrophytes	rbcL	350	(Reef et al., 2017)
Diat_rbcL_708F_1	AGGTGAAGTAAAAGGTTCWTACTTAAA	Phytobenthos	rbcL	312	(Vasselon et al., 2017, Zimmermann et al., 2021)
Diat_rbcL_708F_2	AGGTGAAGTTAAAGGTTCWTAYTTAAA	Phytobenthos	rbcL	312	(Vasselon et al., 2017, Zimmermann et al., 2021)
Diat_rbcL_708F_3	**AGGTGAAACTAAAGGTTCWTACTTAAA**	**Phytobenthos**	**rbcL**	**312**	**(Vasselon et al., 2017, Zimmermann et al., 2021)**
R3_1	**CCTTCTAATTTACCWACWACTG**	**Phytobenthos**	**rbcL**	**312**	**(Vasselon et al., 2017, Zimmermann et al., 2021)**

Primer name	Primer Sequence (5'-3')	Target group	Target gene	Amplicon size (bp)	Reference
R3_2	CCTTCTAATTTACCWACAACAG	Phytobenthos	rbcL	312	(Vasselon et al., 2017, Zimmermann et al., 2021)
F57	GTAATTCCATATGCTAAAATGGG	Rhodophyta	rbcL	Not stated	(Bringloe et al., 2019)
rbcLrevNEW	ACATTTGCTGTTGGAGTYTC	Rhodophyta	rbcL	Not stated	(Bringloe et al., 2019)
NDrbcL2	AAAAGTGACCGTTATGAATC	Phaeophyceae	rbcL	Not stated	(Bringloe et al., 2019)
NDrbcL8	CCAATAGTACCACCACCAAAT	Phaeophyceae	rbcL	Not stated	(Bringloe et al., 2019)
BIVALVE 1F	**TCTAGAGCTAATACATGC**	**Bivalves**	**18S rRNA gene**	**162–196**	**(Espiñeira et al., 2009)**
BIVALVE 1R	**ATAGGKCAGACAYTTGAAAG**	**Bivalves**	**18S rRNA gene**	**162–196**	**(Espiñeira et al., 2009)**
BIVALVE 2Fmod	AAATTAGAGTGYTCAAAGCAGGC	Bivalves	18S rRNA gene	148–151	(Espiñeira et al., 2009)
DtCed18S	CACCTCTCSCGCCGCARTACGT	Bivalves	18S rRNA gene	148–151	(Espiñeira et al., 2009)

2.4 Development of Reference Library

The downloaded databases (Section 2.3) were examined and sequences with uncertain taxonomic labels were removed, for example, sequences labelled ‘Environmental samples’, ‘Unassigned’ or ‘Synthetic’. Additionally, sequences associated with taxonomic groups that had been downloaded due to mislabelling in NCBI and belonged to a group that was of no interest to this study were removed, for example, ‘Archaea’, ‘Bacteria’ or ‘Nematoda’ as these were not taxa targeted by the primer sets. Taxonomic rank labels were also edited for accuracy and consistency (for example, if a taxonomic label consisted of ‘Kingdom Eukaryota, Phylum Eukaryota and class Phaeophyceae’ this was edited to ‘Kingdom Eukaryota, Phylum Ochrophyta and class Phaeophyceae’ as Eukaryota is not a Phylum). All sequences and taxonomic groups were then dereplicated, i.e. all identical sequence reads were combined into a unique sequence with a corresponding abundance. Sequences were removed if they contained degenerate bases (>1) and/or homopolymers (sequences of identical repeats, >8 bp).

Databases were then used to create QIIME2 amplicon-specific taxonomic classifiers to be used in the analysis. The chosen primers were used to generate an initial set of sequences from the raw database to create a primer extracted reference group. Any low-quality sequences were then removed. In order to expand our reference database, the original raw database sequences (the same sequence set from which we initially extracted our primer amplicon region) were queried back against the primer extracted reference sequences for any sequences that were >90% identity match (Robeson et al., 2018, Robeson et al., 2021).

The final set of sequences were used to construct the native bayes taxonomic classifier, which was then evaluated for optimal classification accuracy (F measure) (Robeson et al., 2021).

2.5 PCR amplification

Further experimental details can be found in the protocols in Appendix 1.2 – 1.5 and our Quality Assurance detailed in Section 2.9. Overhang adapter sequences were added to the locus-specific sequences as follows: Forward overhang: 5’-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-[locus-specific sequence]-3’ and Reverse overhang: 5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-[locus-specific sequence]-3’.

PCRs were done using the Platinum™ Hot Start PCR Master Mix (Invitrogen) according to manufacturer’s instructions. Each 25 µl reaction consisted of 12.5 µl Platinum™ Hot Start PCR Master Mix, 0.5 µl of 10 µM forward primer, 0.5 µM reverse primer, 6.5 µl nuclease-free water and 5 µl template DNA. This contains a high-fidelity DNA polymerase that has

low error rate and produces amplicons suitable for next generation sequencing. Initially, PCR optimisation was required through annealing temperature gradient PCR using DNA from the positive controls; four positive controls were run in duplicate. The conditions were: 94ºC for 2 min, followed by 40 cycles of 94ºC for 30 s, 50ºC / 53ºC / 55.9ºC / 59.5ºC / 62.5ºC / 64.1ºC for 30 s and 72ºC for 30 s, with a final extension step of 72ºC for 10 min. Subsequently, an optimal annealing temperature of 55ºC was chosen for both targets. Forty cycles were used due to low DNA yields. Positive control samples were run alongside the sediment samples consisting of soil, seaweed and bivalve DNA. Negative control samples consisted of molecular grade UltraPure Distilled Water, Dnase and Rnase Free (Invitrogen) in place of DNA . DNA extraction blanks were also run. All samples and controls (positive and negative) were amplified in triplicate and pooled by sample after PCR and prior to gel electrophoresis. PCR products (amplicons) were visualised by agarose gel electrophoresis using a 1.5% gel, 1.5% TBE buffer and 6X gel loading dye (New England Biolabs) for 45 min at 80 V.

Amplicons were cleaned using the Ampliclean magnetic bead PCR cleanup (Nimagen) according to the manufacturer's instructions to remove free primers, primer dimers and salts. According to Ampliclean instructions a ratio of 1.8x should recover and purify amplicons over 100 bp. A further round of PCR was conducted to attach dual indices using the Nextera XT Index Kit (Illumina) according to manufacturer's instructions, which included dual-index barcodes to allow samples to be multiplexed. The PCR conditions were: 95ºC for 3 mins, followed by 8 cycles of 95ºC for 30s, 55ºC for 30s, 72ºC for 30s, then a final extension step of 72ºC for 5 min. The amplicons were cleaned using the Ampliclean magnetic bead PCR cleanup at a ratio of 1.8x according to manufacturer's instructions. This ratio was chosen as Ampliclean states that amplicons over 100 bp will be recovered and purified whilst other contaminants such as primer dimers and dNTPs are removed. Indices were determined to be attached from successful qPCR during library preparation (Section 2.6). The libraries were normalised to 10 nM and then pooled in equimolar concentrations for sequencing.

2.6 Library preparation and sequencing

The pooled multiplexed amplicon libraries were quality checked using standard procedures in the Genomics Facility of the James Hutton Institute. Firstly, the library pools were quantified using qPCR (NEBNext Library Quant kit, NEB; as recommended) and Qubit 3 fluorimetry, and the integrity checked on a Bioanalyzer 2100 (Agilent). The high-quality library pools were subsequently diluted and combined with 20% PhiX Control v3 (Illumina) prior to loading on a MiSeq Reagent Kit v2 (500 cycles) at a final concentration of 6 pM, using manufacturer (Illumina) recommended procedures. Sequencing was performed using standard cycling conditions on an Illumina MiSeq, generating paired-end 250 bp reads, along with the associated dual index reads to enable deconvolution into a

pair of FASTQ format files for each sample. Details of our Quality Assurance are in Section 2.9.

2.7 Sequence analysis

All sequence data was analysed using the James Hutton Institute's high performance computing cluster. All scripts and classifiers are available at the GitHub (for more details see Section 2.10).

Sequence data was analysed using the QIIME2 platform (Bolyen et al., 2019). Raw data was imported into QIIME2 and visualised at the QIIME2 view website (https://view.qiime2.org). Quality plots were viewed where the reads were plotted by average quality score (based on 10,000 randomly selected reads) against the base pairs in the forward and then reverse reads and assessed for sequencing depth (per-sample sequence counts). For this proof-of-concept dataset, all samples with >10,000 reads were taken forward through the pipeline in order to retain as many samples as possible. This was to ensure quality and confidence in the dataset (as recommended in the QIIME2 tutorials available at https://docs.qiime2.org/2024.2/tutorials/filtering/).

Reads that did not contain primer sequences were discarded and subsequently, primers were removed from reads using the cutadapt plugin (Martin, 2011).The DADA2 plugin (Callahan et al., 2016) was then used to denoise the datasets. Denoising truncation (trimming from the 3' end of the read) was also performed and reads with errors were removed along with any remaining PhiX control (see Section 2.9) and singletons (reads that only appear once in a sample). Reads were dereplicated, merged and chimeric reads removed. Chimeric reads are formed during PCR when an incomplete amplicon primes the next amplification step, resulting in a spurious read that consists of a forward read of one biological entity and the reverse of another. Therefore, DADA2 uses the dataset itself to remove chimeric reads. These reads are identified by performing an alignment of each sequence to more abundant sequences searching for the left and right combinations (Callahan et al., 2016). Details of our Quality Assurance are in Section 2.9.

2.8 Carbon origin analysis via taxonomic assignment

The two custom databases were trained to create taxonomic classifiers using a I Bayes algorithm (q2-feature-classifier) (Bokulich et al., 2018). The rbcL and the 18S rRNA gene databases were trained on the exact region sequenced by trimming the sequences inside the databases using our primer sequences, so that the database only contained the targeted amplicon region (feature-classifier extract-reads). When a I Bayes classifier is trained on the region of the target sequences, it can improve data classification (Werner et al., 2012). Taxonomy was then assigned to the sequences using the amplicon specific classifier (classify-sklearn). After taxonomic assignment, read abundance tables were

generated in Microsoft Excel (including confidence values and number of reads), and a proportion of identified taxa was sense checked against the National Biodiversity Network Trust Atlas. Details of our Quality Assurance are in Section 2.9.

2.9 Quality assurance

The labs are certified to ISO9001 standard. Moreover, the labs are compartmentalised so that DNA extractions (considered ‘dirty’ lab work) were done in a different room to subsequent steps (i.e. PCR and amplicon library preparation, considered ‘clean’ lab work). Benches and other workspaces (e.g. hoods and work cabinets) were cleaned with disinfectant (50% Chemgene) and DNAse and RNAse inhibitors. Biological safety cabinets were additionally disinfected with ozone. Plastic consumables were certified DNAse and RNAse free. Pipettes are regularly serviced to ensure accuracy. Labs of the Genomics Facility are designated for sole use of staff within the facility. Equipment for QC and sequencing are routinely serviced as required, by either the original supplier or certified service companies. Controls and standards were used throughout to ensure confidence in the methods and quality of the data.

For molecular work, quality checks were done at each stage to ensure the integrity of the results. To account for sediment heterogeneity, the samples were extracted in triplicate. One DNA extraction blank consisting of sterile water was extracted alongside each batch of samples, and each batch consisted of between 8 and 22 samples (see Section 2.2 for details). None of the extraction blanks contained significant concentrations of DNA, giving confidence that there was no contamination or cross-contamination of the samples during extraction. A positive control sample (known soil) was also extracted alongside the samples, which gave reasonable DNA yields (12.9 – 67.7 ng/µl), giving confidence that the DNA extraction methods were working. Quality checks of the extracted eDNA included measurements of yield and quality (260/280 ratio). Normally, if the DNA yield is not sufficient (<10 ng/µl), then repeat extractions are done. However, repeated extractions did not improve the yield and, in most cases, the sediment samples were exhausted so no further extractions could be done. It was therefore assumed that the samples contained low biomass, and the low concentrations of eDNA were used for subsequent analysis anyway. Moreover, when the quality of the DNA is poor (i.e. 260/280 ratio is not 1.8 +/- 0.2), then the DNA is usually further cleaned using an appropriate kit. However, due to the low DNA concentrations, it was not possible to obtain an accurate reading of the 260/280 ratio, and further cleaning was deemed inappropriate as this would have further reduced DNA yields. Replicate DNA extractions were pooled and stored at -20ºC. Freeze-thawing cycles were avoided to reduce DNA degradation.

