THE UNIVERSITY OF HULL

DEVELOPMENT OF AN ENVIRONMENTAL DNA METHOD FOR MONITORING FRESHWATER FISH COMMUNITIES USING METABARCODING

being a thesis submitted for the Degree of
Doctor of Philosophy
in the University of Hull

by

Jianlong Li (B.Sc., M.Sc.)

May 2019

Supervisors: Dr Bernd Hänfling; Dr Lori Lawson Handley

Summary

Monitoring current global biodiversity decline is essential to maintain ecosystem functioning, especially freshwater ecosystems. However, the conventional physical, acoustic and visual-based methods for monitoring biodiversity have some limitations such as morphological identification bias, recording small-bodied, rare and/or elusive species, and destructive impacts on the environment. The significant “game-changer” in biodiversity monitoring is environmental DNA (eDNA) metabarcoding, which refers to the simultaneous identification of a multitude of species from environmental samples. However, developing, validating and improving eDNA-based metabarcoding monitoring methods is not trivial. Firstly, for further eDNA metabarcoding studies focusing on freshwater fish communities, two marker-specific reference databases were compiled and two metabarcoding primer pairs were rigorously tested. Subsequently, a PCR-based metabarcoding approach is applied to investigate (1) the effect of filtration method on the efficiency of eDNA capture and quantification, (2) the spatial and temporal distribution of eDNA in fish ponds, and (3) the potential of eDNA as a tool for biodiversity monitoring in diverse lakes with characterised fish faunas. The results show that the 0.8 μm filters are advocated for turbid and eutrophic water such as ponds to reduce the filtration time, the 0.45 μm filters are appropriate for clear water sampling to obtain consistent results, and the 0.45 μm Sterivex enclosed filters are suitable in situations where on-site filtration is required. Furthermore, eDNA distribution in ponds was highly localised in space and time, and 10 shore samples distributed along the full perimeter of lakes is adequate for capturing the majority of species. Lastly, this thesis provides further evidence that eDNA metabarcoding could be a powerful monitoring tool for freshwater fish communities, considerably outperforming other established survey techniques whether in species detection, relative abundance estimate or characterisation ecological fish communities. These outcomes constitute a significant advance towards a standardised and efficient assessment procedure for the ecological monitoring of aquatic ecosystems.

Keywords: eDNA, filtration, eDNA dynamics, community ecology, fish monitoring, lentic systems

Acknowledgments

Time is ticking away; it has been four years since 2014. I am proud of myself I did manage to finish my PhD project supported by the University of Hull and the China Scholarship Council. Of course, people never walk alone, and I had great support in my scientific and private life which made this thesis possible. Here I would like to this opportunity to express gratitude to everyone who helped me out in the last four years.

Firstly, I would like to express my sincere gratitude to my supervisors, Dr Bernd Hänfling and Dr Lori Lawson Handley, for the continuous support of my PhD study and related research, for their patience, encouragement and immense knowledge, and trust and freedom I could wish for.

Besides my supervisors, I would like to thank my thesis examiners, Prof Douglas W. Yu and Dr Domino A. Joyce, for their insightful comments which incented me to widen my research from various perspectives. My special thanks are also extended to Prof Roland Ennos, Dr David H. Lunt, Dr Africa Gomez and Prof Ian J. Winfield for their professional advice during the progress of the project.

I would like to express my sincere gratitude to my co-authors of published or forthcoming manuscripts, Dr Daniel S. Read, Dr Tristan W. Hatton-Ellis, Dr Helen S. Kimbell, Dr Lynsey R. Harper, Dr Marco Benucci, Dr Graeme Peirson, Dr Hayley V. Watson and Dr Rein Brys, for their valuable and constructive advice.

My grateful thanks are extended to the post doctors, technicians and lab mates of EvoHull who I have been working with, Dr Christoph Hahn, Dr Amir Szitenberg, Dr Peter Shum, Dr James J. N. Kitson, Dr Rob Donnelly, Dr Paul Nichols, Dr Cristina Di Muri, Dr Harriet Johnson, Dr Rose Wilcox, Dr Graham S. Sellers, Dr Daniel L. Jeffries, Dr Rosetta C. Blackman and Mr Robert Jaques, for their assistance whether in bioinformatics analysis, lab work or fieldwork.

I am particularly grateful to staff of the National Coarse Fish Rearing Unit of the UK Environment Agency, Mr Alan Henshaw, Mr Richard Pitman, Mr Nick Gill, Mr Dan Clark and Mr James Rabjohns, for their help with sampling and providing fish stock information; to staff of the Natural Resources Wales, Mr Peter Clabburn, Mr Rob Evans, Ms Sophie Gott, Ms Emma Keenan, Mr Ian Sims, Ms Emma Brown, Mr Jason Jones and Ms Dawn Parry, for their assistance with access of sampling sites, providing boat support and additional data interpretation and fisheries advice of Welsh lakes; and to Ecological Consultancy Limited for providing fishery survey data and cooperation in data collation of Cheshire meres.

I wish to thank Prof Min Liu, Prof Shaoxiong Ding, Dr Xiang Zhang, Dr Guihua Wang, Dr Wenjia Du, Dr Jie Ma, Dr Jinru He, Dr Ze Qian, Dr Hao Hu, Dr Shanshan Guan, Dr Lu Wang, Dr Kai Wang, Ms Xinye Wang, Mr Silong Wang and Mr Rui Luo for supporting me spiritually throughout my life in general. There are still lots of people who helped me in the past four years, but I am truly sorry for not being able to list their names in detail.

Finally, I wish to thank my parents for their support and encouragement throughout my study.

“***Life sucks, but you are gonna love it***”

“人生不易，且行且珍惜”

Table of Contents

List of Tables

Tables in Supporting Information

List of Figures

Figures in Supporting Information

List of Abbreviations

12S	mitochondrial 12S ribosomal RNA gene
16S	mitochondrial 16S ribosomal RNA gene
ANOSIM	Analysis of similarities
ANOVA	Analysis of variance
BLAST	Basic Local Assignment Search Tool
BOD	Biochemical Oxygen Demand
BOLD	Barcode of life data system
bp	base pair
BQEs	Biological Quality Elements
CBoL	Consortium for the Barcode of Life
CEH	Centre for Ecology and Hydrology [UK]
CN filter	Cellulose Nitrate
COD	Chemical Oxygen Demand
COI	mitochondrial cytochrome coxidase subunit I gene
Cytb	mitochondrial cytochrome *b* gene
DAFOR scale	D=Dominant; A=Abundant; F=Frequent; O=Occasional; R=Rare
DDBJ	DNA Data Bank of Japan
ddPCR	droplet digital Polymerase Chain Reaction
DNA	deoxyribonucleic acid
DO	Dissolved Oxygen
EA	Environment Agency [UK]
EBV	Essential Biodiversity Variables
ECON	Ecological Consultancy Ltd [UK]
eDNA	environmental DNA
FIL2	Fish In Lakes classification tool
GEO BON	Group on Earth Observations Biodiversity Observation Network
GF filter	Glass Fibre
GPS	Global Positioning System
HDPE	High-density polyethylene
HEPA	High-Efficiency Particulate Air
HSP	High-scoring Segment Pairs
HTS	High-Throughput Sequencing
INNS	Invasive Non-Native Species
INSDC	International Nucleotide Sequence Database Collaboration
ITS	Internal Transcribed Spacer
LCA	Lowest Common Ancestor
matK	maturase K gene
MCE filter	Mixed Cellulose Ester
metaBEAT	metaBarcoding and Environmental DNA Analysis Tool
mtDNA	mitochondria DNA
MOTUs	Molecular Operational Taxonomic Units
NCBI	National Center for Biotechnology Information [USA]
NCFRU	National Coarse Fish Rearing Unit [UK Environment Agency]
NMDS	Non-metric multidimensional scaling

NRW	Natural Resources Wales [UK]
NTC	No Template Control
PASE	Point Abundance Sampling by Electro-Fishing
PC filter	Polycarbonate
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational multivariate analysis of variance
PES filter	Polyethersulfone
pHMM	profile Hidden Markov Model
PVDF filter	Polyvinylidene difluoride
qPCR	real-time quantitative Polymerase Chain Reaction
rbcL	rbcL plastid ribulose 1,5-bisphosphate carboxylase gene
RSPB	Royal Society for the Protection of Birds
SAC	Special Areas of Conservation
SML	Supervised Machine Learning
SPM	Suspended Particulate Matter
SSSI	Sites of Special Scientific Interest
STC	Single Template Control
UK	United Kingdom
UoH	University of Hull [UK]
UV	ultraviolet
UV-B	ultraviolet B
WFD	Water Framework Directive [2000/60/EC]

Chapter 1 General introduction

1.1 Freshwater biodiversity and monitoring

Freshwater habitats contain 0.01% of the world's water and occupy approximately 0.8% of the Earth's surface (Gleick 2011), but this tiny fraction of global water supports 6–10% of all described species (Hassan et al. 2005) including 10,000 fish species, and thus accounts for at least 40% of global fish diversity and one quarter of global vertebrate diversity (Lundberg et al. 2000). However, freshwater ecosystems are among the most endangered habitats on the Earth due to continuous decline in biodiversity resulting from anthropogenic disturbances such as overexploitation of wild species, water pollution, flow modification, destruction or degradation of habitat, the introduction of invasive non-native species (INNS), and climate change (Dudgeon et al. 2006; Strayer & Dudgeon 2010; Vörösmarty et al. 2010; Collen et al. 2014). As a result of this global crisis, monitoring biodiversity change, diagnosing the causes and finding solutions have become a major part of the contemporary freshwater ecology.

To assess biodiversity change at the regional or global level, researchers and conservation managers face a number of obstacles including insufficient and uneven geographic coverage, and lack of spatial and temporal change monitoring (Proença et al. 2017). In an effort to integrate existing and emerging biodiversity monitoring initiatives, the Group on Earth Observations Biodiversity Observation Network (GEO BON) (Scholes et al. 2012) has developed the Essential Biodiversity Variables (EBVs) framework that could form the basis of efficient and coordinated global biodiversity observation systems (Pereira et al. 2013). Under this framework, six broad classes of EBVs have been proposed for global freshwater biodiversity monitoring including genetic composition, species populations, species traits, community composition, ecosystem structure and ecosystem function, each representing a major component of biodiversity (Pereira et al. 2013; Turak et al. 2017). Analysis of genetic composition is rarely included as an objective in freshwater biodiversity monitoring programmes. However, recent advances in DNA High-Throughput Sequencing (HTS) technologies are providing us with rapid and affordable access to obtain vast amounts of genetic information (Shendure & Ji 2008; Taberlet et al. 2012b; Cristescu 2014). In addition to this increase in throughput, another revolution is ongoing in the field of environmental DNA (eDNA) which is a significant "game-changer" in biological monitoring within the last decade (Lawson Handley 2015; Thomsen & Willerslev 2015; Taberlet et al. 2018).

1.2 What is environmental DNA

eDNA defined in this thesis is restricted to the DNA that is obtained directly from environments (e.g., ice, sediments or water) (Figure 1.1) without first isolating any target organisms (Taberlet et al. 2012a; Thomsen & Willerslev 2015). This is distinct from DNA extracted from bulk mixtures of organisms, which can sometimes be called "community DNA" (Creer et al. 2016; Deiner et al. 2017a). Total eDNA contains intracellular DNA originating from living cells or tissue such as faeces, urine, saliva, gametes, shed skin, feathers and carcasses (Pedersen et al. 2015), and extracellular DNA (dissolved DNA) resulting from natural cell death and subsequent destruction of cell structure (Levy-Booth et al. 2007; Pietramellara et al. 2009). The extracellular DNA can be degraded through physical, chemical or biological processes, which then may be absorbed by inorganic or organic surface-reactive particles such as clay, sand, silt, and humic substances (Levy-Booth et al. 2007; Pietramellara et al. 2009).

The concept of eDNA has been used for several decades in microbiology and concerns a method of obtaining microbial DNA directly from sediments (Ogram et al. 1987), which has given microbiologists access to the genetics of uncultivable micro-organisms. For macro-organismal communities, eDNA as a method was first applied to assess the ancient sediment samples from Siberia and New Zealand revealing the past of extinct and extant mammals, birds, and plants (Willerslev et al. 2003). Since then, the approach has been successfully used on both ancient and contemporary environmental samples including sediments (e.g., Hofreiter et al. 2003; Parducci et al. 2012; Willerslev et al. 2014), ice cores (e.g., Willerslev et al. 2007), soil (e.g., O'Brien et al. 2005; Taberlet et al. 2012b), freshwater (e.g., Ficetola et al. 2008; Takahara et al. 2012; Thomsen et al. 2012b), and seawater (e.g., Foote et al. 2012; Thomsen et al. 2012a; Kelly et al. 2014).