PCRs were conducted in a separate room to DNA extraction, and within a PCR cabinet, which has its own dedicated set of pipettes to avoid contamination. Plastic consumables and some reagents (e.g. PCR-grade water) were treated with UV for 15 min to remove exogenous contaminant DNA. Positive controls were run alongside the samples,

consisting of plant and bivalve tissue DNA. Negative control samples consisting of sterile UltraPure Distilled Water, Dnase and Rnase Free (Invitrogen) and the DNA extraction blanks were also run. Triplicate technical replicates for each sample were run and pooled, and PCR products were checked by agarose gel electrophoresis using a 1.5% gel, 1.5% TBE buffer and 6X gel loading dye (New England Biolabs) for 45 min at 80 V.

A modified version of the standard recommended Illumina Metagenomics Sequencing Library protocol (15044223 Rev. B) was used, and indexing of samples allowed sample multiplexing prior to sequencing since amplicons were of similar size. Amplicon library preparation was done in the PCR flow hood to avoid contamination. Amplicons were cleaned after both initial PCR and indexing PCR using the Ampliclean Magnetic bead PCR cleanup (Nimagen), and then quantified using a Qubit Flex fluorometer and the dsDNA High Sensitivity assay to ensure that DNA was not lost during clean-up.

Sequencing was conducted at the Genomics Facility of the James Hutton Institute, allowing better control over the quality of the sequencing. Further quality checks of the library were performed. Firstly, the multiplexed library was quantified using qPCR and Qubit fluorimetry, and then integrity checked on a Bioanalyzer 2100 (Agilent). Library profiles were compared to expected size ranges to ensure there was no unusual size distribution. The high-quality library pool was diluted and combined with 20% control library PhiX (Illumina) to ensure that increased base diversity was introduced for efficient sequencing. The final library pool was run on a MiSeq v2 500 bp kit at a final concentration of 6 pM. Routine checks were carried out during the sequencing run to ensure the expected quality of the sequencing data was obtained.

Raw sequence data was backed up and stored on the James Hutton Institute's high performance computing cluster. Bioinformatics quality check was performed on all raw data to ensure data were of sufficient depth and quality for profiling the communities from the amplicon sequencing. Raw data was assessed for the number of sequences in each sample and quality for the forward and reverse reads. This information was then used to remove samples that did not contain a sufficient number of reads. For this study, we lowered the threshold to 10,000 reads to allow more samples to pass quality control, however, a minimum of 50,000 read is usually recommended. Quality parameters allowed the appropriate trimming of low-quality ends (where the quality falls below Q20) from reads as well as removing reads with excessive error rates (>2 errors per read). Additionally, the merging of paired-end reads allowed reconstitution of the full amplicon, as well as give confidence in the accuracy of the reads. Given that there are pre-existing biases in available sequence databases, with some orders having significantly fewer genomic resources, species allocation was restricted. As noted by others who have completed similar studies (Ortega et al., 2020), it is not possible to identify all sequences down to species level; however, organisms were identified to the lowest taxonomic classification possible. The level of confidence in identification of organisms was enhanced by increasing the stringency of the match but allowing some ambiguity for sequences with few genomic sequences in the databases.

A second analysis of the raw Illumina sequencing data was performed with an alternative pipeline to confirm the main findings using QIIME2, which had required considerable optimisation in the read processing. Specifically, THAPBI PICT version v1.0.11 (Cock et al., 2023) was used, which starts with merging overlapping reads with Flash v1.2.11 (Magoc and Salzberg, 2011) and then primer identification, trimming and demultiplexing with Cutadapt v4.5 (Martin, 2011). This produces amplicon sequence variant (ASV) or unique amplicon sequence counts per sample, and was done with and without read-correction using the UNOISE algorithm (Edgar, 2016).

2.10 Open science and data availability

All scripts are available on GitHub at https://github.com/HuttonICS/blue-carbon-db. This page provides instructions on how to install, download and build the databases, process the data, and classify taxonomy alongside the scripts used to analyse the datasets. Additionally, links to the custom taxonomic classifiers in the QIIME2 format are available via the GitHub repository.

QIIME2 has automated provenance tracking, which aids reproducibility by viewing the provenance replay of QIIME2 files (Keefe et al., 2023). Therefore, the information of how classifiers were made is not only available via the scripts but also by placing the classifiers into QIIME2 view website (https://view.qiime2.org/), where users can view the provenance graphs and text.

3 Results

3.1 Reference Library

Database downloads resulted in 63,535 sequences across 49,023 taxonomic groups for the rbcL gene database and 3,558 sequences across 2,972 taxonomic groups for the 18S rRNA gene database. These numbers were obtained in October 2023; a more up to date download may result in different numbers as more data is uploaded to NCBI.

For the rbcL gene, primer set Diat_rbcL_708F_3 & R3_1 for rbcL (Vasselon et al., 2017, Zimmermann et al., 2021) performed best, retaining a larger number of species and sequences from the database as well as providing an appropriate fragment length for sequencing (Table 3). For the 18S rRNA gene database, primer set BIVALVE 1F & BIVALVE 1R (Espiñeira et al., 2009) was the most appropriate (Table 4).

Table 3. *In silico* digest of rbcL database to determine the most appropriate primer set for the rbcL gene target.

Forward primer (5'-3')	Reverse primer (5'-3')	Reference	No. unique sequences	No. taxa (at species level)	Amplicon fragment length (bp)
F – GCGGGTGTTAAAGAGTACAA	R – AGTAGAAGATTCGGCAGCTA	(Ortega et al., 2019)	23,671	45,141	Two fragments of <150 and >300
Z1aF – ATGTCACCACCAACAGAGACTAAAGC	R604 – CTGRGAGTTMACGTTTTCATCATC	(Reef et al., 2017)	39,864	52,026	560
F52_tag - GTTGGATTCAAAGCTGGTGTTA	rcblB_tag – AACCYTCTTCAAAAAGGTC	(Reef et al., 2017)	26,808	36,345	720
F57 – GTAATTCCATATGCTAAAATGGG	rbcLrevNEW – ACATTTGCTGTTGGAGTYTC	(Bringloe et al., 2019)	11,748	12,113	1323
NDrbcL2 – AAAAGTGACCGTTATGAATC	NDrbcL8 – CCAATAGTACCACCACCAAAT	(Bringloe et al., 2019)	15,699	16,613	1100

Forward primer (5'-3')	Reverse primer (5'-3')	Reference	No. unique sequences	No. taxa (at species level)	Amplicon fragment length (bp)
Diat_rbcL_708F_1 – AGGTGAAGTAAAAGGTTCWTACTTAAA	R3_1 – CCTTCTAATTTACCWACWACTG	(Vasselon et al., 2017)	38,129	50,617	Two fragments of <400 and >800
Diat_rbcL_708F_2 – AGGTGAAGTTAAAGGTTCWTAYTTAAA	R3_2 – CCTTCTAATTTACCWACAACAG	(Vasselon et al., 2017)	39,333	51,700	Two fragments of <400 and >800
Diat_rbcL_708F_3 – AGGTGAAACTAAAGGTTCWTACTTAAA	R3_1 – CCTTCTAATTTACCWACWACTG	(Vasselon et al., 2017)	31,787	53,609	266

Table 4. *In silico* digest of Mollusc database to determine the most appropriate primer set for the 18S rRNA gene target.

Forward primer (5'-3')	Reverse primer (5'-3')	Reference	No. unique sequences	No. taxa (at species level)	Amplicon fragment length (bp)
BIVALVE 1F – TCTAGAGCTAATACATGC	BIVALVE 1R – ATAGGKCAGACAYTTGAAAG	(Espiñeira et al., 2009)	1,776	2520	133
BIVALVE 2Fmod – AAATTAGAGTGYTCAAAGCAGGC	DtCed18S – CACCTCTCSCGCCGCARTACGT	(Espiñeira et al., 2009)	867	2576	106

Classifiers were then evaluated for optimal classification (F-measure). This simulates best possible classification accuracy when the true label is known but classification accuracy may be confounded by other similar hits in the database. The F-measure can be thought of as a classification accuracy measure i.e. the ability of a classifier to classify sequences to taxa. The highest measure of an F-measure value is 1, which would indicate perfect accuracy, while an F-measure value of 0 would indicate that the classifier has no ability to recall taxa for a sequence accurately.

Both databases resulted in lower F-measure values at species level indicating that classifying reads to species level may be less accurate (Table 5).

Table 5. F-measure values at different taxonomic levels for the rbcL and 18S rRNA databases.

Level	Level analogous to taxa description	F-measure for rbcL database	F-measure for 18S rRNA database
1	Kingdom	1.000	1.000
2	Phylum	1.000	1.000
3	Class	0.999	0.999
4	Order	0.997	0.987
5	Family	0.990	0.937
6	Genus	0.855	0.832
7	Species	0.498	0.638

3.2 DNA extraction efficiency

DNA was initially extracted using the CTAB method; however, the DNA yields were low (an average of 0.3 ng/µl for 45 site B samples). A trial of the Dneasy PowerSoil Pro DNA extraction kit gave similar results (an average of 0.37 ng/µl for 14 site B samples) and was therefore used for further extractions since it is a less labour-intensive method. This gave an average yield of 0.56 ng/µl across 114 samples. There was a lower DNA yield from site B than site C (an average of 0.17 ng/µl over 90 site B samples vs an average of 2.02 ng/µl over 24 site C samples). DNA yields from the soil positive controls were all high (an

average of 43.07 ng/µL for 5 soil positive controls), so we do not believe there to be issues with the methods.

Consequently, the Dneasy PowerMax Soil DNA extraction kit was used to extract DNA from the remaining site C samples because this method is recommended for low biomass samples (Qiagen, 2023). This improved the DNA yield because the DNA is eluted in a much larger volume (5 ml rather than 50 µl for the other methods), but concentrations were still low (an average of 2.36 ng/µl for 16 site C samples).

Overall, in 63% of cases, DNA concentrations were lower than the negative controls (an average of 0.18 ng/µl across 8 negative controls). The 260/280 ratios were outside the acceptable range of 1.8 ± 0.2, but this is because it is difficult to get an accurate reading if the DNA content is low.

Taken together, DNA yields were low, likely due to low biomass in the samples rather than the methods used. For low biomass samples, the Dneasy PowerMax Soil DNA extraction kit gave higher DNA yields (Table 6).

Table 6. Results of DNA extraction using the three DNA extraction methods.

Sample	Method used	No. samples	DNA yield range (ng/ul)	Average DNA yield (ng/ul)	260/280 ratio range	Average 260/280 ratio
B	CTAB	45	0.11 – 1.79	0.3	-15.5 – 22.98	1.01
B	PowerSoil	90	0.0002 – 2.13	0.17	-39.58 – 142.61	3.06
C	PowerSoil	24	0.01 – 9.25	2.02	0.94 – 4.03	1.74
C	PowerMax	16	0.16 – 5.34	2.36	-4.69 – 38	3.42
Positive controls	CTAB	1	NA	33.1	NA	11.64
Positive controls	PowerSoil	5	20.6 – 67.7	54.84	1.83 – 1.89	1.87
Positive controls	PowerMax	2	12.9 – 14.4	13.65	1.99 – 2.04	2.02

Sample	Method used	No. samples	DNA yield range (ng/ul)	Average DNA yield (ng/ul)	260/280 ratio range	Average 260/280 ratio
Negative controls	CTAB	2	0.11 – 0.18	0.14	0.01 – 21.79	10.89
Negative controls	PowerSoil	6	0.004 – 0.35	0.19	-0.76 – 1.56	0.26
Negative controls	PowerMax	2	0.16 – 3.67	1.91	-16.06 - -0.12	-8.09

3.3 PCR Positive Controls

Several types of plant material were trialled as potential positive controls. PCR amplification of the rbcL gene using our selected primer set showed that genomic DNA from strawberry fruit and strawberry leaf, grass (potential *Holcus sp.*), spider plant leaf (*Chlorophytum comosum*) and a green leaf taken from an unknown tree were not good controls as they showed no amplification. This is despite the fact that *in silico* analysis demonstrated that these primers could detect terrestrial plants in our rbcL database (Section 3.1). However, the seaweed mix (Dulse, 'ocean greens' and Nori) and tropical fish food (Bloodworms (family Glyceridae), Mysis (Mysidae), Daphnia (Daphniidae), Cyclops (Cyclopidae), Brine shrimp/Artemia (Artemiidae), and algae (unknown family)) amplified successfully, however, only the DNA from the seaweed mix (Dulse, 'ocean greens' and Nori) was used as the rbcL positive control. This gave an expected amplicon size of ~300 bp (Vasselon et al., 2017).