1.3 Strategies of eDNA analysis

Broadly, two main PCR-based strategies of biodiversity monitoring can be deployed for eDNA analysis (reviewed in Rees et al. 2014; Barnes & Turner 2016; Deiner et al. 2017a). The first one consists of targeting single species using standard PCR, real-time quantitative PCR (qPCR), or droplet digital PCR (ddPCR) (hereafter referred to as "eDNA single-species detection"). The second strategy aims to simultaneously detect multiple species relying on HTS technologies (hereafter referred to as "eDNA metabarcoding") (a basic overview of the strategies is provided in Figure 1.1).

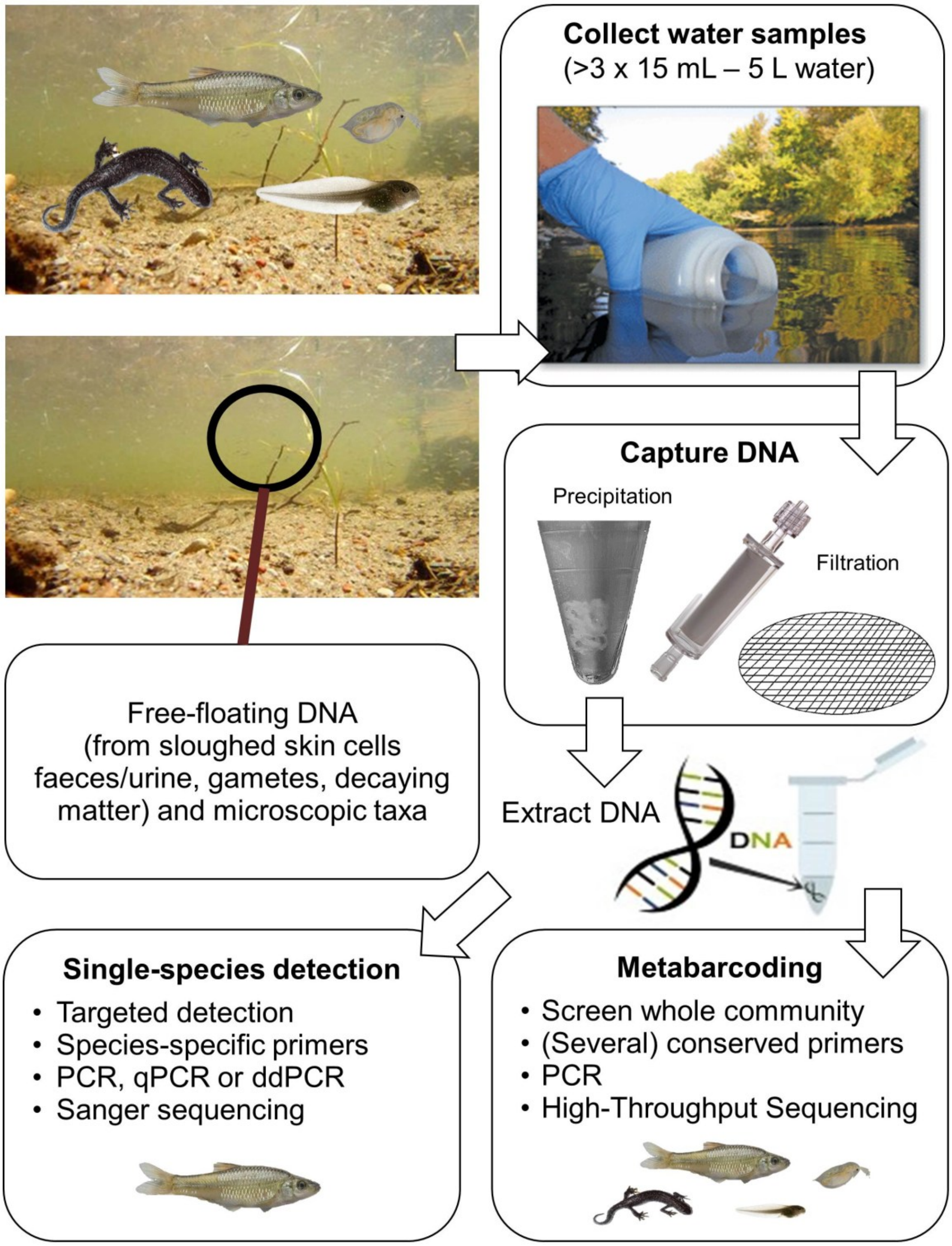


Figure 1.1 Overview of the steps taken in environmental DNA (eDNA) studies from water samples adapted from Lawson Handley (2015).

1.3.1 eDNA single-species detection in freshwater

The target of eDNA single-species detection is one or a few known species, for which species-specific primers and probes can be developed (Figure 1.1). Ficetola et al. (2008) applied this approach to detect the presence at low densities of the American bullfrog *Rana catesbeiana*, invasive in Western Europe, through PCR with a short fragment of the mitochondrial cytochrome *b* gene (Cytb) then sequencing with the 454 pyrosequencing technology. After this study, eDNA single-species identification has been tested and used successfully for discovery, surveillance and monitoring of invasive, rare, or threatened species in freshwater environments such as amphibians (e.g., Dejean et al. 2011; Goldberg et al. 2011; Dejean et al. 2012; Biggs et al. 2015), fish (e.g., Jerde et al. 2011; Jerde et al. 2013; Mahon et al. 2013; Takahara et al. 2013; Wilcox et al. 2013; Keskin 2014; Fernandez et al. 2018), reptiles (e.g., Piaggio et al. 2014), gastropods (e.g., Goldberg et al. 2013), and crustaceans (e.g., Mächler et al. 2014; Tréguier et al. 2014).

Encouragingly, species abundance estimate based on eDNA single-species detection is an intriguing possibility in freshwater ecosystems. Several studies have found positive relationships between eDNA concentration from qPCR or ddPCR and organism abundance and/or biomass density in aquariums, mesocosms and experimental ponds (e.g., Takahara et al. 2012; Thomsen et al. 2012b; Doi et al. 2015; Lacoursière-Roussel et al. 2016b), and in natural freshwater systems such as streams, rivers (e.g., Pilliod et al. 2013; Baldigo et al. 2017; Doi et al. 2017a), and lakes (e.g., Lacoursière-Roussel et al. 2016a).

Although eDNA single-species detection has high specificity, sensitivity, and quantification ability, it is hampered by the limit to detect only one or a few target organisms at a time. For more diverse systems, this approach quickly becomes inefficient, expensive and even impossible due to lack of DNA extract for multiple reactions (Taberlet et al. 2012b; Comtet et al. 2015).

1.3.2 eDNA metabarcoding in freshwater

Rather than focussing on single species, the target of eDNA metabarcoding is whole communities with one or several generic primers to monitor biodiversity for nature conservation of different types of ecosystems (Epp et al. 2012; Taberlet et al. 2012b) (Figure 1.1). The emergence of eDNA metabarcoding was facilitated by breakthroughs in DNA sequencing technology. In the last decade, HTS technologies led to a breakthrough in DNA-based taxon identification (Ellegren 2008; Shendure & Ji 2008; Taylor & Harris 2012; Cristescu 2014). HTS technologies can produce billions of sequence reads in a single run, which makes DNA-based taxon identification possible while bypassing the expensive and time-consuming steps of cloning and sequencing PCR products using Sanger sequencing (Loman et al. 2012; Shokralla et al. 2012).

In freshwater ecosystems, eDNA metabarcoding can complement and overcome the limitations of conventional physical, acoustic and visual-based methods (hereafter referred to as “conventional survey methods”) by targeting different species, sampling greater diversity, and increasing the resolution of taxonomic identifications such as macroinvertebrates, fish, amphibians, and chironomids (e.g., Deiner et al. 2016; Hänfling et al. 2016; Lim et al. 2016; Olds et al. 2016; Shaw et al. 2016a; Valentini et al. 2016; Evans et al. 2017). Estimating abundance information using eDNA metabarcoding for whole communities still lacks substantial evidence, but some studies in freshwater environments have shown positive relationships between the relative sequencing read counts and relative abundance and/or biomass density or rank abundance estimated with conventional survey methods (e.g., Evans et al. 2016;

Hänfling et al. 2016; Lawson Handley et al. 2019), which demonstrated that at least semi-quantitative estimates could potentially be obtained from eDNA data. These results are promising, but not all studies support such findings (e.g., Lim et al. 2016).

1.4 Challenges with eDNA analysis

Despite the obvious perspectives of using eDNA analysis for freshwater biodiversity monitoring, it is affected by a number of precision and accuracy challenges distributed throughout the workflow including DNA capture, preservation and isolation, the choice of metabarcodes and PCR primers, bioinformatics analysis pipelines, and reference databases for taxonomic assignment (reviewed in Thomsen & Willerslev 2015; Deiner et al. 2017a; Pawlowski et al. 2018; Harper et al. 2019). Here, I focus on the following three aspects: marker selection and primer design, eDNA capture, and sampling design.

1.4.1 Marker selection and primer design

In any eDNA analysis study, the choice of the marker is crucial and can greatly impact the biological results and conclusions. The standard barcoding markers recommended by the Consortium for the Barcode of Life (CBoL) are the cytochrome c oxidase subunit I gene (COI) for animals (Hebert et al. 2003b), the plastid ribulose 1,5-bisphosphate carboxylase gene (rbcL) and the maturase K gene (matK) for plants (Hollingsworth et al. 2009), and the internal transcribed spacer (ITS) for fungi (Schoch et al. 2012). These target genes are preferred over single-copy nuclear DNA in eDNA analysis as the high copy number per cell of these genes increases the chance of detection in environmental samples (Thomsen & Willerslev 2015). COI is also recommended by Andújar et al. (2018) for metazoan community metabarcoding study with bulk samples considering the substantial COI specific reference databases, the broader taxonomic coverage and resolution, combined with recent improvements of primer design in COI region. In the case of environmental samples, most primers targeting the COI region amplify large proportions of microbial species. For instance, Stat et al. (2017) demonstrated that less than 0.03% of 3.1 million sequencing read counts from seawater are assigned to COI of fish compared to 94.5% assigning to bacteria. This fact remains the strongest reason for the use of mitochondrial 12S and 16S rRNA genes or other mitochondrial protein-coding genes such as Cytb (see more detail in Chapter 2, Section 2.1) that are much less affected by this type of cross-amplification (Andújar et al. 2018).

The amplicon size of markers is also an important consideration because there may be a trade-off in species detection with amplicon length. DNA from environmental samples, especially ancient samples, is often fragmented and in low concentrations (e.g., Willerslev & Cooper 2005; Deagle et al. 2006). Thus, researchers assumed that eDNA was highly degraded, and eDNA analysis should rely on shorter DNA fragments rather than the traditionally defined barcoding regions (e.g., Hajibabaei et al. 2006; Hajibabaei et al. 2011). More recently, Deiner et al. (2017b) demonstrated the eDNA in freshwater is available in the genomic state indicating that eDNA in water exists in both undegraded and degraded forms. However, short fragments may persist longer in the water and increase the inference in space or time that can be made from environmental samples (e.g., Bista et al. 2017; Jo et al. 2017).

Overall, for eDNA single-species detection, the primers are designed to ensure high target specificity with no base pair mismatches for the target species and as many mismatches as possible for any closely-related or co-occurring species (e.g., Wilcox et al. 2013), while the eDNA metabarcoding primers have to be versatile enough to amplify equally and exhaustively different targeted groups. Additionally, the metabarcode has to have good taxonomic resolution and be discriminative, ideally to the species level (Clarke et al. 2014; Deagle et al. 2014). Therefore, when designing new primers or choosing previously designed primers, it is important to perform rigorous testing, *in silico*, *in vitro*, and *in situ* to infer their utility for eDNA analysis in a new study system based on the taxonomic group(s) of interest (Freeland 2016; Goldberg et al. 2016).

1.4.2 eDNA capture

Filtration and precipitation are two broad methods that are used in the capture of eDNA from water (Figure 1.1). Comparative studies have generally shown that filtration approaches have higher sample throughput and can process greater water volumes than precipitation approaches, thereby increasing the potential to recover greater amounts of DNA (e.g., Hinlo et al. 2017; Spens et al. 2017). A wide range of filter types have been applied for the capture of eDNA from water (see more detail in Chapter 3, Section 3.1). The pore size of filters ranges from 0.22 to 20 µm with various materials such as cellulose nitrate, nylon, glass fibre, polycarbonate, polyvinylidene difluoride, and polyethersulfone (Turner et al. 2014; Mächler et al. 2015; Spens et al. 2017).