Genomic DNA from Chilean mussel (*Mytilus chilensis*) amplified strongly for the 18S rRNA gene using the selected primer set. DNA from mussel was subsequently used as the positive control throughout. This gave an expected amplicon size of ~200 bp (Espiñeira et al., 2009).

3.4 PCR amplification

Gradient PCR was employed to find the optimal annealing temperature, which showed that 55ºC was the most appropriate annealing temperature for both rbcL and 18S rRNA gene. Moreover, 40 PCR cycles were used instead of 30 because of the low DNA yields. It

has been shown that higher PCR cycles allows greater coverage from samples with low biomass (Witzke et al., 2020).

From the 54 samples, 23 samples gave amplicons for rbcL (1 for site B and 22 for site C) and 15 samples gave amplicons for 18S rRNA gene (10 for site B and 5 for site C). The positive controls showed good amplification of these genes, therefore increasing confidence that samples that failed to give amplicons were because they either did not contain those genes or they were present below the detection limit of the PCR assay. The majority of the positive results came from surface samples. Only five samples from site C2 from 20 – 30 cm and two samples from site C1 from >43 cm gave positive results (for rbcL only). Only four samples (1 from site B1 and 1 from each of the sites C1-C3) gave positive results for both target genes. The amplicons were about 50 bp larger than the expected sizes due to the addition of the sequencing adapters.

3.5 Library preparation

During library preparation, DNA from two site C samples for the18S rRNA gene amplicon were lost, likely from the PCR clean-up step. Since DNA concentration of the amplicons werelow, these were not taken forward for sequencing. In total, 23 rbcL amplicons (1 for site B and 22 for site C) and 13 18S rRNA gene amplicons (10 for site B and 3 for site C) were pooled for sequencing.

3.6 Sequencing

The final library pool profile assessed using the Bioanalyzer 2100 (Agilent) was as expected (Fig 1), with two discrete bands of c. 490 bp and c. 310 bp, corresponding to the rbcL and 18S rRNA gene amplicons, respectively, including sequencing adapters. The baseline was flat and no small fragments visible, meaning that the library pool was suitable for sequencing. Accurate quantification of the library pool was measured at 2.8 nM using Qubit fluorimetry and confirmed by qPCR, prior to dilution for sequencing. PhiX control library was added to 20% to increase nucleotide diversity, and the final library was loaded at 6 pM and run on a 500 cycle v2 kit on the MiSeq using recommended settings for paired-end 250 bp sequencing.

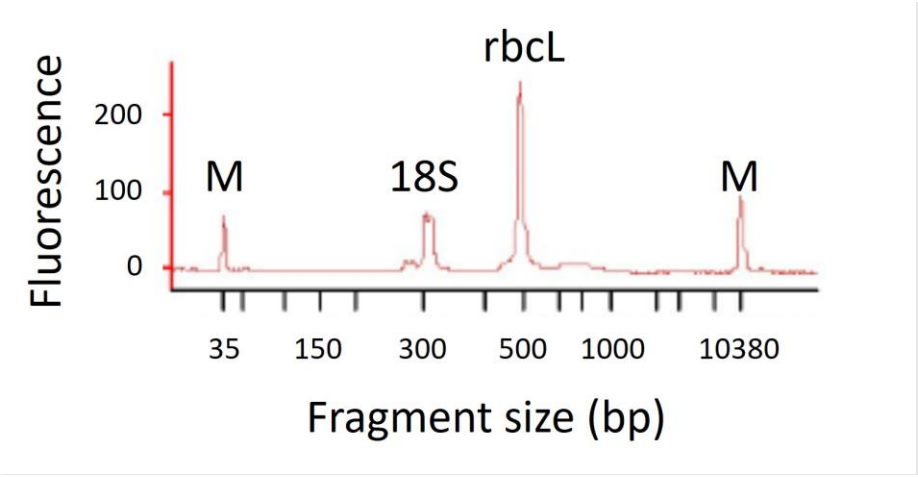

Figure 1. Library pool profiled on a Bioanalyzer 2100: two discrete bands of c. 490 bp and c. 310 bp were visible, corresponding to the rbcL and 18S rRNA gene amplicons, respectively. M, markers.

A total of 13.8M raw paired-end reads were generated, of which 8.2M passed quality filters. Quality, as measured by Q30 algorithm, was 82% overall (Fig 2), which is high for this type of sequencing. On-board deconvolution of data assigned reads to individual samples based upon their indexing (Fig 3). There was clear uneven balance of reads between samples, but this can be explained by low concentration amplicons being difficult to quantify accurately and subsequently normalise. Data was exported in FASTQ format for downstream processing.

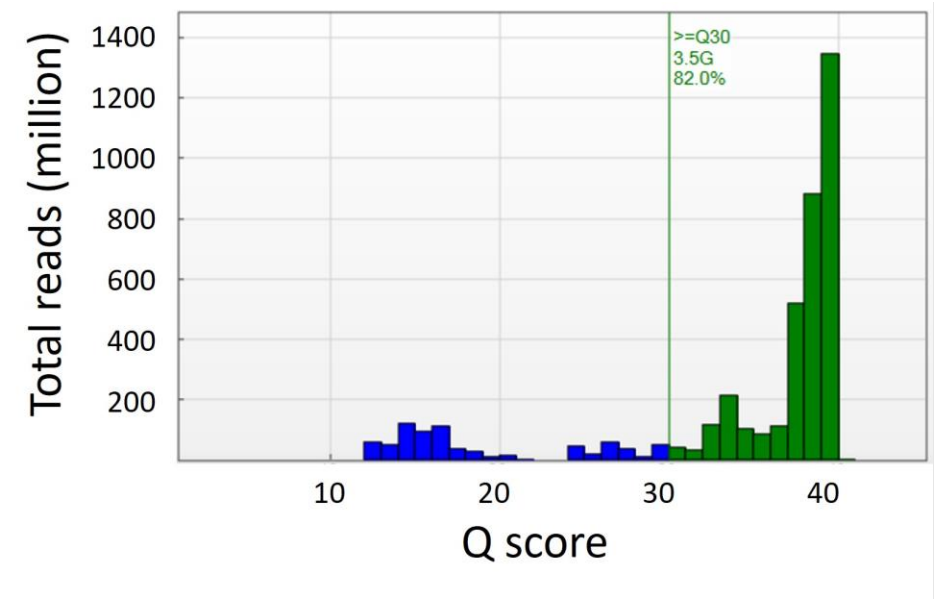

Figure 2. Quality (Q) scores of raw sequencing reads. Overall, reads with high-quality data (>Q30) comprised 82%.

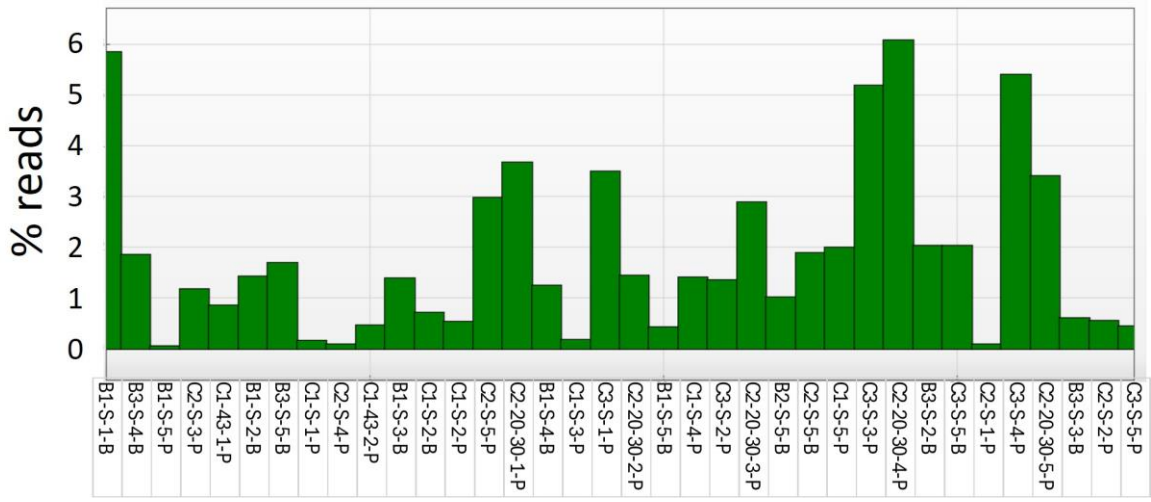

Figure 3. Proportion of raw sequencing reads from each sample. There is a clear uneven balance of reads between samples.

3.7 Sequence filtering

3.7.1 Sequence processing for rbcL gene

Raw sequence numbers for the rbcL targeted samples ranged from 4406 to 497,076 reads per sample (Table 7). Samples with read counts <10,000 were removed from further analysis before denoising, which removed three samples. Filtering (including all primer filtering, and denoising) resulted in retention on average of 83% of reads per sample (Table 7).

Table 7. Read numbers from the rbcl gene amplicon data before (raw read count input) and after (filtered reads output) denoising and filtering.

Sample ID	Description	Raw read count input	Filtered reads output	% reads recovered after filtering
C1-S-1-P	Site C1, surface sample, sample 1 of 1, positive for rbcL	13,215	11,970	90.58
C1-S-2-P	Site C1, surface sample, sample 2 of 5, positive for rbcL	43,806	39,078	89.21

Sample ID	Description	Raw read count input	Filtered reads output	% reads recovered after filtering
C1-S-3-P	Site C1, surface sample, sample 3 of 5, positive for rbcL	15,518	13,662	88.04
C1-S-4-P	Site C1, surface sample, sample 4 of 5, positive for rbcL	114,051	100,753	88.34
C1-S-5-P	Site C1, surface sample, sample 5 of 5, positive for rbcL	161,770	143,776	88.88
C1-43-1-P	Site C1, depth of 43 cm, sample 1 of 5, positive for rbcL	67,230	61,410	91.34
C1-43-2-P	Site C1, depth of 43 cm, sample 2 of 5, positive for rbcL	37,722	34,421	91.25
C2-20-30-1-P	Site C2, depth of 20-30 cm, sample 1 of 5, positive for rbcL	293,959	233,028	79.27
C2-20-30-2-P	Site C2, depth of 20-30 cm, sample 2 of 5, positive for rbcL	116,683	97,621	83.66
C2-20-30-3-P	Site C2, depth of 20-30 cm, sample 3 of 5, positive for rbcL	231,199	182,630	78.99
C2-20-30-4-P	Site C2, depth of 20-30 cm, sample 4 of 5, positive for rbcL	488,400	380,935	78
C2-20-30-5-P	Site C2, depth of 20-30 cm, sample 5 of 5, positive for rbcL	272,821	218,214	79.98
C2-S-2-P	Site C2, surface sample, sample 1 of 5, positive for rbcL	44,943	40,002	89.01
C2-S-3-P	Site C2, surface sample, sample 3 of 5, positive for rbcL	81,364	68,968	84.76

Sample ID	Description	Raw read count input	Filtered reads output	% reads recovered after filtering
C2-S-5-P	Site C2, surface sample, sample 5 of 5, positive for rbcL	233,843	185,707	79.42
C3-S-1-P	Site C3, surface sample, sample 1 of 5, positive for rbcL	277,933	223,110	80.27
C3-S-2-P	Site C3, surface sample, sample 2 of 5, positive for rbcL	103,894	57,690	55.53
C3-S-3-P	Site C3, surface sample, sample 3 of 5, positive for rbcL	416,032	319,118	76.71
C3-S-4-P	Site C3, surface sample, sample 4 of 5, positive for rbcL	430,807	335,870	77.96
C3-S-5-P	Site C3, surface sample, sample 5 of 5, positive for rbcL	34,492	28,706	83.23
B1-S-5-P	Site B1, surface sample, sample 5 of 5, positive for rbcL	4,406	NA	NA
C2-S-1-P	Site C2, surface sample, sample 1 of 5, positive for rbcL	7,284	NA	NA
C2-S-4-P	Site C2, surface sample, sample 4 of 5, positive for rbcL	7,042	NA	NA

3.7.2 Sequence processing for mollusc 18S rRNA gene

Raw sequence numbers for the mollusc targeted samples ranged from 36,103 and 478,134 reads per sample (Table 8). All samples met the minimum threshold of >10,000 read counts and were taken through to denoising. Filtering (including primer filtering, and denoising) resulted in retention on average of 34% of reads per sample (Table 8). However, the denoising process had variable results and four samples resulted in <10% of

reads being retained for downstream analysis. Samples with <10,000 reads were removed from further analysis before taxonomic assignment.