Most studies that have investigated the impact of different types and the pore sizes of filter on DNA quantity, focussed on individual target species using qPCR (e.g., Eichmiller et al. 2016; Lacoursière-Roussel et al. 2016b; Minamoto et al. 2016; Robson et al. 2016). Several studies have investigated if and how the choice of filtration method affects the estimate of community composition. Overall, different filter membrane materials do not affect estimates of species richness and community composition across multiple trophic levels (Djurhuus et al. 2017). The performance of 0.45 µm filters in representing the community composition is more consistent (Miya et al. 2016; Majaneva et al. 2018), while it is unclear if pre-filtration (i.e., size fractioning of particles through filters of different pore sizes) significantly affects the detected community composition. The key question is that the suitability of various pore sizes of the filter to capture eDNA may be heavily influenced by the heterogeneous nature of aquatic ecosystems.

1.4.3 Sampling design

As for any field study, the sampling design is of paramount importance, as it will impact the downstream statistical power and analytical interpretation of any eDNA study. In freshwater ecosystems, the detection probability of eDNA is influenced by the characteristics of its ecology, including the origin (physiological sources), state (physical forms), transport (physical movement), and fate (degradation) of eDNA molecules (see more detail in Chapter 4, Section 4.1). Spatial heterogeneity of eDNA within water bodies has been reported in several studies (e.g., Pilliod et al. 2013; Civade et al. 2016; Hänfling et al. 2016), which will result in imperfect species detection. Additionally, Sato et al. (2017) indicated that sample pooling from the whole lake reduces the detection probability of fish species. Site occupancy models have been developed in ecological studies to cope with multiple levels of bias and uncertainty (e.g. imperfect detection) (MacKenzie et al. 2002; MacKenzie & Royle 2005). The key data requirement in these analyses is that there are multiple visits with a number of sampling sites during a period when the true occurrence state (presence or absence) of a site is unlikely to change. Thus, the detection probability based on numbers of detection and non-detection sites allows estimating the true proportion of occupied sites. The site occupancy model has been applied to the analysis of eDNA data for imperfect species detection or estimating abundance (Pilliod et al. 2013; Schmidt et al. 2013; Ficetola et al. 2015; Hänfling et al. 2016; Valentini et al. 2016). Consequently, the better understanding of the spatial and temporal distribution of eDNA will greatly inform eDNA sampling strategies and ensure the accuracy and reliability of eDNA biodiversity assessments (Goldberg et al. 2018).

1.5 Integrating eDNA metabarcoding in biological monitoring of freshwater ecosystems

To determine the consequences of anthropogenic disturbances on ecosystem health throughout the world, the structure, function or some other characteristic of organisms (i.e., biological indicators) have been used for biological monitoring to define ecological status (Adams 2002; Bonada et al. 2006). For instance, the main European initiative for water quality assessment and improvements, Water Framework Directive 2000/60/EC (WFD), requires all member states to reach “good” ecological status of lakes, rivers and ground waters based on biological elements including phytoplankton, macrophytes and phytobenthos, benthic invertebrates, and fish (CEC 2000).

To address the requirements of WFD, a large number of biotic metrics/indices, based on the morphological identification of various groups of aquatic indicator organisms, have been developed in different countries (reviewed in Birk et al. 2012). However, there are some limitations when using conventional survey methods to obtain these biotic metrics/indices, such as morphological identification bias, recording small-bodied, rare and/or elusive species, and destructive impacts on the environment (Stribling et al. 2008; Deiner et al. 2017a). eDNA metabarcoding could potentially alleviate some of these limitations by using genetic information from environmental samples instead of morphology to identify organisms and to characterise a given ecosystem, which is a promising tool for rapid, non-invasive biodiversity monitoring (reviewed in Rees et al. 2014; Barnes & Turner 2016; Jackson et al. 2016; Deiner et al. 2017a; Hering et al. 2018).

Different aspects of the metabarcoding approach are disputable as discussed above (Section 1.4). In addition to these technical challenges, application of eDNA metabarcoding used in biological monitoring requires the parallel application of morphometric and eDNA-based identification in order to learn whether species richness, abundance, diversity, and assemblage composition estimates derived from the two methods result in a similar measure for metrics/indices of interest. Alternatively, new biotic indices need to be developed which simultaneously consider morphological and metabarcoding data (Deiner et al. 2017a; Pawlowski et al. 2018). In freshwater habitats, the comparison of the biotic indices inferred from morphological and metabarcoding data has been made using diatoms with a relatively good correlation (Apothéloz-Perret-Gentil et al. 2017; Vasselon et al. 2017). An effective fish-based biodiversity assessment tool for lakes remains problematic (Winfield 2002; Kelly et al. 2012) (see more detail in Chapter 5, Section 5.1). Therefore, it is difficult to compare the biotic indices inferred from morphological and metabarcoding data using fish. However, eDNA metabarcoding analysis can complement and overcome the limitations of conventional survey methods by targeting different species, sampling greater diversity, and increasing the resolution of taxonomic identifications (e.g., Hänfling et al. 2016; Lim et al. 2016; Olds et al. 2016; Shaw et al. 2016a; Evans et al. 2017). However, in order to integrate eDNA metabarcoding data to calculate conventional metrics/indices or for future intercalibration, a much larger dataset from a broader geographic area and more diverse habitats is required.

1.6 Thesis organisation and objectives

The main objective of this thesis is to develop an eDNA metabarcoding method for routine monitoring of freshwater fish communities to improve assessments of lentic ecosystem quality to fulfil WFD requirements. The reliable assessment is a key element for successful management of freshwater resources and restoration of damaged ecosystems. However, developing, validating and improving eDNA-based metabarcoding monitoring methods is not trivial, as it involves many steps from sample collection, laboratory protocols, bioinformatics analysis, as well as practical implementation and policy questions (reviewed in Deiner et al. 2017a; Hering et al. 2018; Pawlowski et al. 2018). Thus, my thesis mainly focus on the choice of metabarcodes and compilation of reference databases for UK freshwater fish (Chapter 2), determination of the suitable pore size of the filter for eDNA capture through metabarcoding (Chapter 3), understanding spatial and temporal distribution of eDNA inferred by metabarcoding (Chapter 4), and validation of eDNA metabarcoding analysis in diverse lakes with characterised fish faunas (Chapter 5).

1.6.1 Chapter 2

In Chapter 2, I compiled two curated reference databases (Cytb and 12S) of the UK freshwater fish with a reproducible workflow. After that, I *in silico* evaluated three and six metabarcoding primer pairs targeting Cytb and 12S, respectively, against curated reference databases and determined the suitable metabarcoding primer pairs for each locus. The successfully selected metabarcoding primer pairs were then *in vitro* tested in single DNA PCR amplifications on 22 common freshwater fish species and 10 mock communities via metabarcoding. The metabarcoding data of 10 mock communities were analysed with the custom reproducible metabarcoding bioinformatics analysis pipeline (metaBEAT) against curated reference databases. These two metabarcoding primer pairs targeting different mitochondrial genes, custom-made reference databases, and metaBEAT pipeline are used to investigate the other eDNA questions in the rest of this thesis.

1.6.2 Chapter 3

In Chapter 3, six treatments, with differing filter types and pore sizes for eDNA capture, were compared for their efficiency and accuracy to assess fish community structure through metabarcoding. Specifically, I investigated the impact of different pore sizes of the membrane filter, different types of filter, and pre-filtration on eDNA capture and community diversity estimation. I evaluated the effect on filtration time, total eDNA recovered, the probability of species detection, repeatability, and the relationship between read counts and known fish abundance or biomass in four fish ponds with differing assemblages. The filter type determined in this chapter to have the most suitable pore size is applied to all further studies in this thesis.

1.6.3 Chapter 4

In Chapter 4, I explored the spatial and temporal distribution of eDNA in fish ponds and evaluated the detection sensitivity of eDNA metabarcoding for low-density species. Specifically, I examined the shedding and decay rates of eDNA in fish ponds following the introduction and removal of two rare species. I also tried to understand the spatial and temporal changes of fish communities after the rare species introduction and removal. The results of this research are critical to understand the ecological characteristics of eDNA in ponds, including production, degradation and transport, and to inform effective sampling strategies for eDNA study.

1.6.4 Chapter 5

In Chapter 5, I explored the potential of eDNA metabarcoding as a tool for WFD status assessment by collecting and analysing water samples from eight Welsh lakes and six meres in Cheshire, England, with well-described fish faunas. In this study, I tried to use a possible approach to evaluate the confidence of species presence based on site occupancy and read counts and proposed to use a five-level classification scale to estimate species relative abundance, so that they can be used to compute biotic metrics/indices for WFD assessment approaches in the future. I also investigated the effectiveness of different spatial sampling approaches, particularly comparing shore and offshore sampling to understand the sampling effort required for eDNA study.

1.6.5 Chapter 6

In Chapter 6, I discussed the main findings of the studies with current knowledge of eDNA from other studies to fully exploit the potential of metabarcoding data and improve the accuracy and precision of their analysis for the future integration of eDNA metabarcoding to routine biological monitoring programmes.

Chapter 2 Development of the reference database and metabarcoding primers for targeting UK freshwater fish communities[1]

Abstract

The choice the metabarcode and compilation of the curated reference database are important methodological considerations for metabarcoding. Two curated reference databases (Cytb and 12S) of UK freshwater fish werecompiled with a reproducible workflow. The reproducible workflow allows updating or adding new sequences into the reference databases at any time and can also be applied to generate local reference databases for other fish communities and in fact other taxonomic groups and markers. Moreover, two metabarcoding primer pairs targeting different mitochondrial genes have been fully *in silico* tested against the curated reference databases, and *in vitro* on 22 species and 10 mock communities. The data of mock communities were analysed with a custom reproducible metabarcoding bioinformatics analysis pipeline (metaBEAT). Except the potential primer bias with nine-spined stickleback *Pungitius pungitius* when using the Cytb primes targeting 412-bp region, together with two marker-specific curated reference databases and metaBEAT, these two metabarcoding primer pairs are selected for metabarcoding studies to investigating other eDNA questions.

[1] Partial results of this chapter have been published as Supporting Information in Hänfling, B., Lawson Handley, L., Read, D.S., Hahn, C., ***Li, J.***, Nichols, P., Blackman, R.C., Oliver, A. & Winfield, I.J. (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, 25, 3101-3119. https://doi.org/10.1111/mec.13660

2.1 Introduction

DNA barcoding is a diagnostic technique, which can be used for taxonomic identification of species usually using Sanger sequencing of short standard DNA sequences (i.e., DNA barcodes) from individual specimens. This technique is now considered to be a powerful tool both for taxonomical and ecological studies (Hebert et al. 2003a; Hebert et al. 2003b; Valentini et al. 2009; Pečnikar & Buzan 2014). DNA metabarcoding is an extension of this approach for taxonomic identification of multiple species extracted from a mixed sample (community DNA or environmental DNA "eDNA") coupled with sequencing on a high-throughput sequencing (HTS) platform (e.g., Illumina, Ion Torrent) (Taberlet et al. 2012b; Yu et al. 2012). Within the last decade, metabarcoding has increasingly been applied to biodiversity monitoring, with enormous potential to inform nature conservation and management (e.g., Ji et al. 2013; Deiner et al. 2017a; Elbrecht et al. 2017).

DNA barcodes must contain enough information to discriminate between closely related species and to discover new ones. Similarly, metabarcodes need to provide taxonomic resolution for all species of the target group, and be flanked by two conserved regions to enable simultaneous amplification of all target taxa without bias while preventing that of non-target organisms (Clarke et al. 2014; Deagle et al. 2014). In the case of eDNA studies, several metabarcoding primers amplifying different short regions of multi-copy DNA have been tested to minimise taxonomic biases when targeting genetically diverse taxonomic groups, while facilitating the amplification of potentially degraded eDNA (Coissac et al. 2012; Yoccoz 2012).

The mitochondrial cytochrome c oxidase subunit I gene (COI) is widely accepted as the standard barcoding marker for most metazoans (Hebert et al. 2003b) and recommended by the Consortium for the Barcode of Life (CBoL). However, for DNA metabarcoding, COI is also widely, but not unanimously accepted as the standard metabarcode for metazoans (Clarke et al. 2014; Deagle et al. 2014; Elbrecht et al. 2016), although it is recommended by Andújar et al. (2018) for metazoan community metabarcoding study with bulk samples considering the large COI specific reference databases, the broader taxonomic coverage and resolution, combined with recent improvements of primer design in COI region. However, with environmental samples, most primers targeting the COI region amplify large proportions of microbial species (e.g., Yang et al. 2014; Stat et al. 2017). This fact remains the strongest reason for the use of other genes such as mitochondrial cytochrome *b* gene (Cytb), mitochondrial 12S and 16S rRNA genes (hereafter referred to as 12S and 16S, respectively). A number of studies have adopted the universal primers described by Kocher et al. (1989) to target the Cytb locus (e.g., Irwin et al. 1991; Burgener & Hübner 1998; Hsieh et al. 2001). The "12S" and "16S" regions are more highly conserved than Cytb and COI loci in vertebrates due to the stem-loop structures, leading to a variation of short stretches of highly conserved and stretches of highly variable DNA (Hickson et al. 1996; Springer & Douzery 1996; Burk et al. 2002). An increasing number of studies have indicated that 12S and 16S could be good candidate regions for metabarcodes in animals (Riaz et al. 2011; Epp et al. 2012; Clarke et al. 2014; Deagle et al. 2014; Kocher et al. 2017).