Table 8. Read numbers from the Mollusc 18S rRNA gene amplicon data before (Raw read count input) and after (filtered reads output) denoising and filtering.

Sample ID	Description	Raw read count input	Filtered reads output	% reads recovered after filtering
B1-S-1-B	Site B1, surface sample, sample 1 of 5, positive for 18S rRNA gene	461,339	216,681	46.97
B1-S-2-B	Site B1, surface sample, sample 2 of 5, positive for 18S rRNA gene	114,008	5,396	4.73
B1-S-3-B	Site B1, surface sample, sample 3 of 5, positive for 18S rRNA gene	111,154	65,805	59.20
B1-S-4-B	Site B1, surface sample, sample 4 of 5, positive for 18S rRNA gene	99,624	7,247	7.27
B1-S-5-B	Site B1, surface sample, sample 5 of 5, positive for 18S rRNA gene	34,023	14,317	42.08
B2-S-5-B	Site B2, surface sample, sample 5 of 5, positive for 18S rRNA gene	80,933	5,045	6.23
B3-S-2-B	Site B3, surface sample, sample 2 of 5, positive for 18S rRNA gene	162,440	78,367	48.24
B3-S-3-B	Site B3, surface sample, sample 3 of 5, positive for 18S rRNA gene	48,551	34,009	70.05
B3-S-4-B	Site B3, surface sample, sample 4 of 5, positive for 18S rRNA gene	143,142	57,453	40.14

Sample ID	Description	Raw read count input	Filtered reads output	% reads recovered after filtering
B3-S-5-B	Site B3, surface sample, sample 5 of 5, positive for 18S rRNA gene	134,454	25,954	19.30
C1-S-2-B	Site C1, surface sample, sample 2 of 5, positive for 18S rRNA gene	56,267	25,159	44.71
C2-S-5-B	Site C2, surface sample, sample 5 of 5, positive for 18S rRNA gene	150,020	11,867	7.91
C3-S-5-B	Site C3, surface sample, sample 5 of 5, positive for 18S rRNA gene	162,368	84,963	52.33

3.8 Taxonomic assignment

Taxonomic profiles for both rbcL and 18S rRNA gene are presented using relative abundance at several “levels”. To reduce noise in the graphs and for ease of viewing taxonomic profiles, any taxa that were present at a level < 1% relative sequence abundance across all samples have been removed. However, those at low abundance were examined when searching the data for the presence or absence of taxa in all read assignments. While taxonomic profiles are presented at specific levels (i.e. Order level 4 or Genus level 6), some classifications will be at higher taxonomic levels due to the classifier only assigning taxonomy to a lower level if confidence values were >0.7. It should be noted that the data is based on read abundance, which does not necessarily reflect organism abundance due to variation in gene copy numbers within different organisms.

3.8.1 Macrophyte (rbcL) taxonomy

The taxonomic profiles were similar across samples (Figs 4 and 5) and were dominated by diatoms. At level 4, analogous to Order, taxonomic profiles resulted in 7 major groups, but one taxonomic group was classified only to Class (Bacillariophyceae) (Fig 4). At level 6, analogous to Genus, the taxonomies delineated out into 11 groups (Fig 5). Both graphs show community compositions of different orders and genera of diatoms, respectively. Detailed rbcL amplicon sequence variant (ASV) tables have been provided in a separate spreadsheet called “Macrophyte (rbcl) gene taxonomy”.

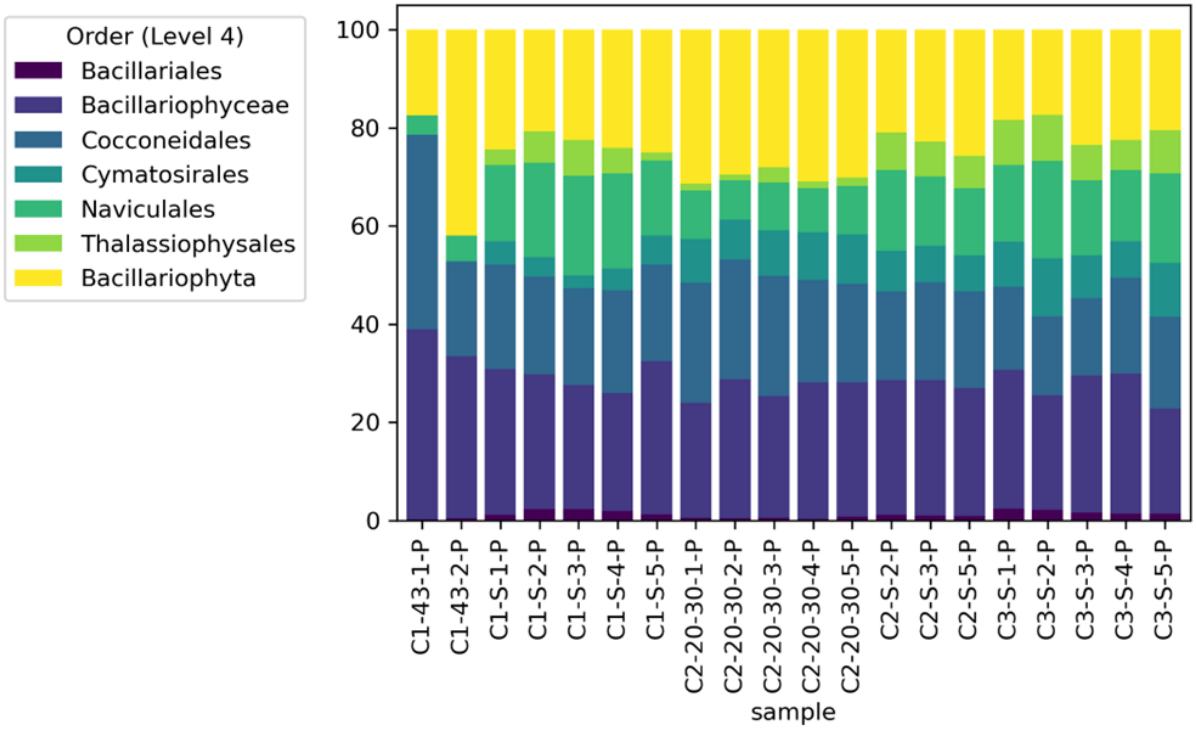

Figure 4. A stacked bar chart of relative abundance of groups from the rbcL sequencing at the level 4, analogous to Order.

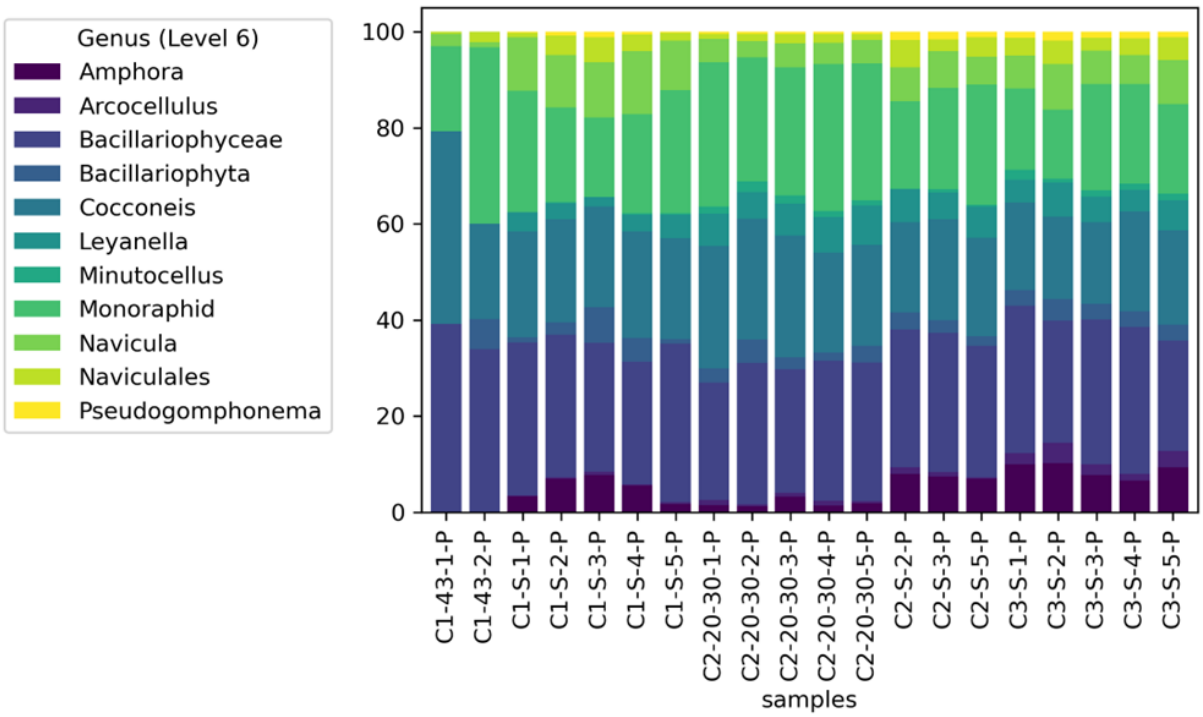

Figure 5. A stacked bar chart of relative abundance of groups from the rbcL sequencing at the level 6, analogous to Genus.

The full set of read assignments (including those of relative abundance <1%) were used to confirm the presence of specific species. To ascertain if any non-native species were present, the dataset was filtered for any matches against the UK Marine Non-Indigenous Species Priority List supplied by Natural England (UKNNSS, 2020). No matches were found between the rbcL taxonomic list, and the Non-Indigenous Species Priority List and all taxa found were marine or freshwater taxa, i.e. no terrestrial taxa were discovered in this dataset.

3.8.2 Mollusc (18S rRNA gene) taxonomy

The mollusc dataset resulted in an uneven spread of taxonomy across samples and many reads were assigned only at the two highest levels in the database, ‘Metazoa’ or ‘Mollusc’ (Fig 6). Level 4 (Order) resulted in 5 groupings, including the targeted Bivalves and Gastropods (Fig 6).

To further investigate the dataset, the commonly employed SILVA database (Quast et al., 2013) was used as a classifier. Taxonomic classification using SILVA resulted in many of the sequences that the custom classifier only classified as far as Metazoa or Mollusc being assigned to worm groups (Nematoda, Annelida) as well as Arthropods. Some remained difficult to assign and are labelled as ‘Eukaryota’ (Fig 7). Detailed 18S rRNA gene ASV tables have been provided in a separate spreadsheet called “Mollusc 18s rRNA gene taxonomy”.