Once the locus or loci are chosen, primers are then designed based on the taxonomic group(s) of interest within a study. The ecoPrimers software (Riaz et al. 2011) could be a useful and efficient tool for identifying new barcode markers and their associated PCR primers. After designing new primers or choosing previously designed primers, it is important to perform rigorous testing, *in silico*, *in vitro*, and *in situ* to infer their utility for metabarcoding (Freeland 2016; Goldberg et al. 2016). Currently, ecoPCR (Ficetola et al. 2010) is the most popular and versatile bioinformatics tool for evaluating taxonomic coverage and resolution of available metabarcodes considering the target and non-target groups of interest.

One prerequisite for *in silico* design and test of metabarcoding primers is to have a reliable, curated reference database to which the unknown DNA sequences are compared against to retrieve the taxonomic composition from sequence data. There are three major public databases under the International Nucleotide Sequence Database Collaboration (INSDC) which are the European Molecular Biology Laboratory (EMBL) database and National Center for Biotechnology Information (NCBI) GenBank database, as well as the DNA Data Bank of Japan (DDBJ). Unfortunately, there is a growing problem of taxonomic misidentification and insufficient annotations in these public DNA databases due to mislabelling, PCR-based errors including chimeric sequences, sequencing errors, and contamination of environmental sequences (Vilgalys 2003; Nilsson et al. 2006; Mioduchowska et al. 2018). Limiting the impact of such errors is essential for both DNA metabarcoding surveys and phylogenetic studies. Therefore, researchers have started to establish gene-specific reference databases with high-quality and reliable sequences such as SILVA for nuclear rRNA genes (Quast et al. 2012), the Barcode of Life Data System (BOLD) for COI (Ratnasingham & Hebert 2007), and the UNITE for the internal transcribed spacer (ITS) region (Kõljalg et al. 2005). Another drawback of these large public databases is that sequence redundancy within them increases computational resource use, and is particularly problematic for software programmes that classify sequences based on a set number of top alignments. To limit these drawbacks, a curated non-redundant reference database for a given gene is required for metabarcoding.

This study aims to (1) compile curated Cytb and 12S reference databases of UK freshwater fish with a reproducible workflow; (2) *in silico* evaluate species taxonomic resolution of published or new Cytb and 12S metabarcodes against curated reference databases, (3) *in vitro* test the performance of successfully selected metabarcoding primers on individual DNA of 22 species using PCR and 10 mock communities via metabarcoding, and (4) test a custom reproducible metabarcoding bioinformatics analysis pipeline against curated reference databases.

2.2 Materials and methods

2.2.1 Compilation and curation of reference databases

Freshwater fish DNA sequences from two regions of mtDNA (Cytb and 12S) were compiled as reference databases for metabarcoding and to compare the suitability of different markers and primers for such approaches. The workflow of compilation and curation of reference databases is shown in Figure 2.1.

At first, a list of 72 target species was compiled based on expert opinion (Pierson G. & Winfield I.J. pers. comm.). The target species included all fish previously recorded in UK freshwaters and additional non-native species that could potentially be present but have not yet been confirmed (Table 2.1). All available Cytb and 12S sequences of the target species were retrieved from Genbank using E-utilities (Sayers 2008). Subsequently, gaps in coverage of the target species list were identified, and new sequence data were generated from existing tissue/fin clip collections at the University of Hull (UoH) and the Centre for Ecology and Hydrology (CEH), or from freshly collected tissues/fin clips. As no significant gaps were identified among existing Cytb sequences, new sequences were only generated for the 12S database (Table 2.1).

Fish DNA was extracted from fin clips or muscle tissues using a DNeasy Blood & Tissue kit (QIAGEN). A set of novel primers was designed from an alignment of whole mitochondrial fish genomes (12S_30F: CACTGAAGMTGYTAAGAYG and

12S_1380R: CTKGCTAAATCATGATGC) in order to generate reference sequences of the entire 12S region. Polymerase chain reactions (PCRs) were performed in 25 µL volumes containing: 1 × NH_4 Buffer, 2 mM $MgCl_2$, 1 mM total dNTPs, 0.8 µM of each primer, 1 U BIOTAQ polymerase (Bioline), and ~10 ng DNA template. PCRs were performed on an Applied Biosystems Veriti Thermal Cycler with the following profile: 95 °C for 2 min, 30 cycles of 95 °C for 30 sec, 50 °C for 30 sec and 72 °C for 50 sec, followed by a final elongation step at 72 °C for 10 min. Purified PCR products were Sanger sequenced directly (Macrogen Inc., Republic of Korea) in both directions using the PCR primers. Forward and reversed sequences were aligned and edited using CodonCode Aligner (CodonCode Corporation, USA). Eventually, validated *de novo* 12S sequences were submitted to NCBI and combined with those mined from Genbank to create the raw 12S reference database.

The raw reference databases were further processed in the ReproPhylo environment (Szitenberg et al. 2015) in order to produce a set of non-redundant quality checked reference sequences for both markers. Sequences were extracted in FASTA format and clustered at 100% identity to remove redundancy using CD-hit-est v4.6.1 (Fu et al. 2012). After final quality control checking outliers, based on the distribution of sequence length (minimum length of 100 bp for Cytb, 50 bp for 12S), the remaining sequences were aligned using MAFFT v7.0 (Katoh & Standley 2013). For Cytb, nucleotide sequences were translated to protein sequences prior to alignment, and aligned protein sequences were converted back to nucleotide sequences using PAL2NAL v14 (Suyama et al. 2006) to check if any of the sequences were pseudogenes. Alignments were trimmed to remove poorly aligned regions using trimAl v1.2 (Capella-Gutiérrez et al. 2009) based on the consistency across the sequences. A phylogenetic approach was then used to identify potentially erroneous records (i.e., records which were likely mislabelled). First, maximum likelihood trees of trimmed alignments were inferred with RAxML v8.0.2 (Stamatakis 2006) using the GTR + gamma model of substitutions. Resulting trees were investigated to identify any sequence records that were potentially misplaced in the phylogenetic trees using SATIVA v0.9 (Kozlov et al. 2016). Polyphyletic groups identified by SATIVA v0.9 (Kozlov et al. 2016) were then manually inspected to decide whether this could be a biological reality. Sequences deemed to be erroneous were removed from the databases. The remaining sequences (i.e., the curated non-redundant reference databases) were used in downstream analyses. To facilitate full reproducibility of analyses, the Jupyter notebooks are provided to illustrate how to prepare custom curated reference databases in the dedicated GitHub repository (https://github.com/HullUni-bioinformatics/Curated_reference_databases).

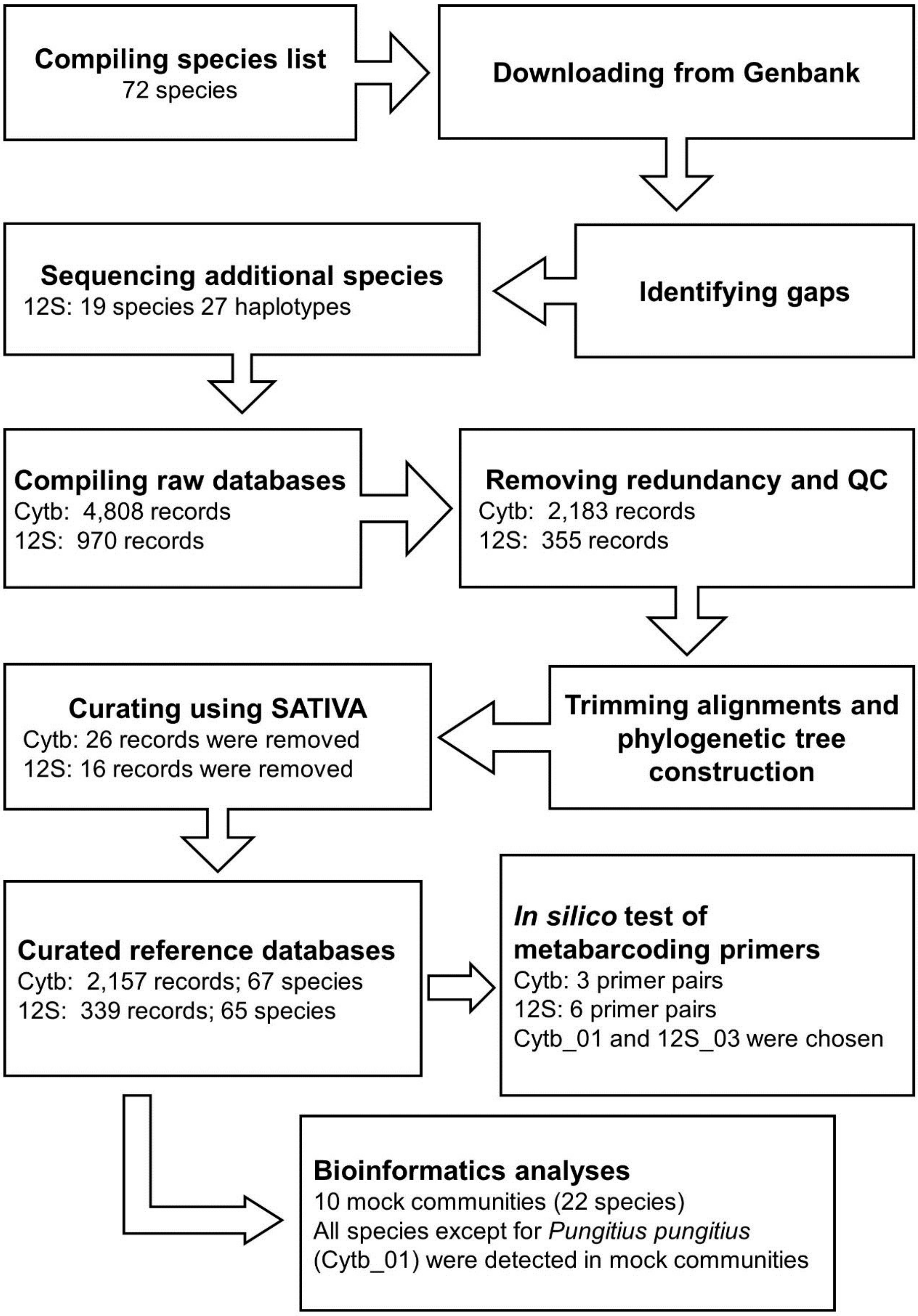

Compiling species list
72 species
Downloading from Genbank
Identifying gaps
Sequencing additional species
12S: 19 species 27 haplotypes
Compiling raw databases
Cytb: 4,808 records
12S: 970 records
Removing redundancy and QC
Cytb: 2,183 records
12S: 355 records
Trimming alignments and phylogenetic tree construction
Curating using SATIVA
Cytb: 26 records were removed
12S: 16 records were removed
Curated reference databases
Cytb: 2,157 records; 67 species
12S: 339 records; 65 species
In silico test of metabarcoding primers
Cytb: 3 primer pairs
12S: 6 primer pairs
Cytb_01 and 12S_03 were chosen
Bioinformatics analyses
10 mock communities (22 species)
All species except for Pungitius pungitius (Cytb_01) were detected in mock communities

Figure 2.1 Overview of the workflow of the compilation and curation of reference databases for UK freshwater fish and downstream analysis in this study.

Table 2.1 List of species included in curated reference databases.