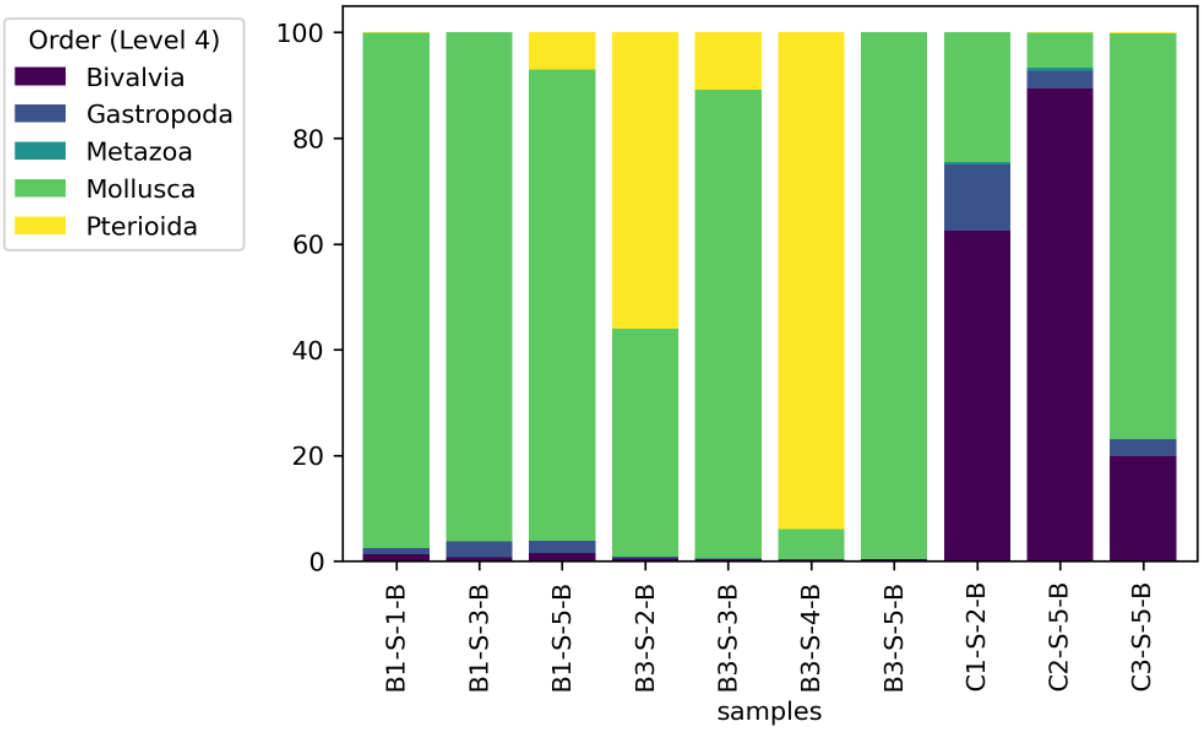

Figure 6. A stacked bar chart of relative abundance of groups from the Mollusc 18S rRNA gene sequencing at the level 4, analogous to Order.

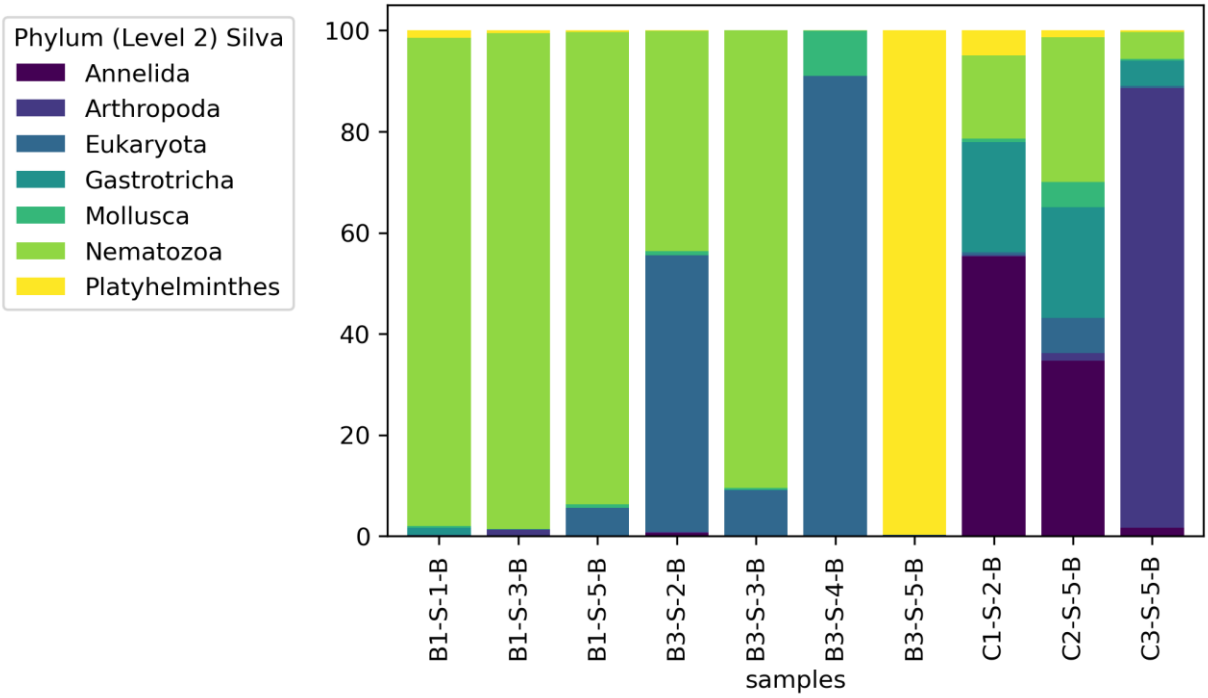

Figure 7. Relative abundance of Mollusc 18S rRNA gene dataset with taxonomy assigned using SILVA database viewed at level 2 (Phylum).

The record of any UK marine non-native species was explored by comparing the UK Marine Non-Indigenous Species Priority List (UKNNSS, 2020) and the Mollusc 18S rRNA gene dataset assigned using the custom database.

One match within the genus *Ensis* was found. However, taxonomic assignment was only possible to genus level and identification to species level was not possible. Therefore, it remains unconclusive as to whether the match corresponds to the American jack knife clam (*Ensis leei*) or the various native species within the same genus. The SILVA taxonomic assignments were also checked against the UK Marine Non-Indigenous Species Priority List and there were no matches.

3.8.3 Sense checking QIIME pipeline and taxonomic assignments

The alternative analysis with THAPBI PICT accepted around 55% of the raw reads for the 18S rRNA gene at the default abundance threshold of 100 copies of any unique sequence, but only 43% of the rbcL gene reads. The marker discrepancy was in line with the QIIME2 results. There were low levels of 18S rRNA gene amplicons in the rbcL reads, suspected to be due to Illumina “tag switching” or “tag jumping” (noise in the multiplexing protocol), perhaps exacerbated by some samples having an order of magnitude more reads than others, or from the inherent risk of background contamination when working with high amplification of trace level samples. The patterns did not suggest simple cross-sample contamination. After denoising, both markers gave around 100 unique sequences, with similar distributions over the samples, again in line with the QIIME2 results. The classifiers included in THAPBI PICT are focused on near-identical matches for species/genus assignment, which was unsuitable for this dataset where most of the sequences are unknown with no close match in the NCBI Nucleotide database. Inspection of the top sequences with NCBI BLAST against the Nucleotide database identified good matches to nematode and flatworm 18S rRNA genes, and diatom rbcL matches, again in line with the QIIME2 results.

4 Discussion

The DNA yields obtained in this experiment were low, with more than half the samples giving yields lower than the negative controls. This was likely to be a consequence of the type of sediments collected rather than the methods used because the positive extraction controls (soil) gave high DNA yields. The low DNA yield was potentially due to the samples being sandy, which is difficult to extract DNA from due to a low organic matter content (Qiagen, 2023). Sediments with higher organic matter content may have yielded more reliable results. There was some improvement in yield when the Dneasy PowerMax Soil DNA extraction kit was used as this is recommended for low biomass samples,

however, it should be noted that only about 3 g of sediment was used for extraction because that was the maximum amount of sediment provided to us from each site, but the kit recommends 10 g of material. Larger amounts of sediment would likely have increased the yield further.

Biases exist between different DNA extraction methods (van der Loos and Nijland, 2021), which should be investigated for different matrices (e.g. sand, mud, clay) so that appropriate methods are used. Nevertheless, low DNA yields present further biases generating lower coverage of the community such that species may be missed. It is possible that samples that failed to give amplicons were because the low DNA yields meant that any gene targets present were below the detection limit of the PCR. It can also lead to higher stochasticity during PCR amplification, which results in community compositions that deviate from the actual community compositions (Biesbroek et al., 2012). We attempted to alleviate both biases by conducting triplicate DNA extractions and pooling, followed by triplicate PCRs and pooling. In addition, we achieved a high sequencing depth, which is recommended to accurately capture community composition (Smith and Peay, 2014). An additional issue with low DNA yield samples is the potential for contaminant OTUs to be overrepresented in the final analysis (Karstens et al., 2019). This can be alleviated in the taxonomic assignment stage where clear contaminant ASVs can be ruled out. Many 18S rRNA gene sequences were classified as ‘Eukaryota’ and this group would be likely to contain the contaminant sequences if any were present. However, due to low taxonomic resolution, we were not able to identify any.

The rbcL gene was selected for characterisation of the macrophyte community because it is an agreed plant DNA barcode by the Consortium for the Barcode of Life, which offers high coverage of plant and algal groups, with an extensive GenBank reference sequence library (Reef et al., 2017). Primer sets were searched for in the literature and validated against the custom database. The chosen primer set showed good coverage of both terrestrial and marine/coastal macrophytes, while also providing a suitable amplicon length for sequencing. Nevertheless, the selected rbcL primers failed to amplify the target gene from several plant materials as positive controls (strawberry fruit and strawberry leaf, grass (*Holcus sp.*), spider plant (*Chlorophytum comosum*) and a green leaf taken from an unknown tree), but they successfully amplified the target gene from a seaweed mix. This suggests that the primers may be biased against terrestrial plants, preferentially amplifying seaweed, as demonstrated by the positive controls. Moreover, diatoms and algae dominated the sequencing dataset as confirmed by both the QIIME2 and THAPBI PICT approaches. This result is reasonable given that diatoms are abundant in marine and freshwater environments, particularly in surface sediment such as the samples used in this study (Fukai et al., 2022), while terrestrial species are in the minority. However, it could be again reflecting biases associated with the selected primers.

Similarly, the primer set chosen to characterise molluscs showed good coverage of bivalves and gastropods *in silico*, while also providing a suitable amplicon length for sequencing. Amplification of the positive control (Chilean mussel, *Mytilus chilensis*) was

also strong. Classification of the Mollusc 18S rRNA gene sequences using the custom classifier resulted in many sequences being classified as Metazoa and Mollusc. However, these classifications were largely uninformative because all sequences in the custom classifier were under these taxonomic headings. Effectively, a classification of 'Mollusc' or 'Metazoa' suggests that the sequences have a similarity to 18S rRNA gene sequences with > 0.7 confidence in the custom database. When examined using the SILVA database, the sequences had a higher confidence match to several worm taxa or Arthropoda taxa than they did to Mollusc sequences in the SILVA database. This suggests that the sequences were in fact more likely to belong to worm groups or Arthropoda. This was confirmed by the THAPBI PICT approach, which also identified good matches to nematode and flatworms. This could again suggest bias in the primer set, preferentially picking up worms, despite verifying *in silico* that they pick up bivalves and gastropods. On the other hand, it could suggest that bivalves and gastropods were absent in the samples. Some sequences were classified as Eukaryotes by the SILVA database, effectively leaving the sequences unassigned. The poor performance of the Mollusc 18S rRNA gene dataset could be the result of low input DNA, as well as a lack of primer specificity.

It should be noted that metabarcoding amplifies target genes present in the sample to indicate which organisms are present that potentially contribute to carbon stocks. Therefore, it cannot give information on the relative abundance of the organisms due to differences in gene copy numbers amongst different organisms.

To gain better coverage and avoid primer bias in future studies, a 'toolbox' approach may be required, where several primer sets for different target genes are employed to specifically amplify different groups. This is especially important when attempting to characterise large taxonomic groups, for example, terrestrial plants. Selected primer sets should be tested *in silico* to ensure good coverage (Clarke et al., 2014). Moreover, experiments using mock communities will help elucidate how different primer sets perform *in vitro* (Braukmann et al., 2019).

Sequencing of the amplicons was successful, yielding 8.2M reads that passed quality filters. There was an uneven balance of reads between samples, due to low concentration amplicons being difficult to quantify accurately and subsequently normalise. Samples with read counts <10,000 were removed from further analysis, which is lower than the recommended <50,000 reads to allow more samples to be analysed further. However, this threshold is reasonable for low biomass samples (e.g. Bender et al. (2018)). Sequencing of the rbcL gene was largely successful with only three samples not meeting the >10,000 reads threshold and filtering resulting in the retention on average of 83% of reads per sample. Sequencing of the Mollusc 18S rRNA gene was less successful, with all samples meeting the >10,000 reads threshold, but with filtering resulting in retention on average of 34% of reads per sample only.

The inclusion of a size filter during the development of the reference database may have omitted partial sequences that include the target genes and therefore, some species from

the database. However, due to the difficulty in identifying organisms down to species level, this may not be an issue for this study.

The positive signals for both gene targets mostly came from surface samples rather than samples at depth. This suggests that DNA preservation in the deeper sediments is poor in this particular system, and therefore, the eDNA approach is most suitable to characterising carbon origins in the biologically active layer rather than in older carbon deposits stored in deeper sediments. This approach has been taken by others, for example, Reef et al. (2017) only sampled the top 1 cm of sediment to characterise recent sources of organic matter (< 10 years) because current sources are more important for carbon accounting (Reef et al., 2017). On the other hand, deeper less mobile sediments contain a carbon sink, which is also important for carbon accounting.