Scientific name	Common name	Reference database	Number in Figure 2.1	12S sequenced during current project
Abramis brama	Common bream	Cytb/12S/ mtDNA	5	Yes
Acipenser sturio	Common sturgeon	Cytb/12S		
Alburnoides bipunctatus	Schneider	Cytb/12S		
Alburnus alburnus	Common bleak	Cytb/12S/ mtDNA	20	
Alosa alosa	Allis shad	Cytb/12S		
Alosa fallax	Twaite shad	Cytb/12S		
Ambloplites rupestris†	Rock bass	Cytb/12S		
Ameiurus melas†	Black bullhead	Cytb/12S		
Ameiurus nebulosus†	Brown bullhead	Cytb/12S	17	Yes
Anguilla anguilla	European eel	Cytb/12S/ mtDNA		
Aspius aspius	Asp	Cytb		
Barbatula barbatula	Stone loach	Cytb/12S		Yes
Barbus barbus	Barbel	Cytb/12S/ mtDNA	21	Yes
Blicca bjoerkna (= *Abramis bjorkna*)	Silver bream	Cytb/12S/ mtDNA		
Carassius auratus†	Goldfish	Cytb/12S/ mtDNA		
Carassius carassius†	Crucian carp	Cytb/12S/ mtDNA		
Chondrostoma nasus	Nase	Cytb		
Cobitis taenia	Spined loach	Cytb/12S		
Coregonus albula	Vendace	Cytb/12S	4	Yes
Coregonus autumnalis	Pollan	Cytb/12S		
Coregonus lavaretus	Whitefish	Cytb/12S/ mtDNA		
Coregonus oxyrinchus	Houting	Cytb/12S/ mtDNA		
Cottus gobio	Bullhead	Cytb/12S	10	Yes
Ctenopharyngodon idella†	Grass carp	Cytb/12S		
Cyprinus carpio†	Common carp	Cytb/12S/ mtDNA	18	Yes
Esox lucius	Pike	Cytb/12S/	1	Yes

		mtDNA		
Gasterosteus aculeatus	Three-spined stickleback	Cytb/12S/ mtDNA		
Gobio gobio	Gudgeon	Cytb/12S/ mtDNA	19	Yes
Gymnocephalus cernua	Ruffe	Cytb/12S/ mtDNA	2	Yes
Hypophthalmichthys molitrix†	Silver carp	Cytb/12S/ mtDNA		
Hypophthalmichthys nobilis†	Bighead carp	Cytb/12S/ mtDNA		
Lampetra fluviatilis	River lamprey	Cytb/12S/ mtDNA		
Lampetra planeri	Brook lamprey	Cytb		
Lepomis gibbosus†	Pumpkinseed	Cytb/12S	11	Yes
Leucaspius delineatus†	Sunbleak	Cytb/12S/ mtDNA	12	
Leuciscus idus†	Orfe	Cytb/12S/ mtDNA		
Leuciscus leuciscus	Dace	Cytb/12S	14	Yes
Lota lota	Burbot	Cytb/12S/ mtDNA		
Micropterus salmoides†	Largemouth bass	Cytb/12S/ mtDNA		
Misgurnus bipartitus	Northern weatherfish	12S		
Misgurnus fossilis†	Weather loach	Cytb/12S		
Neogobius fluviatilis	Monkey goby	12S		
Neogobius melanostomus†	Round goby	Cytb/12S		
Oncorhynchus gorbuscha†	Pink salmon	Cytb/12S/ mtDNA		
Oncorhynchus mykiss†	Rainbow trout	Cytb/12S/ mtDNA		
Osmerus eperlanus	Smelt	Cytb/12S		
Perca fluviatilis	Perch	Cytb/12S/ mtDNA	3	Yes
Petromyzon marinus	Sea lamprey	Cytb/12S/ mtDNA		
Phoxinus phoxinus	Minnow	Cytb/12S/ mtDNA	8	
Pimephales promelas	Fathead minnow	Cytb/12S		
Platichthys flesus	Flounder	Cytb/12S		
Ponticola kessleri	Bighead goby	Cytb/12S		
Pomatoschistus minutus	Sand goby	12S		

Proterorhinus semilunaris†	Western tubenose goby	Cytb		
Pseudorasbora parva†	Topmouth gudgeon	Cytb/12S/mtDNA	13	Yes
Pungitius pungitius	Nine-spined stickleback	Cytb/12S/mtDNA	15	Yes
Rhodeus amarus†	Bitterling	12S		
Rhodeus sericeus†	Amur bitterling	Cytb/mtDNA		
Rutilus rutilus	Roach	Cytb/12S	6	
Salmo salar†	Atlantic salmon	Cytb/12S/mtDNA		
Salmo trutta	Brown trout	Cytb/12S/mtDNA	7	Yes
Salvelinus alpinus	Arctic charr	Cytb/12S/mtDNA		
Salvelinus fontinalis†	Brook charr	Cytb/mtDNA		
Sander lucioperca†	Pikeperch (zander)	Cytb/12S/mtDNA		
Scardinius erythrophthalmus	Rudd	Cytb/12S		Yes
Silurus glanis†	Wels catfish	Cytb/12S/mtDNA		
Solea solea	Common sole	12S		
Squalius cephalus (=*Leuciscus cephalus*)	Chub	Cytb/12S	22	Yes
Thymallus thymallus	Grayling	Cytb/12S/mtDNA		
Tinca tinca	Tench	Cytb/12S/mtDNA	9	Yes
Umbra pygmaea	Mudminnow	Cytb/12S/mtDNA	16	
Vimba vimba	Vimba bream	Cytb		

Notes: "†" indicates non-native species in UK according to the Non-native Species Secretariat (http://www.nonnativespecies.org/)

2.2.2 *In silico* test of metabarcoding primers

To test the suitability of primers for eDNA-based metabarcoding of UK freshwater fish communities, a total of nine primer combinations (three for Cytb, six for 12S) combined from published and novel primers (Table 2.2) were evaluated *in silico* against the curated reference databases. The relative location of these primer pairs is shown in

Figure 2.2. Although some primer combinations have previously been evaluated against a broader database by Taberlet et al. (2018), the purpose of this study is to explicitly test the utility for metabarcoding of UK fish communities. The programme ecoPCR v0.2 (Ficetola et al. 2010) of OBITools v1.01.22 (Boyer et al. 2016) was used to evaluate the conservation of primer binding sites and species resolution of the metabarcodes. When using ecoPCR (Ficetola et al. 2010), the maximum number of mismatches allowed per primer was set to three, in order to later evaluate how conserved the primers are across fish taxa in the curated reference databases. The results from ecoPCR (Ficetola et al. 2010) were checked and visualised in R v3.5.0 (R_Core_Team 2018). The full R script is available on the GitHub repository (https://github.com/HullUni-bioinformatics/Curated_reference_databases/tree/master/R_script).

To evaluate conservation of the metabarcoding primer pairs and the capacity of the metabarcode to discriminate between taxa, two indices were used in this study: the coverage index (*Bc*), corresponding to the ratio of the number of amplified target taxa (i.e., species level for this analysis) to the total number of target taxa in the curated database, and the specificity index (*Bs*), defined as the ratio of taxonomically discriminated taxa to the number of amplified taxa. It must be noted that these two ratios are highly dependent upon the reference database they are estimated from, such as the length of sequence. For instance, this approach did not work with the Cytb_01 primer pairs (L14912 and H15149) (Table 2.2; Kocher et al. 1989), since a large proportion of the sequences in the Cytb curated database did not cover the forward primer binding sites used for their amplification. Therefore, this primer pair was evaluated with a reduced database which only included complete mitochondrial genomes (mitogenomes).

Table 2.2 Metabarcoding primer sequences of Cytb and 12S tested in this study.

Primer ID	Primer	Sequence 5’–3’	Reference
Cytb_01	L14912	AAAAACCACCGTTGTTATTCAACTA	Kocher et al. (1989)
	H15149	GCDCCTCARAATGAYATTTGTCCTCA	Kocher et al. (1989)
Cytb_02	Fish2CBL	ACAACTTCACCCCTGCAAAC	Thomsen et al. (2012a)
	Fish2bCBR	GATGGCGTAGGCAAACAAGA	Thomsen et al. (2012a)
Cytb_03	Fish2degCBL	ACAACTTCACCCCTGCRAAY	Thomsen et al. (2012a)
	Fish2CBR	GATGGCGTAGGCAAATAGGA	Thomsen et al. (2012a)
12S_01	Tele02_F	AAACTCGTGCCAGCCACC	Taberlet et al. (2018)
	Tele02_R	GGGTATCTAATCCCAGTTTG	Taberlet et al. (2018)
12S_02	MiFish-U_Fa†	GCCGGTAAAACTCGTGCCAGC	this study
	MiFish-U_R	CATAGTGGGGTATCTAATCCCAGTTTG	Miya et al. (2015)
12S_03	12S_V5_F	ACTGGGATTAGATACCCC	Riaz et al. (2011)
	12S_V5_R	TAGAACAGGCTCCTCTAG	Riaz et al. (2011)
12S_04	12S-V5_F	ACTGGGATTAGATACCCC	Riaz et al. (2011)
	12S-V5_R2	CTACACCTCGACCTGACG	this study
12S_05	12S-V5_F	ACTGGGATTAGATACCCC	Riaz et al. (2011)
	Ac12s_R	GAGAGTGACGGGCGGTGT	Evans et al. (2016)
12S_06	Teleo_F	ACACCGCCCGTCACTCT	Valentini et al. (2016)
	Teleo_R	CTTCCGGTACACTTACCATG	Valentini et al. (2016)

Notes: “†” indicates this primer adapted from MiFish-E_F in Miya et al. (2015).

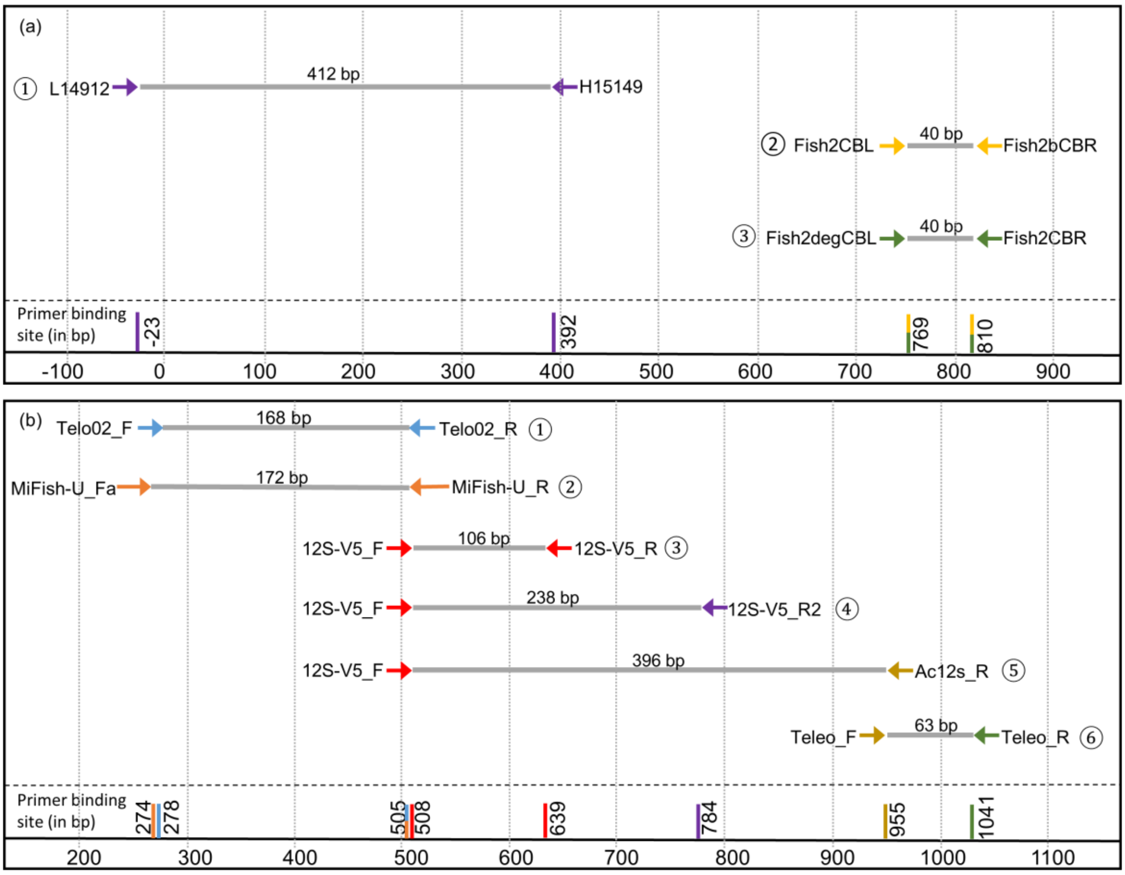


Figure 2.2 The relative location of (a) Cytb and (b) 12S metabarcoding primer pairs tested in this study. The primer binding site is according to the position in the alignments of reference databases constructed by this study. The primer IDs and sequences are described in Table 2.2.

2.2.3 *In vitro* test results of metabarcoding primers

Primer pairs selected based on *in silico* results (i.e., Cytb_01 and 12S_03, Table 2.3, see more detail in Section 2.3.2) were then tested *in vitro* on 22 species (Table 2.1): first in single DNA PCR amplification to check consistency of amplification across taxa, and second in 10 mock communities to evaluate whether all species amplified in competitive mixed assemblages. In single DNA PCR amplification, DNA concentration of each species was normalised to 5 ng μL^{-1} based on NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific) readings. PCR reagent concentrations were identical to those given above (see more detail in Section 2.2.1). PCRs were performed on an Applied Biosystems Veriti thermal cycler with the gradient profile to determine the optimal annealing temperature consisted of an initial denaturation at 98 °C for 5 min followed by 30 cycles with 15 sec at 98 °C, 20 sec at the annealing temperature (48, 50, 52, 54, 56, and 58 °C) and 30 sec at 72 °C, and a final extension step of 7 min at 72 °C. The 10 mock communities were generated from three different DNA concentrations (10 ng μL^{-1}, 5 ng μL^{-1} and 0.5 ng μL^{-1}) of the 22 species (Table 2.1) with different species composition (Table 2.4 & Table 2.5).