The community compositions found using the rbcL gene showed little inter-site variability, with communities dominated by different genera of diatoms. By contrast, community compositions found using the Mollusc 18S rRNA gene showed high inter-site variability, containing different relative abundances of ASVs assigned to worms (Nematoda and Annelida) and arthropods. It is possible that variability between samples came from the different extraction methods used. However, the inter-site variability may not be surprising given many of the samples failed to produce enough raw reads and some had highly reduced read numbers per sample after denoising. Better coverage will more accurately capture the actual community compositions. All classified rbcL ASVs were associated with marine or freshwater environments, again, unsurprising given the potential primer biases discussed previously.

Sequencing data was assessed for non-native species, which identified a genus of razor clams (*Ensis* sp.). However, care should be taken in interpreting this result since this group contains both native and non-native species. Moreover, assessing sequencing data for non-native species can be challenging. The F measures (Section 3.1) showed that our databases would be less accurate at lower levels of taxonomy, namely in the genus (level 6) and species level (level 7). The classifier will only assign taxonomy to the level at which it has a confidence level of 0.7 or higher (the default and recommended setting in QIIME2). As a result, delineating out species from amplicon data can be difficult as sequences are rarely classified to species level confidently and often non-native species may be in the same genus, family or order as native species.

Stable isotope analysis is a popular technique for tracing carbon origins. It is based on the discrimination against heavier isotopes in metabolic processes, therefore identifying organisms from different trophic levels along axes of 15N/14N and separates organisms with different types of primary productivity and carbon sources along axes of 13C/12C (Post, 2002). However, the taxonomic resolution of this technique is limited between similar trophic levels and carbon-fixing pathways. For example, terrestrial plants and mangrove trees both have lower d13C values, so it would not be possible to distinguish between carbon contributions of terrestrial plants from that of mangrove trees (Rodelli et

al., 1984). The eDNA approach provides greater resolution between taxa, and has been used in combination with stable isotope analysis to get a better understanding of carbon stocks (e.g. Reef et al. (2017)). There is therefore potential to apply the eDNA approach to characterise carbon stocks in other marine habitats, including subtidal sediments.

It should be noted that we have not performed analysis to determine the percentage contribution of species to intertidal sediment stocks or organic carbon contribution from each species, due to the reasons outlined above.

5 Recommendations

Based on the outcomes of this project, we recommend that various steps be optimised to allow accurate characterisation and quantification of carbon stocks, as outlined below.

5.1 Sampling

The sampling plan should be reviewed to allow collection of samples from areas with higher organic matter content, such as stable muddy sediments, to give better DNA yields. The sampling mass collection should also be increased to at least 30 g to allow DNA extraction replication and compatibility with DNA extraction methods, particularly the Dneasy PowerMax kit for low biomass samples (see Section 5.2). Sandy samples tend to have low biomass (Qiagen, 2023), so it is difficult to get a good DNA yield for metabarcoding studies. The issues of low DNA yields were clearly carried through to subsequent stages of analysis. In addition, the results of this study showed that mostly surface sediment (i.e. containing recent carbon sources) yielded DNA to conduct metabarcoding, while only a few deeper sediments (i.e. older carbon sinks) yielded DNA. This may be due to poor DNA preservation in deeper sediments in that particular system. We therefore recommend testing DNA recovery from various sediments at depth. If indeed, DNA preservation is poor and only present at very low concentrations, then forensic DNA kits such as the QIAamp DNA Investigator Kit (Qiagen) may need to be employed.

5.2 DNA extraction

Different DNA extraction methods should be tested to select the best method for the given sample matrix (van der Loos and Nijland, 2021). This will avoid biases associated with extraction. For low biomass samples, this study showed that the Dneasy PowerMax Soil DNA extraction kit (Qiagen) performed the best. However, larger sediment volumes should be processed if this method is to be employed because the kit can process up to 10 g

sediment in a single extraction. Increasing samples to at least 30 g of sediment will allow triplicate extractions to be performed. In this study, we performed triplicate DNA extractions to account for heterogeneity in sediment samples, and we recommend that this practice continue in future studies. Moreover, for saline samples, higher DNA yields may be obtained if the salts are removed by washing with sterile PBS prior to extraction (Qiagen, 2023). This could be performed if issues with low DNA yields persist. As part of our quality control procedure, we measured DNA purity (260/280 ratio), which should be 1.8 for pure DNA. If pure DNA is not obtained, we recommend further DNA cleaning using kits such as the NucleoSpin Gel and PCR Clean-up (Macherey-Nagel), which is also suitable for cleaning genomic DNA. This will help reduce PCR inhibition (see Section 5.4).

5.3 Primer selection

This study focussed on the rbcL gene to characterise macrophytes and 18S rRNA gene to characterise molluscs; however, these markers may not provide the best coverage. A more in-depth literature review with a wider scope (we only searched for primers for marine organisms, but primers for freshwater organisms may also be appropriate) may identify better markers and corresponding primer sets to characterise these organisms. Alternatively, a toolbox approach could be taken where several markers are employed to allow better coverage of the communities and reduce primer bias. For example, the ITS and trnL markers could be used to characterise terrestrial plants (Cheng et al., 2016, Taberlet et al., 2007) and the mitochondrial 12S rRNA gene and the COI gene could be used to characterise bivalves (Barucca et al., 2004, Reunov et al., 2021). In any case, *in silico* analysis similar to that performed in this study should be conducted to verify coverage and identify biases. Moreover, the primers should be validated *in vitro* using mock communities to elucidate how they perform on mixed communities (Braukmann et al., 2019). Similarly, pre-extraction biases exist with larger organisms being overrepresented due to having higher biomass and therefore higher target gene copies (Elbrecht and Leese, 2015); this could also be investigated *in vitro*. In addition, an optimised eDNA approach could be complemented by stable isotope analysis to quantify carbon content, which is becoming a common approach for trophic ecology studies (Compson et al., 2019, Whitaker et al., 2019, Hoenig et al., 2022).

5.4 PCR

We recommend alleviating PCR inhibition by further cleaning of eDNA (Section 5.2). Moreover, PCR enhancers, for example, those tested by Kemp et al. can help reduce the effect of PCR inhibitors, particularly in aged, degraded and low copy number DNA (Kemp et al., 2020). Annealing temperature is an important factor in ensuring specificity of the primers (van der Loos and Nijland, 2021). In this study, we performed temperature gradient PCR to find the optimal annealing temperature for the selected primer sets. We recommend that this be performed prior to the use of any new primer sets. We also

performed triplicate reactions to account for stochasticity during PCR amplification (Bourlat et al., 2016), and we recommend that this practice be continued in future studies. Since DNA yields were low, we used higher PCR cycles (40 cycles) and we recommend that this practice be continued for all low biomass samples. In this study, we used Platinum™ Hot Start PCR Master Mix (Invitrogen), but there are other polymerases on the market that will influence primer bias as certain polymerases have a preference for specific GC contents. In addition, hi-fidelity polymerases can reduce PCR error rates (Liu et al., 2020). Moreover, additives such as dimethyl sulphoxide (DMSO) for GC-rich templates or betaine for AT-rich templates can reduce PCR bias (Nichols et al., 2018). Increased template volumes can increase PCR sensitivity (Hodgson et al., 2015), as can reduced reaction volumes (Leclair et al., 2003). If the choice of polymerase or PCR conditions are a concern, then we recommend trialling different polymerases and PCR reagents and volumes against mock communities.

5.5 Sequencing

There was a clear imbalance in amplicon representation across the samples, possibly due to low concentrations of amplicons. Therefore, we recommend increasing amplicon product concentration prior to pooling. Improvements in DNA yields will help to improve amplicon concentration as well as increased PCR cycle numbers. We also recommend that high sequencing depths be used to provide good coverage of the communities (Smith and Peay, 2014).

5.6 Sequence processing

During sequence processing, we removed samples that had read counts <10,000 from further analysis. However, this was only performed for this study to allow more samples to be taken forward for further analysis and because this is a reasonable threshold for low biomass samples (Bender et al., 2018). In future studies, we recommend sticking to a higher threshold of 50,000 high quality reads.

5.7 Taxonomic assignment

Including mock communities in the sequencing and/or DNA from single species of interest would allow a direct test of the taxonomic assignment. In a large-scale monitoring experiment, such samples might be best restricted to an initial pilot sequencing run to avoid the small but non-zero risk of cross-talk though “tag switching” (imperfect demultiplexing) wrongly contributing reads from these mocks to biological samples.

5.8 Statistical considerations

Due to low DNA yields, many of the samples were dropped because they failed to give amplicons. The small number of remaining samples means that that statistical power was lost. In future studies, if statistical differences between taxa associated with depth or location is required, then multiple replicates (at least 10) per site should be taken, and more sampling sites (at least 10) should be considered.

6 Conclusions

This pilot study demonstrated that eDNA methods could potentially be used to characterise recent carbon stocks in surface intertidal sediments. It is likely that this study only captured a small proportion of the true diversity of carbon sources given the issues encountered with low biomass and potential primer bias. Nevertheless, the dominance of diatoms in the macrophyte dataset was reasonable given their wide distribution in surface sediments. Likewise, the dominance of worms and arthropods in the mollusc dataset is reasonable, especially if bivalves and gastropods are scarce. Optimisation of the process will yield results that better reflect the true community compositions contributing to stored carbon. Moreover, it provides higher taxonomic resolution compared to popular methods such as stable isotope analysis.

Optimisation of the different steps in this process would be required for future studies to fully characterise the origins of UK carbon stocks. These are:

1. Review of sampling protocols to capture samples with high organic matter and provide sufficient material to allow replication, especially for low biomass samples.
2. Exploration of DNA preservation in deeper sediments containing locked-in carbon.
3. Consider higher sample replication for statistical analysis (10 sampling stations and 10 replicate samples per station).
4. Exploration of DNA extraction methods to reduce extraction bias.
5. Perform technical replicates at DNA extraction and PCR amplification stages to account for sediment heterogeneity and PCR stochasticity, respectively.
6. Review target genes or consider a ‘toolbox’ approach (selection of multiple target genes) to provide better coverage.
7. Test new selected primer sets *in silico* and *in vitro* (using mock communities) to determine primer bias. Test different PCR reagents, conditions, and volumes *in vitro* to determine species coverage.
8. Conduct temperature gradient PCR with new selected primer sets to obtain optimal PCR conditions. Use 40 PCR cycles for low biomass samples.
9. Include mock communities and/or DNA from single species in the sequencing to allow a direct test of the taxonomic assignment.

10. Ensure sufficient amplicon concentration for sequencing to ensure even distribution of reads amongst samples.
11. If the data allows, use the recommended threshold of >50,000 quality reads for further analysis during sequence processing. Otherwise, a threshold of >10,000 reads is reasonable for low biomass samples.

7 Appendices

Appendix 1 Detailed Protocols

1.1 CTAB DNA extraction

1.1.1 Reagents

- 10% (weight/volume) CTAB in 0.7 M NaCl, autoclaved
- Potassium phosphate buffer (240 mM pH 8), autoclaved
- Phenol, equilibrated (tris buffered 10 mm, Sigma)
- Chloroform:Isoamyl alcohol (24:1); Sigma)
- 30% Poly (ethylene) Glycol (PEG) 6000 in 1.6 M NaCl, autoclaved
- 70% Ethanol made with ultra-pure nuclease free water
- Ultra-Pure nuclease free H_2O (Sigma)
- TE Buffer pH 7-8 (optional alternative to H_2O), autoclaved
- Linear Acrylamide (5 mg/ml, Life technologies,)
- RNAseIn (Optional)

1.1.2 Consumables

- 2.0 ml Lysing matrix “B” tubes (MP Biomedicals)
- 2.0 ml MaXtract Tubes (Qiagen)

1.1.3 Initial Steps

The following steps are designed to allow the user to prepare and save time ahead of extraction:

- Mix the 10% (wt./vol.) CTAB buffer in 0.7 M NaCl and Potassium phosphate buffer (240 mM pH 8) in equal amounts (1:1). Each sample requires 0.5 ml of the mixture. If a crystalline deposit is observed in the CTAB bottle before use: warm on a hotplate or heat briefly (<20 seconds) in a microwave at full power to re-dissolve the precipitated material.
- Load Lysing matrix “B” tubes with 0.5 ml of CTAB/potassium phosphate mixture, then briefly centrifuge and freeze the tubes at -20ºC. These prepared tubes can be kept in the freezer indefinitely.
- When weighing out materials for extraction, ensure the sample defrosts as little as possible – try to keep the tubes or packets on dry ice or in the vapour phase of LN2 at all times. This is very important for RNA work, but less so for DNA work.