2.2.3.1 Library preparation and sequencing

The 10 mock community samples were sequenced via metabarcoding with the Cytb_01 and 12S_03 primer pairs. Both libraries were PCR-amplified with a one-step library preparation protocol (Kozich et al. 2013). Three PCR technical replicates were performed for each mock community sample then pooled to minimise noise (see more detail in Box 1) in individual PCRs. All PCRs were set up in a PCR workstation with UV hood and high-efficiency particulate air (HEPA) filter in the eDNA laboratory at UoH. Eight-strip PCR tubes with individually attached lids and mineral oil (Sigma-Aldrich) were used to reduce cross-contamination between samples. PCR reactions were carried out in 25 μL volumes containing: 12.5 μL of 2 × Q5 High-Fidelity PCR Kit (New England Biolabs), 0.5 μM of each tagged primer, 2.5 μL template DNA and 7.5 μL molecular grade water. PCRs were performed on an Applied Biosystems Veriti thermal cycler with the following profile: 98 °C for 5 min, 35 cycles of 98 °C for 10 sec, 50 °C (Cytb_01) or 58 °C (12S_03) for 20 sec and 72 °C for 30 sec, followed by a final elongation step at 72 °C for 7 min. PCR products were purified and normalised using the SequalPrep Normalisation Plate Kit (Invitrogen) and subsequently pooled in equal volume (i.e., 5 uL per sample). The pooled library for each locus was further purified using the QIAquick Gel Extraction Kit (Qiagen), and each library was resuspended in 20 µL elution buffer. Libraries were then quantified by QUBIT v3.0 using the dsDNA HS Assay Kit (Thermo Fisher Scientific) and the KAPA Illumina Library Quantification Kit (KAPA Biosystems) on a Roche LightCycler Real-Time PCR machine using manufacturer's guidelines. Libraries were respectively run at a 6 pM concentration on an Illumina MiSeq at CEH using the v3 chemistry (2 × 300 cycles). To improve clustering during the initial sequencing cycles, 10% PhiX genomic control was added to each library. The custom sequencing and index primers were added to the appropriate wells of the MiSeq reagent cartridge as described in Kozich et al. (2013).

2.2.3.2 Bioinformatics and data analyses

Raw read data of the 10 mock communities from Illumina MiSeq sequencing have been submitted to NCBI (BioProject: PRJNA313432; BioSample accession numbers: SAMN04530501–SAMN04530510). Bioinformatics analysis was implemented following a custom reproducible metabarcoding pipeline (metaBEAT v0.80) with custom-made reference databases (Cytb and 12S) (see more detail in Appendix S2.1 Section S2.1.1; Hänfling et al. 2016). Sequences for which the best BLAST (see more detail in Box 2) hit had a bit score below 80 or had less than 95%/100% (Cytb/12S) similarity to any sequence in the curated databases were considered non-target sequences. The different level of similarity for each locus depended on the length of metabarcode and the knowledge of intraspecific diversity of the studied taxon. To assure full reproducibility of our bioinformatics analysis, the Jupyter notebooks for data processing have been deposited in an additional dedicated GitHub repository (https://github.com/HullUni-bioinformatics/Haenfling_et_al_2016). Filtered data were summarised into the number of sequence reads per species for downstream analyses. All statistical analyses were performed in R v3.3.2 (R_Core_Team 2016), and graphs were plotted using GGPLOT2 v2.2.1 (Wickham & Chang 2016). Relationships between observed and expected read count proportions for the two loci were investigated by calculating the Pearson's product-moment correlation coefficient.

Box 1. PCR artefacts

- PCR noise and bias

The basic types of PCR artefacts can be divided into two categories: those resulting in sequence artefacts (PCR noise or known as PCR error) which lead to a misrepresentation of an actual sequence, and those skewing the relative abundances of PCR products due to unequal amplification (PCR bias). Sequence artefacts may arise due to (1) the formation of chimerical molecule, (2) the formation of heteroduplex molecules, and (3) nucleotide transition probabilities resulting from *Taq* DNA polymerase error (Qiu et al. 2001; Edgar et al. 2011). PCR bias is thought to be due to PCR primer mismatches in the amplification efficiency of mixed templates (Polz & Cavanaugh 1998) or to the inhibition of amplification by the self-annealing of the most abundant templates in the late stages of amplification (Suzuki & Giovannoni 1996).

- PCR chimeras

PCR chimeras are by-products of the PCR amplification process from two or more parental sequences (chimeric), most commonly produced through an incomplete extension step (Edgar et al. 2011). Theoretically, PCR-generated chimeras should be fewer in amplifications with DNA polymerases with higher processitivity and decrease as elongation time increases and cycle number decreases (Qiu et al. 2001). It has been shown that when unique reads, such as chimeras and singletons, are withheld in analysis, the estimation of diversity can be severely inflated (Kunin et al. 2010). The nature of the chimeric sequences, which can be present as high-quality reads, does not enable their removal directly through quality-based end trimming (Coissac et al. 2012).

Box 2. Taxonomic assignment

- blastn

Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The programme compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families (Altschul et al. 1990). Using a heuristic method, BLAST finds similar sequences, by locating short matches between the two sequences. This process of finding similar sequences is called seeding. It is after this first match that BLAST begins to make local alignments. The main idea of BLAST is that there are often high-scoring segment pairs (HSP) contained in a statistically significant alignment. BLAST searches for high scoring sequence alignments between the query sequence and the existing sequences in the database. By default, BLAST reports all the database sequences that match a query sequence sufficiently well to within a specified level of quality (usually defined through an E-value cutoff).

BLAST is actually a family of programmes include: nucleotide-nucleotide BLAST (blastn) which given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies such as GenBank (Benson et al. 2013); protein-protein BLAST (blastp) which given a protein query, returns the most similar protein sequences from the protein database that the user specifies such as Pfam (Finn et al. 2010), position-specific iterative BLAST (PSI-BLAST), nucleotide 6-frame translation-protein (blastx), nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx), protein-nucleotide 6-frame translation (tblastn), and large numbers of query sequences (megablast).

Box 2. Taxonomic assignment (continued)

- Taxonomic assignment

Taxonomic assignment is that identification of High-Throughput Sequencing (HTS) reads is achieved through a comparison of anonymous Molecular Operational Taxonomic Unit (MOTU) clusters/centroid sequences or direct comparisons of reads remaining after quality filtering against a reference database. Depending on the taxon of study and the marker used, the reference database may consist of publicly available sequences or study-generated reference sequences.

There are three main approaches that compare query sequences to database sequences for the purpose of assigning a taxon label: (1) sequence similarity search via alignment methods like BLAST or similarity searches using profile hidden Markov model (pHMM) such as jMOTU/Taxonerator (Jones et al. 2011) and MG-RAST (Glass et al. 2010), and several programmes incorporate of the lowest-common ancestor (LCA) algorithm first pioneered by MEGAN (Huson et al. 2007); (2) sequence composition approaches using interpolated Markov models (IMMs), naive Bayesian classifiers, and k-means/k-nearest-neighbour algorithms such as RDP (Wang et al. 2007), PhymmBL (Brady & Salzberg 2009), and TACOA (Diaz et al. 2009); and (3) phylogenetic methods which attempt to “place” a query sequence on a phylogenetic tree and determine where the query best “fits” in the phylogeny according to a model of evolution using maximum likelihood (ML), Bayesian methods, or other methods such as neighbour-joining (NJ) such as EPA (Berger et al. 2011), FastTree (Price et al. 2009), and pplacer (Matsen et al. 2010). A number of widely used programmes use combinations of these methods; for example, the programme SAP (Munch et al. 2008) uses BLAST searches of the NCBI database and phylogenetic reconstruction to establish taxonomic identity of query sequences.

2.3 Results and discussion

2.3.1 Reference databases

The validated *de novo* complete 12S sequences of 27 haplotypes from 19 species were submitted to NCBI (Genbank accession numbers: MH918114–MH918140) and combined with those mined from Genbank to create the raw 12S reference database. The raw reference database included 4,808 and 997 partial or complete sequences for Cytb and 12S, respectively. After removing redundant sequences (i.e., identical haplotypes) the reference databases contained 2,183 and 355 sequences for Cytb and 12S. A further 26 sequences in the Cytb database and 16 sequences in the 12S database were identified as likely mislabelled records based on phylogenetic tree inference and removed (the removed records can be found in Appendix S2.1 Section S2.1.2). The final curated reference database contained 2,157 and 339 sequences for Cytb (Appendix S2.2) and 12S (Appendix S2.3) and covered 67 species for Cytb and 65 species for 12S, respectively (Table 2.1). Additionally, there were 125 mitogenomes of 38 species (Table 2.1) in the reduced reference database (Appendix S2.4) for *in silico* testing with the Cytb_01 primer pairs (L14912 and H15149) (Kocher et al. 1989).

2.3.2 *In silico* test results of metabarcoding primers

The summary statistics obtained from ecoPCR (Ficetola et al. 2010) are shown in Table 2.3. Among the three Cytb primer pairs, the Cytb_01 was *in silico* tested with the reduced reference database including 38 species, since a large proportion of the sequences in the Cytb curated database did not cover the forward primer binding sites. Six species (pike *Esox lucius*, three-spined stickleback *Gasterosteus aculeatus*, river lamprey *Lampetra fluviatilis*, orfe *Leuciscus idus*, sea lamprey *Petromyzon marinus*, and nine-spined stickleback *Pungitius pungitius*) potentially cannot be *in silico* amplified due to one or two insertions at the forward primer binding site (see more detail in Appendix S2.5 Section S2.5.1). The Cytb_02 and Cytb_03 are located in the same position (Figure 2.2a). Both of these two Cytb primer pairs resulted in low coverage (≤ 50%) of species-level, even though the forward primer of the Cytb_03 is degenerate (Table 2.3). The coverage of the Cytb_01 was 84.21% although *in silico* testing with 38 species. Moreover, the Cytb_01 primer pair should be universal among vertebrates (Kocher et al. 1989; Burgener & Hübner 1998), and has been successfully tested in small mesocosms with eight fish and one amphibian via eDNA metabarcoding (Evans et al. 2016). It was therefore selected for metabarcoding of mock communities.

The coverage of the six 12S primer pairs was more than 95% apart from the 12S_02 (Table 2.3). The 12S_01 optimised from 12S_02 resulted in better coverage. However, the low coverage (83.07%) of the 12S_02 could be explained by the sequences of those species in the curated 12S database are missing part of the forward primer binding site (see more detail in Appendix S2.5 Section S2.5.5). The metabarcodes amplified by the six 12S primer pairs ranged in length from 63 to 396 bp with high specificity species resolution (*Bs* > 92%) apart from the 12S_06. The 12S_06 was ruled out due to the potential problem in the taxonomic resolution of the metabarcode (*Bs* = 85.24%) (Table 2.3). Compared to the 12S_03, the metabarcodes which were amplified by the 12S_01 and 12S_02 offered higher taxonomic resolution power since they can distinguish perch *Perca fluviatilis* and pikeperch *Sander lucioperca*. Unfortunately, these two primer pairs were unavailable when I performed the *in silico* test in 2014. As far as I know, several research groups are *in situ* evaluating the 12S_02 and 12S_03 including UoH and the University of Salford. The preliminary results from UoH indicate that there is no significant difference in species detection probability between the 12S_02 and 12S_03 in Lake Windermere (England, UK) and New Lake of Marchamley Pools

(England, UK), but the difference of community composition is under investigation (Di Muri C. pers. comm.). The forward primer among the 12S_03, 12S_04 and 12S_05 are same, but the reverse primer of 12S_03 is more conserved (i.e., fewer mismatches) than 12S_04 and 12S_05 (see more detail in Appendix S2.5 Section S2.5.6–8). Considering primer bias (i.e., reduced amplification of the species during PCR) and taxonomic resolution, the 12S_03 was selected for metabarcoding of mock communities.

Table 2.3 *In silico* test results of metabarcoding primers.