- Inspect the sample(s) to be extracted – if the material is friable (crumbly) and easily mixed, proceed to weigh out 0.25- 0.5 g (± 0.1 g) into a CTAB/phosphate-loaded Lysing matrix B tube, and record the weights.
- Keep the Lysing matrix tubes on dry ice or in the vapour phase of liquid nitrogen whilst weighing out samples.
- If the sample is hard to break up when frozen, use a pre-chilled mortar filled with LN2 and a pestle to break up the material. If this is not possible, a cryomill may be employed to mechanically pulverise the samples.
- The samples in Lysing matrix tubes may be stored at -80ºC (RNA) or -20ºC (DNA) until required.
- Examine the bottle of Phenol to be used for the extraction: if the tris buffer has not been added, follow the instructions on the bottle to add and mix the buffer. Leave the mixture to settle for at least the minimum time period suggested on the bottle (ideally, leave the bottle overnight).
- On the day of extraction: If doing an RNA extraction, add 1.5 µl of Linear Acrylamide (LA) to the tubes and keep on ice – the LA is a co-precipitant and can help remove contaminants.
- On the day of extraction: For final resuspension of DNA/RNA; prepare solution of 30 µl RNAse free H_2O + 1.5 µl RNAseIn per sample. Adding RNAseIn to the tubes is not necessary if the extraction is aimed at DNA only.
- Prepare the 70% ethanol in nuclease free water and store at -20ºC in a spark free freezer until required.

1.1.4 Extraction

1. Work in a fume cupboard for all steps until precipitation is complete. Change gloves frequently.
2. Estimate the amount of phenol and CHCl3:IAA (24:1) required for the day of extraction and then measure the reagents out into labelled glass tubes: 0.85 ml per sample for the CHCl3:IAA (24:1) and 0.35 ml for the phenol (round up to the nearest 0.5 ml to allow for evaporation). Place the tubes on ice until required.
3. Centrifuge a suitable number of Maxtract tubes at maximum speed for one minute, then add 0.1 phenol and CHCl3:IAA (24:1) to each tube. Place the tubes on ice about 20 min before use.
4. Remove the pre-filled Lysing matrix tubes from the freezer and place on ice for 5-10 min. Do not allow complete thawing – proceed to the next step when the CTAB begins to turn back into a liquid.
5. Add 0.25 ml each of ice cold phenol and chloroform:isoamyl alcohol, 24:1 to the Lysing matrix tubes. If the phenol appears cloudy before use, hand-warm the tube until the solution becomes clear.
6. Ensure the lid caps are tightly closed, and then place the tubes in the sample homogeniser – activate the homogeniser for 1 x 15 s at speed 5600 rpm.

7. Remove the tubes and cool them on ice for 5 min.
8. Place tubes in sample homogeniser and repeat lysing for another cycle, then place on ice for a further 5 min.
9. Centrifuge the tubes at 16000 x g for 10 min.
10. Aspirate the entire upper aqueous layer, and transfer to the correspondingly labelled pre-centrifuged and preloaded Maxtract tubes.
11. Clamp tubes tightly between two tube racks and shake gently by inversion for 30 s. Take care that all tubes are tightly shut to prevent leakage.
12. Centrifuge for 10 min at maximum speed.
13. The silica layer in the MaXtract tubes should have begun to separate the aqueous from the organic layer, plus any contamination (proteins, general detritus) from the interface should also have been trapped below the silica layer. Add 0.5 ml of ice-cold Chloroform: Isoamyl alcohol into the same tubes on top of aqueous layer and mix by gentle inversion; clamped between two tube racks for 10-20 s.
14. Centrifuge for 10 min at maximum speed. The aqueous and organic phases should now be completely separated by the silica layer.
15. Extract the entire aqueous phase (top) layer from MaXtract tubes into prepared 1.5 Eppendorf-style tubes (containing Linear Acrylamide if required), thereby estimating the volume of the layer. Precipitate the RNA/DNA by adding approximately two volumes of 30% PEG6000/1.6M NaCl solution, then mix well by inverting tubes repetitively, clamped between two racks for 30 s. For example: if you recover ~500 µl aqueous phase from the tube, then add 2 x 500 µl or 1 ml of PEG6000/1.6M NaCl solution to the tube.
16. Leave tubes on ice for 2 h to allow precipitation of nucleic acids.
17. Centrifuge the tubes at max speed for 30 min. A pellet (invisible at this stage) will form at the bottom/outside of the tube.
18. Carefully aspirate and discard the supernatant with a 1 ml pipette, trying not to touch the tube wall. Experienced users may pour out the supernatant, taking care that the pellet does not slide out.
19. Add 1 ml of 70% ice cold Ethanol; then mix well by clamping the tubes between two racks and repetitively inverting for 20 s.
20. At this point the samples may be stored at -20ºC for up to 72 h, or at -80ºC for longer term storage.
21. When ready to proceed, remove the tubes from the freezer (if stored in the last step) and place them in the centrifuge; spin at maximum speed for 10 min.
22. Carefully aspirate and discard the supernatant with a 1 ml pipette, trying not to touch the tube wall. Experienced users may pour out the supernatant, taking care that the pellet does not slide out.
23. Add 1 ml of 70% ice cold Ethanol; then mix well by clamping the tubes between two racks and repetitively inverting for 30 s. Centrifuge the tubes at maximum speed for 10 min.

24. Remove the supernatant as before, but at this step take care to remove as much volume as possible: briefly centrifuge a second time to collect the remaining liquid and using a smaller volume pipette tip to get everything out (avoiding the pellet)
25. Dry the pellet in a pre-warmed drying block (tube lids open) at 55ºC. If there is no liquid visible, drying will take only a few minutes (1-3 min) – if there is too much liquid left, it may take a while.
26. Tubes are dried if there is no liquid left and the pellets are white. Tubes are over-dried if the pellet turns clear. RNA/DNA becomes difficult to re-dissolve if over-dried; however, residual EtOH and salts from wet pellets will inhibit downstream analysis.
27. If both DNA and RNA are required, add 30 µl of prepared nuclease free water/ RNAseIn solution and incubate for 5-15 min at 55ºC on drying block (lids closed) to re-dissolve. If only DNA is required, just add 30 µl of nuclease free water and incubate at 55ºC in the same way. If pellets were over-dried, i.e. had become clear, re-dissolve for 15 min, if pellets were just on the brink between white and becoming clear re-dissolve for 5 min. Carefully flick and shake down every few minutes. Check that the entire DNA pellet is dissolved.
28. If both DNA and RNA are required, spin down and split extracts into two sets of labelled tubes: one for RNA, and one for DNA, 15 µl each.
29. Flash freeze RNA tubes in LN2 if possible and transfer to -80ºC. For long term storage, RNA solutions can be precipitated again with NaOAc and stored in EtOH.

1.2 PCR

1. Thaw the Invitrogen Platinum Hot Start PCR 2X Master Mix Reagents (master mix, forward primer and reverse primer) on ice, mix and briefly centrifuge beforehand. Keep reagents on ice once thawed and when not in use.
2. Add the following components to each PCR tube for a final reaction volume of 25 µL: 12.5 µL Invitrogen Platinum Hot Start PCR 2X Master Mix, 0.5 µL forward primer (10 µM), 0.5 µL reverse primer (10 µM), 6.5 µL nuclease-free water and 5 µL template DNA.
3. Add the Platinum GC enhancer for targets with >65% GC.
4. Mix and briefly centrifuge PCR mix prior to thermocycling using the Eppendorf Mastercycler PCR System.
5. PCR conditions: initial denaturation at 94ºC for 2 min, followed by 40 cycles of denaturing at 94ºC for 30 s, annealing at 55ºC for 30 s and extension at 72ºC for 30 s.
6. Keep PCRs at 4ºC until further analysis.

1.3 Amplicon clean-up

1. Bring AmpliClean beads up to room temperature and make fresh 80% ethanol.
2. Centrifuge the amplicon PCR plate at 1,000 x g for 1 min to collect reaction to the bottom of the well. Transfer the PCRs from the PCR plate to a MIDI deep-well plate compatible with using a shaker for mixing.
3. Vortex the AmpliClean beads for 30 secs and then add 45 µL of AmpliClean beads to each well of the MIDI plate. This ratio (1.8X) of AmpliClean to PCR product should recover all amplicons >100 bp allowing the removal of smaller contaminants.
4. Seal the plate using a plate seal and place on a shaker (Eppendorf MixMate) for 2 min at 1800 rpm.
5. Incubate the plate at room temperature without shaking for 5 min.
6. Placed the plate on a magnetic stand for 2 min allowing the supernatant to clear so that it could be discarded using a pipette.
7. With the plate still on the magnetic stand, wash beads twice with the fresh 80% ethanol as follows: add 200 µL of 80% ethanol to each sample well, incubate for 30 s at room temperature and then remove supernatant and discard using a pipette.
8. Remove excess 80% ethanol and leave plates to air dry on the magnetic stand for 10 min.
9. Remove the plate from the magnetic stand and add 52.5 µL of Tris (10 mM pH8.5) to each sample well. Seal the plate using a plate seal and place on a shaker (Eppendorf MixMate) for 2 min at 1800 rpm to ensure beads are fully resuspended.
10. Incubated for a further 2 min at room temperature.
11. Placed plate back on the magnetic stand for 2 min to allow the supernatant to clear.
12. Remove 50 µL of supernatant and place into a new PCR plate ready for the index PCR.

1.4 Index PCR

1. Thaw the Invitrogen Platinum Hot Start PCR reagents (Invitrogen Platinum Hot Start PCR 2X Master Mix, Nextera XT Index 1 Primers (N7XX) from the Nextera XT Index kit and reverse primer, and Nextera XT Index 2 Primers (S5XX) from the Nextera XT Index kit) on ice, mix and briefly centrifuged beforehand.
2. Transfer 5 µL of the cleaned amplicon to a new 96-well plate and store the remaining 45 µL at -20ºC.
3. Arrange Index 1 and 2 primers in a TruSeq Index Plate Fixture as per the Illumina 16S metagenomic library prep guide: place Index 1 primer tubes with orange caps horizontally, aligned with the appropriate number of columns, and place Index 2 primer tubes with white caps vertically, aligned with rows A through to H. Also place the plate containing 5 µL of amplicon on the Index Plate Fixture.
4. To each well, add the following reagents: 5 µL Nextera XT Index 1 Primers (N7XX), 5 µL Nextera XT Index 2 Primers (S5XX), 25 µL Invitrogen Platinum Hot Start PCR

2X Master Mix and 10 µL nuclease-free water for a total reaction volume of 50 µL including the DNA.

5. Mix the reaction by pipetting up and down 10 times and then cover the with a plate seal before centrifuging at 1,000 x g for 1 min.
6. Index PCR conditions: initial denaturation at 95ºC for 3 min, followed by 8 cycles of denaturation at 95ºC for 30s, annealing at 55ºC for 30 s and extension at 72ºC for 30 s, and a final extension step at 72ºC for 5 min. Hold the plate at 4ºC until further clean-up.

1.5 Index PCR clean-up

1. Protocol is the same as the amplicon PCR clean-up. Differences are:
    a. the addition of 90 µL of the AmpliClean (instead of 45 µL) due to the increased volume of product;
    b. the addition of 27.5 µL of Tris (10 mM pH8.5) to give a final volume of 25µL clean product.

8 References

Anglès d'Auriac, M., Hancke, K., Gundersen, H., Frigstad, H. and Borgersen, G. 2021. Blue Carbon eDNA–A novel eDNA method to trace macroalgae carbon in marine sediments. *NIVA-rapport*.

Barucca, M., Olmo, E., Schiaparelli, S. and Canapa, A. 2004. Molecular phylogeny of the family Pectinidae (Mollusca: Bivalvia) based on mitochondrial 16S and 12S rRNA genes. *Mol Phylogenet Evol,* 31**,** 89-95.