Primer ID	**Metabarcode length**	**T$_A$**	**Coverage *Bc***	**Specificity *Bs***	**Unresolved species pairs**
Cytb_01§	*ca.* 412 bp	50 °C	84.21%	93.75%	1
Cytb_02	*ca.* 40 bp	50 °C	37.31%	92.00%	2
Cytb_03	*ca.* 40 bp	50 °C	44.78%	86.67%	2, 3
12S_01	*ca.* 168 bp	54 °C	95.38%	95.16%	1
12S_02	*ca.* 172 bp	61 °C	83.07%†	94.44%	1
12S_03	*ca.* 106 bp	58 °C	95.38%	92.18%	1, 4
12S_04	*ca.* 238 bp	55 °C	95.38%	95.16%	1
12S_05	*ca.* 396 bp	55 °C	95.38%	96.77%	1
12S_06	*ca.* 63 bp	55 °C	93.84%	85.24%	1, 2, 5, 6

Notes: T$_A$ represents recommended annealing temperature. Unresolved species pairs: 1 = Coregonus; 2 = *Ameiurus melas*, *A. nebulosus*; 3 = *Hypophthalmichthys nobilis*, *H. molitrix*; 4 = *Perca fluviatilis*, *Sander lucioperca*; 5 = *Leuciscus idus*, *L. leuciscus*; 6 = *Ctenopharyngodon idella*, *H. molitrix*. "†" indicates this value could increase (see more detail in Appendix S2.5 Section S2.5.5). "§"indicates this primer pair is *in silico* tested with the reduced reference database including 38 species. The primer IDs and sequences are described in Table 2.2.

2.3.3 *In vitro* test of metabarcoding primers

The gradient PCR with single DNA PCR amplification indicated that the optimal annealing temperature is 50°C and 58°C for the Cytb_01 and 12S_03, respectively. The 22 freshwater fish species can be successfully amplified by the Cytb_01 and 12S_03 primer pairs under optimal annealing temperature (Figure 2.3).

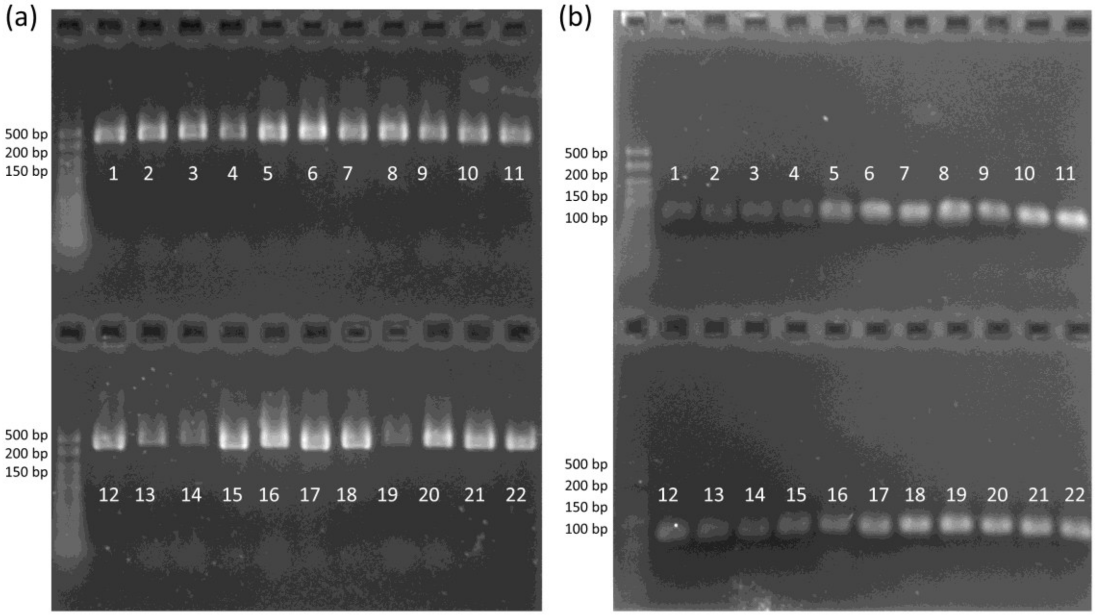


Figure 2.3 Results of *in vitro* tests (single species amplification) of the two chosen primer pairs (a) Cytb_01 (L14841 and H15149) and (b) 12S_03 (12S_V5_F and 12S_V5_R) under optimal annealing temperature. The optimal annealing temperature is 50°C and 58°C for the Cytb_01 and 12S_03, respectively. PCR products were run on 2.5% agarose gels, and stained with ethidium bromide. The detail information of these two primer pairs are given in Table 2.2. Numbers indicate different species and correspond to those in Table 2.1.

All 22 species were detected in the mock communities with the exception of nine-spined stickleback for the Cytb_01 (Table 2.4 & Table 2.5). The reason for nine-spined stickleback reduced detection probability with the Cytb_01 could be primer bias due to one insertion at the forward primer binding site of this species (see more detail in Appendix S2.5 Section S2.5.1). Two other species were represented by a very low number of sequence read counts: gudgeon *Gobio gobio* (32 read counts) for the Cytb_01 and pumpkinseed *Lepomis gibbosus* for the 12S_03 (125 read counts). There were false positives across all mock communities and two loci which could be from low potential contamination during the library construction process or sequencing error (barcode misassignment, Deakin et al. 2014; or “tag jumps” Schnell et al. 2015). In the Cytb_01 dataset, the read count proportion in the sample of most false positive records was less than 0.2% excluding common bream *Abramis brama* in mock community 5, brown bullhead *Ameiurus nebulosus* in mock community 9, and common carp *Cyprinus carpio* in mock communities 8 and 9 (Table 2.4). In the 12S_03 dataset, the read count proportion in the sample of most false positive records was less than 0.1% excluding common bleak *Alburnus alburnus* in mock communities 3 and 8, common carp in mock community 8, and ruffe *Gymnocephalus cernua* in mock community 6 (Table 2.5).

On the whole, there was significant correlation between observed read count proportions per species and expected read count proportions per species (based on DNA concentrations) of the combined 10 mock communities, for both loci (Figure 2.4a; Cytb_03 Pearson’s $r = 0.53$, $df = 20$, $p = 0.011$; 12S_01 Pearson’s $r = 0.43$, $df = 20$, $p = 0.044$). There also was significant correlation between the Cytb_03 and 12S_01 in terms of species read count proportions in the combined 10 mock communities (Figure 2.4b; Pearson’s $r = 0.60$, $df = 20$, $p = 0.003$).

These two primer pairs have also been tested *in situ* on Lake Windermere (England, UK) and detected 14 of the 16 previously recorded species using the 12S_03 and 12 recorded species using the Cytb_01 (Hänfling et al. 2016). The species that are not detected with either locus were the river lamprey and sea lamprey. These two lampreys could not be amplified by the Cytb_01 due to primer bias (see more detail in Box 1) with two insertions at the forward primer binding site of these species (see more detail in Appendix S2.5 Section S2.5.1). However, the reason for lampreys reduced detection with the 12S_01 could be specific lotic habitat requirements for these species instead of primer bias, since lamprey eDNA in Lake Windermere are detected in other sampling campaigns (Lawson Handley et al. 2019). In addition to lampreys, tench *Tinca tinca* and rudd *Scardinius erythropthalmus* are not detected with the Cytb_01 in Lake Windermere. Tench are detected with the Cytb_01 in five mock communities which contained this species (Table 2.4); therefore the possibility that the Cytb_01 is unsuitable for metabarcoding is ruled out. eDNA is often highly degraded into mostly small fragments over time (Deagle et al. 2006). Nevertheless, tench and rudd are found at very low site occupancy in the 12S_03 dataset, and hence this discrepancy between markers could be explained by the shorter fragment size of metabarcode amplified from the 12S_03 (106 bp) compared to the Cytb_01 (414 bp). Shorter fragments are likely to be present for longer in the environment after being shed from the source (e.g., Bista et al. 2017; Jo et al. 2017), which in turn, would make the 12S_03 more likely to pick up scarce species compared to the Cytb_01.

Table 2.4 Sequence read counts in individual Cytb mock communities with the primer pair Cytb_01 (L14841 and H15149) via metabarcoding.

Species	MC01	MC02	MC03	MC04	MC05	MC06	MC07	MC08	MC09	MC10
A. brama	8193	26	10769	15074	393	18219	14	4124	22141	36
A. alburnus	1867	765	30	7	2232	139	70	46	99	7615
A. nebulosus	12168	0	42	37	9492	1678	10	107	373	2850
B. barbus	0	2526	3533	0	0	3	5619	1486	0	0
C. albula	76	0	124	170	4	285	0	39	23	0
C. gobio	1864	0	5	0	4704	7118	0	4	63	779
C. carpio	7059	9349	80	180	16713	2874	4030	314	442	17813
E. lucius	2467	16	4642	9733	0	60	16	4784	6859	3
G. gobio	0	32	0	0	0	0	0	0	0	0
G. cernua	3	2851	3123	5646	0	0	5306	7071	510	0
L. gibbosus	4	8828	59	0	5196	0	15178	100	4	1133
L. delineatus	0	10	8216	0	0	0	0	7624	0	3
L. leuciscus	0	368	0	0	106	0	34	0	0	621
P. fluviatilis	10	2197	0	3521	0	0	4792	0	3782	0
P. phoxinus	3	3	2307	2295	0	0	0	138	3130	0
P. parva	0	1340	0	0	923	0	2158	7	11	185
P. pungitius	0	0	0	0	0	0	0	0	0	0
R. rutilus	0	149	0	155	0	0	23	0	13	0
S. trutta	6	10	2312	2383	0	0	7	4171	3460	0
S. cephalus	2046	0	0	2920	83	3898	0	0	400	7
T. tinca	273	0	345	296	0	55	0	33	37	4
U. pygmaea	1788	10	4827	13	26	860	13	12943	49	0
A. anguilla	0	3	0	0	0	0	0	0	0	0
Cyprinidae	216	0	276	352	0	300	0	63	371	0
Percidae	0	57	46	119	0	0	196	157	18	0
Salmonidae	0	0	3	6	0	0	0	0	0	0
Clupeocephala	0	0	0	0	135	0	0	9	0	0
Percinae	0	208	17	481	0	0	401	0	511	0
nohit	6921	4190	2638	5295	3285	5336	3519	3362	1615	3929
Total	44964	32938	43394	48683	43292	40825	41386	46582	43911	34978

🟩	20 ng DNA added to mock community
🟩	10 ng DNA added to mock community
🟩	1 ng DNA added to mock community
🟧	Read count proportion in the sample of false positive < 0.2%
🟫	Read count proportion in the sample of false positive > 0.2% but < 1.5%

Notes: Green palettes indicate species which were added to the community and the amount of DNA added (see legend above). Orange and brown colours indicate false positives and their frequencies (see legend above).

Table 2.5 Sequence read counts in individual 12S mock communities with the primer pair 12S_03 (12S_V5_F and 12S_V5_R) via metabarcoding

Species	MC01	MC02	MC03	MC04	MC05	MC06	MC07	MC08	MC09	MC10
A. brama	8985	3	4736	5238	0	19987	12	1224	13454	4
A. alburnus	3786	2748	119	8	5142	440	531	133	3	6379
A. nebulosus	7913	0	13	14	6582	1618	0	29	0	1655
B. barbus	0	3611	4935	0	3	0	7686	1062	0	0
C. albula	290	0	329	451	0	2475	0	63	38	3
C. gobio	3678	0	7	5	3022	10055	0	4	0	986
C. carpio	3284	4392	31	30	6250	692	848	101	33	8654
E. lucius	3541	9	3187	6142	0	409	14	4224	4588	0
G. gobio	0	2606	0	0	10	0	347	0	0	14
G. cernua	8	8172	6421	10379	0	59	9925	14600	1159	0
L. gibbosus	0	64	0	0	38	0	15	0	0	8
L. delineatus	0	4	11889	0	0	0	5	23635	0	0
L. leuciscus	5	7029	0	0	6297	3	1598	41	0	8380
P. fluviatilis	0	2847	0	3319	0	5	2504	0	4780	0
P. phoxinus	4	5	5279	6505	0	13	8	624	6746	3
P. parva	19	3447	15	33	3475	0	5731	29	34	342
P. pungitius	4	0	0	9	4128	4	0	11	3	882
R. rutilus	0	3124	3	2163	6	0	973	0	396	4
S. trutta	3	19	4219	5802	0	10	7	8005	6454	0
S. cephalus	5248	6	0	6354	7	9891	5	0	648	15
T. tinca	983	3	527	667	0	1094	3	125	77	0
U. pygmaea	1426	3	1793	6	0	461	3	2174	0	3
A. anguilla	0	0	7	0	0	0	0	0	0	0
B. bjoerkna	3	0	3	0	0	4	0	0	7	0
H. molitrix	8	0	0	0	0	6	0	0	0	0
L. idus	0	0	0	0	3	0	0	0	0	4
nohit	13948	23891	12393	10343	18476	9657	19320	7433	10847	11031
Total	53136	61983	55906	57468	53439	56883	49535	63517	49267	38367

■ 20 ng DNA added to mock community
🟩 10 ng DNA added to mock community
□ 1 ng DNA added to mock community
🟧 Read count proportion in the sample of false positive < 0.1%
🟫 Read count proportion in the sample of false positive > 0.1% but < 0.3%

Notes: Green palettes indicate species which were added to the community and the amount of DNA added (see legend above). Orange and brown colours indicate false positives and their frequencies (see legend above).