Bender, J. M., Li, F., Adisetiyo, H., Lee, D., Zabih, S., Hung, L., Wilkinson, T. A., Pannaraj, P. S., She, R. C., Bard, J. D., Tobin, N. H. and Aldrovandi, G. M. 2018. Quantification of variation and the impact of biomass in targeted 16S rRNA gene sequencing studies. *Microbiome,* 6**,** 155.

Biesbroek, G., Sanders, E. A., Roeselers, G., Wang, X., Caspers, M. P., Trzcinski, K., Bogaert, D. and Keijser, B. J. 2012. Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection. *PLoS One,* 7**,** e32942.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A. and Gregory Caporaso, J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome,* 6**,** 90.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodriguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C.,

Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y. X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., 2nd, Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vazquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R. and Caporaso, J. G. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol,* 37**,** 852-857.

Bourlat, S., Haenel, Q., Finnman, J. and Leray, M. 2016. Preparation of amplicon libraries for metabarcoding of marine eukaryotes using Illumina MiSeq: The Dual-PCR Method. *In:* S., B. (ed.) *Marine Genomics: Methods and protocols.* New York, NY: Springer. .

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., de Waard, J. R., Sones, J. E., Zakharov, E. V. and Hebert, P. D. N. 2019. Metabarcoding a diverse arthropod mock community. *Mol Ecol Resour,* 19**,** 711-727.

Bringloe, T. T., Sjøtun, K. and Saunders, G. W. 2019. A DNA barcode survey of marine macroalgae from Bergen (Norway). *Marine Biol Res,* 15**,** 580-589.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. and Holmes, S. P. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods,* 13**,** 581-3.

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. *Mol Ecol Resour,* 16**,** 138-49.

Clarke, L. J., Soubrier, J., Weyrich, L. S. and Cooper, A. 2014. Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Mol Ecol Resour,* 14**,** 1160-70.

Cock, P. J. A., Cooke, D. E. L., Thorpe, P. and Pritchard, L. 2023. THAPBI PICT-a fast, cautious, and accurate metabarcoding analysis pipeline. *PeerJ,* 11**,** e15648.

Compson, Z. G., Monk, W. A., Hayden, B., Bush, A., O'Malley, Z., Hajibabaei, M., Porter, T. M., Wright, M. T. G., Baker, C. J. O., Al Manir, M. S., Curry, R. A. and Baird, D. J. 2019. Network-Based Biomonitoring: Exploring Freshwater Food Webs With Stable Isotope Analysis and DNA Metabarcoding. *Frontiers in Ecology and Evolution,* 7.

Edgar, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv preprint*.

Elbrecht, V. and Leese, F. 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass--Sequence Relationships with an Innovative Metabarcoding Protocol. *PLoS One,* 10**,** e0130324.

Espiñeira, M., González-Lavín, N., Vieites, J. M. and Santaclara, F. J. 2009. Development of a method for the genetic identification of commercial bivalve species based on mitochondrial 18S rRNA sequences. *J Agric Food Chem,* 28**,** 495-502.

Fukai, Y., Matsuno, K., Fujiwara, A. and Suzuki, K. 2022. Photophysiological response of diatoms in surface sediments to light exposure: A laboratory experiment on a diatom community in sediments from the Chukchi Sea. *Frontiers in Marine Science,* 9.

Griffiths, R. I., Whiteley, A. S., O'Donnell, A. G. and Bailey, M. J. 2000. Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Appl Environ Microbiol,* 66**,** 5488-91.

Hodgson, S. H., Douglas, A. D., Edwards, N. J., Kimani, D., Elias, S. C., Chang, M., Daza, G., Seilie, A. M., Magiri, C., Muia, A., Juma, E. A., Cole, A. O., Rampling, T. W., Anagnostou, N. A., Gilbert, S. C., Hoffman, S. L., Draper, S. J., Bejon, P., Ogutu, B., Marsh, K., Hill, A. V. and Murphy, S. C. 2015. Increased sample volume and use of quantitative reverse-transcription PCR can improve prediction of liver-to-blood inoculum size in controlled human malaria infection studies. *Malar J,* 14**,** 33.

Hoenig, B. D., Trevelline, B. K., Kautz, A., Latta, S. C. and Porter, B. A. 2022. Two is better than one: Coupling DNA metabarcoding and stable isotope analysis improves dietary characterizations for a riparian-obligate, migratory songbird. *Molec Ecol,* 31**,** 5635-5648.

IPCC 2014. 2013 Supplement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories: Wetlands. *In:* HIRAISHI, T., KRUG, T., TANABE, K., SRIVASTAVA, N., BAASANSUREN, J., FUKUDA, M. & TROXLER, T. G. (eds.). IPCC, Switzerland.

Karstens, L., Asquith, M., Davin, S., Fair, D., Gregory, W. T., Wolfe, A. J., Braun, J. and McWeeney, S. 2019. Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. *mSystems,* 4.

Keefe, C. R., Dillon, M. R., Gehret, E., Herman, C., Jewell, M., Wood, C. V., Bolyen, E. and Caporaso, J. G. 2023. Facilitating bioinformatics reproducibility with QIIME 2 Provenance Replay. *PLoS Comput Biol,* 19**,** e1011676.

Kemp, B. M., Bingham, B., Frome, R., Labonte, M., Palmer, E., Parsons, E. S., Gobalet, K. W. and Rosenthal, J. 2020. Subduing the influence of PCR inhibitors on amplifying aged, degraded, and low copy number DNA: PCR enhancer cocktail-p and rescue PCR. *PLoS One,* 15**,** e0234745.

Lahiri, D. K. and Schnabel, B. 1993. DNA isolation by a rapid method from human blood samples: effects of MgCl2, EDTA, storage time, and temperature on DNA yield and quality. *Biochem Genet,* 31**,** 321-8.

Leclair, B., Sgueglia, J. B., Wojtowicz, P. C., Juston, A. C., Fregeau, C. J. and Fourney, R. M. 2003. STR DNA typing: increased sensitivity and efficient sample consumption using reduced PCR reaction volumes. *J Forensic Sci,* 48**,** 1001-13.

Liu, M. X., Clarke, L. J., Baker, S. C., Jordan, G. J. and Burridge, C. P. 2020. A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology,* 45**,** 373-385.

Magoc, T. and Salzberg, S. L. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics,* 27**,** 2957-63.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* 17**,** 10-12.

Miyajima, T., Hamaguchi, M., Masakazu, H. O. R. I., Munar, J., Abad, J. and Morimoto, N. 2022. Evaluation of carbon export from blue carbon ecosystems and allochthonous sequestration using eDNA techniques. *Authorea Preprints*.

Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., Green, R. E. and Shapiro, B. 2018. Minimizing polymerase biases in metabarcoding. *Mol Ecol Resour*.

Ortega, A., Geraldi, N. R., Díaz-Rúa, R., Ørberg, S. B., Wesselmann, M., Krause-Jensen, D. and Duarte, C. M. 2019. A DNA mini-barcode for marine macrophytes. *Mol Ecol Resour,* 20**,** 920-935.

Ortega, A., Geraldi, N. R. and Duarte, C. M. 2020. Environmental DNA identifies marine macrophyte contributions to Blue Carbon sediments. *Limnology and Oceanography,* 65**,** 3139-3149.

Parliament, U. 2021. Blue carbon. *POSTNOTE 651*.

Post, D. M. 2002. Using stable isotopes to estimate trophic position: models, methods, and assumptions. *Ecology,* 83**,** 703-718.

Qiagen 2023. Microbial DNA isolation from soil samples.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glockner, F. O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res,* 41**,** D590-6.

Reef, R., Atwood, T. B., Samper-Villarreal, J., Adame, M. F., Sampayo, E. M. and Lovelock, C. E. 2017. Using eDNA to determine the source of organic carbon in seagrass meadows. *Limnology and Oceanography,* 62**,** 1254-1265.

Reunov, A., Lutaenko, K., Vekhova, E., Zhang, J., Zakharov, E., Sharina, S., Alexandrova, Y., Reunova, Y., Akhmadieva, A. and Adrianov, A. 2021. In the Asia-Pacific region,

the COI DNA test revealed the divergence of the bivalve mollusc Mactra chinensis into three species; can these species be distinguished using shell coloration and sperm structure? *Helgol Mar Res,* 75**,** 7.

Robeson, M. S., 2nd, Khanipov, K., Golovko, G., Wisely, S. M., White, M. D., Bodenchuck, M., Smyser, T. J., Fofanov, Y., Fierer, N. and Piaggio, A. J. 2018. Assessing the utility of metabarcoding for diet analyses of the omnivorous wild pig (Sus scrofa). *Ecol Evol,* 8**,** 185-196.

Robeson, M. S., 2nd, O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T. and Bokulich, N. A. 2021. RESCRIPt: Reproducible sequence taxonomy reference database management. *PLoS Comput Biol,* 17**,** e1009581.

Rodelli, M. R., Gearing, J. N., Gearing, P. J., Marshall, N. and Sasekumar, A. 1984. Stable isotope ratio as a tracer of mangrove carbon in Malaysian ecosystems. *Oecologia,* 61**,** 326–333.

Smith, D. P. and Peay, K. G. 2014. Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS One,* 9**,** e90234.

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C. and Willerslev, E. 2007. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Res,* 35**,** e14.

UKBCEP 2023. UK Blue Carbon Evidence Partnership Evidence Needs Statement. Lowestoft, UK.

UKNNSS 2020. UK Marine Non-Indigenous Species Priority List (updated 2020). *UK Non-Native Secretariat.*

van der Loos, L. M. and Nijland, R. 2021. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Mol Ecol,* 30**,** 3270-3288.

Vasselon, V., Rimet, F., Tapolczai, K. and Bouchez, A. 2017. Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). *Ecological Indicators,* 82**,** 1-12.

Werner, J. J., Koren, O., Hugenholtz, P., DeSantis, T. Z., Walters, W. A., Caporaso, J. G., Angenent, L. T., Knight, R. and Ley, R. E. 2012. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J,* 6**,** 94-103.

Whitaker, M. R. L., Baker, C. C. M., Salzman, S. M., Martins, D. J. and Pierce, N. E. 2019. Combining stable isotope analysis with DNA metabarcoding improves inferences of trophic ecology. *PLoS One,* 14**,** e0219070.

Witzke, M. C., Gullic, A., Yang, P., Bivens, N. J., Adkins, P. R. F. and Ericsson, A. C. 2020. Influence of PCR cycle number on 16S rRNA gene amplicon sequencing of low biomass samples. *J Microbiol Methods,* 176**,** 106033.

Zimmermann, J., Mora, D., Tapolczai, K., Proft, S., Chonova, T., Rimet, F., Bouchez, A., Fidlerova, D., Makovinska, J. and Weigand, A. M. 2021. Metabarcoding of phytobenthos samples. *In:* LIŠKA, I., WAGNER, F., SENGL, M., DEUTCH, K., SLOBODNÍK, J. & PAUNOVIĆ, M. (eds.) *Joint Danube Survey 4. Scientific report: A shared analysis of the Danube river.* International Commission for the Protection of the Danube River.

9 Glossary

12S rRNA – the ribosomal component of the small subunit of the eukaryotic ribosome located on the mitochondrial genome. The gene for this is used for classification of Eukarya.

18S rRNA – the ribosomal component of the small subunit of the eukaryotic ribosome. The gene for this is used for classification of Eukarya.

ASV - amplicon sequence variant

COI – cytochrome c oxidase I. The gene for this is located on the mitochondrial genome and is used for classification of Eukarya.

CTAB – hexadecyltrimethylammonium bromide.

DADA – Divisive Amplicon Denoising Algorithm.

FASTQ –text-based format for storing both a nucleotide sequence and its corresponding quality scores.

Gel electrophoresis – procedure to size-separate DNA molecules

GitHub – code hosting platform for collaboration and version control.

MiSeq – next generation sequencing platform

NCBI – National Centre for Biotechnology Information.

PCR – polymerase chain reaction

PhiX – an icosahedral, non-tailed bacteriophage with a single-stranded DNA, used as a control for Illumina sequencing

QIIME - Quantitative Insights Into Microbial Ecology.

QPCR – quantitative polymerase chain reaction

rbcL – the large-chain gene, is encoded by the chloroplast DNA. This gene is used for classification of macrophytes.

THAPBI – Tree Health and Plant Biosecurity Initiative.

UNOISE – algorithm for denoising (error-correction) sequencing data.

www.gov.uk/natural-england