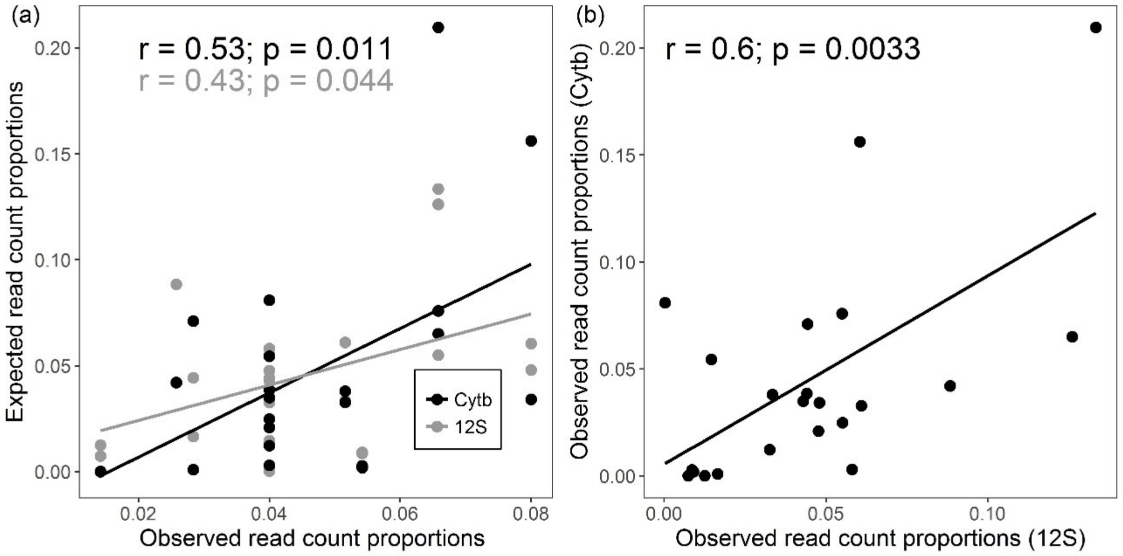


Figure 2.4 Correlations between (a) expected read count proportions per species (based on DNA concentrations) and observed read count proportions per species, and (b) observed read count proportions between Cytb_01 and 12S_03.

2.4 Conclusions

In this study, two primer pairs were selected for metabarcoding of UK freshwater fish communities, based on extensive *in silico* tests. These two metabarcoding primer pairs were also *in vitro* validated on 22 common freshwater fish species and 10 mock communities, which demonstrates their suitability for eDNA metabarcoding of UK lake fish communities. The data of mock communities were analysed with a custom reproducible metabarcoding bioinformatics analysis pipeline (metaBEAT). These results are supported by the *in situ* test results of Lake Windermere (Hänfling et al. 2016). Therefore, these two metabarcoding primer pairs, two custom-made reference databases and metaBEAT are used to investigate the other eDNA questions such as the optimal pore size of filters during filtration (Chapter 3), the eDNA production, degradation and transport (Chapter 4), and development of an eDNA method for monitoring fish communities (Chapter 5). The reproducible workflow of the compilation of curated reference databases presented here allows updating or adding new sequences into reference databases at any time and can also be applied to generate local reference databases for other fish communities and in fact other taxonomic groups and markers.

Supporting Information

Appendix S2.1 Supplementary text

S2.1.1 Custom reproducible metabarcoding pipeline

The bioinformatics analysis was carried out using the reproducible metabarcoding pipeline developed in Hänfling et al. (2016) with a number of minor improvements. The programme Trimmomatic v0.32 (Bolger et al. 2014) was used for quality trimming and removal of adapter sequences from the raw Illumina reads. Average read quality was assessed in sliding windows (window size 5 bp) starting from the 3'-end of the read, and reads were clipped until the average quality per window above a phred score of 30. All reads shorter than a defined minimum read length (100 bp for Cytb, 90 bp for 12S) were discarded. Sequence pairs were then merged into single high quality reads using the programme FLASH v1.2.11 (Magoč & Salzberg 2011). The remaining reads were screened for chimeric sequences (see more detail in Box 1) against the curated reference databases using the "uchime_ref" function implemented in VSEARCH v1.1.0 (Rognes et al. 2016). To remove redundancy, sequences were clustered at 100% identity using VSEARCH v1.1.0 (Rognes et al. 2016). Clusters represented by less than three sequences were considered sequencing error and omitted from further analyses. Non-redundant sets of query sequences were then compared to the respective curated non-redundant reference database using BLAST (Zhang et al. 2000). The BLAST output was analysed using a custom python script, which implements a lowest common ancestor (LCA) approach for taxonomic assignment similar to the strategy used by the programme MEGAN (Huson et al. 2007) (see more detail in Box 2). In brief, after the BLAST search the most significant matches were recorded to the reference database (yielding the top 10% bit-scores) for each of the query sequences. If only a single taxon was present in the top 10% the query was assigned directly to this taxon. If more than one reference taxon was present in the top 10%, the query was assigned to the lowest taxonomic level that was shared by all taxa in the list of most significant hits for this query.

S2.1.2 Removed mislabelled records in the Cytb and 12S reference databases

Removed mislabelled records in the Cytb reference database:

JN995186.1; KP794942.1; EU492281.1; EU224046.1; AJ969128.1; AJ937943.1; JF489783.1; KT275288.1; DQ185405.1; KT275289.1; EU224045.1; JQ346141.1; JQ661401.1; AJ937931.1; AJ937952.1; AJ937925.1; JQ661398.1; AJ937951.1; KP644340.1; DQ664351.1; GQ279764.1; GU182336.1; JQ661399.1; KP452507.1; JQ661400.1; JQ231114.1

Removed mislabelled records in the 12S reference database:

AH013021.2; AH013020.2; AF154850.1; EU048341.1; JN007557.1; JN007558.1; KM052222.1; KU821707.1; DQ447667.1; AJ002633.1; KC292943.1; EU075178.1; JQ231114.1; KJ135626.1; KJ746953.1; Y12671.1

Appendix S2.2 The curated Cytb reference database (.gb; supplied in a separate file)

The curated Cytb reference database can be downloaded use the link:

https://github.com/HullUni-bioinformatics/Curated_reference_databases/blob/master/Cytb_Fish/Cytb_Fish_SATIVA_cleaned_Dec_2017.gb

Appendix S2.3 The curated 12S reference database (.gb; supplied in a separate file)

The curated 12S reference database can be downloaded use the link:

https://github.com/HullUni-bioinformatics/Curated_reference_databases/blob/master/12S_Fish/12S_Fish_SATIVA_cleaned_May_2017.gb

Appendix S2.4 The reduced reference database (.gb; supplied in a separate file)

The reduced reference database which only including the mitogenomes can be downloaded use the link:

https://github.com/HullUni-bioinformatics/Curated_reference_databases/blob/master/fish_mitoDNA.gb

Appendix S2.5 *In silico* test results of metabarcoding primers

S2.5.1 Cytb_01 *in silico* test results

Forward primer:

L14912 AAAAACCACCGTTGTTATTCAACTA Kocher et al. (1989)

Reverse primer:

H15149 GCDCCTCARAATGAYATTTGTCCTCA Kocher et al. (1989)

Recommended annealing temperature (T_A): 50 °C

Min. length: 407 bp **Mean length:** 412 bp **Max. length:** 414 bp

Coverage in Species level (*Bc*): 84.21% (32/38)

Potentially not amplified species: *Esox lucius*§, *Gasterosteus aculeatus*§, *Lampetra fluviatilis*§, *Leuciscus idus*§, *Petromyzon marinus*§, *Pungitius pungitius*§

Specificity in Species level (*Bs*): 93.75% (30/32)

Unresolved species pairs: Coregonus (*Coregonus lavaretus*, *C. oxyrinchus*)

Notes: This primer pair is *in silico* tested with the reduced reference database including 38 species. § This species cannot be *in silico* amplified because there are one or two insertions at the forward primer binding site.

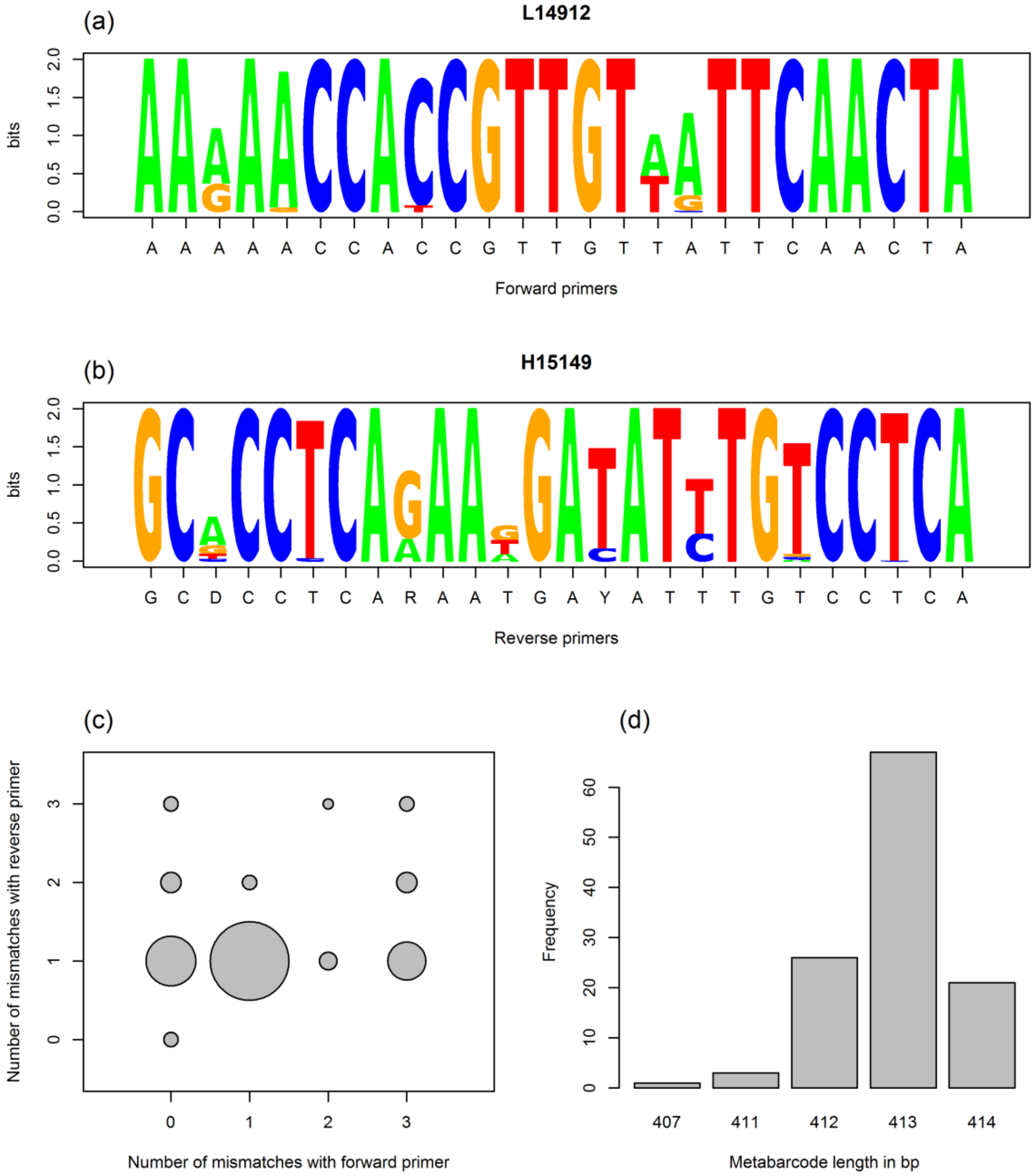


Figure S2.1 *In silico* test results of the Cytb_01 (L14912 and H15149) primer pair. Sequence logo for (a) the forward primer L14912 and (b) the reverse primer H15149; the taller the letter, the more conserved the corresponding nucleotide position. (c) Mismatch analysis of primers for the species in the curated database and (d) Length distribution of the metabarcode.

S2.5.2 Cytb_02 *in silico* test results

Forward primer:

Fish2CBL ACAACTTCACCCCTGCAAAC Thomsen et al. (2012a)

Reverse primer:

Fish2bCBR GATGGCGTAGGCAAACAAGA Thomsen et al. (2012a)

Recommended annealing temperature (T_A): 50 °C

Min. length: 40 bp **Mean length:** 40 bp **Max. length:** 40 bp

Coverage in Species level (*Bc*): 37.31% (25/67)

Specificity in Species level (*Bs*): 92.00% (23/25)

Unresolved species pairs: Ameiurus (*Ameiurus nebulosus* and *A. melas*)

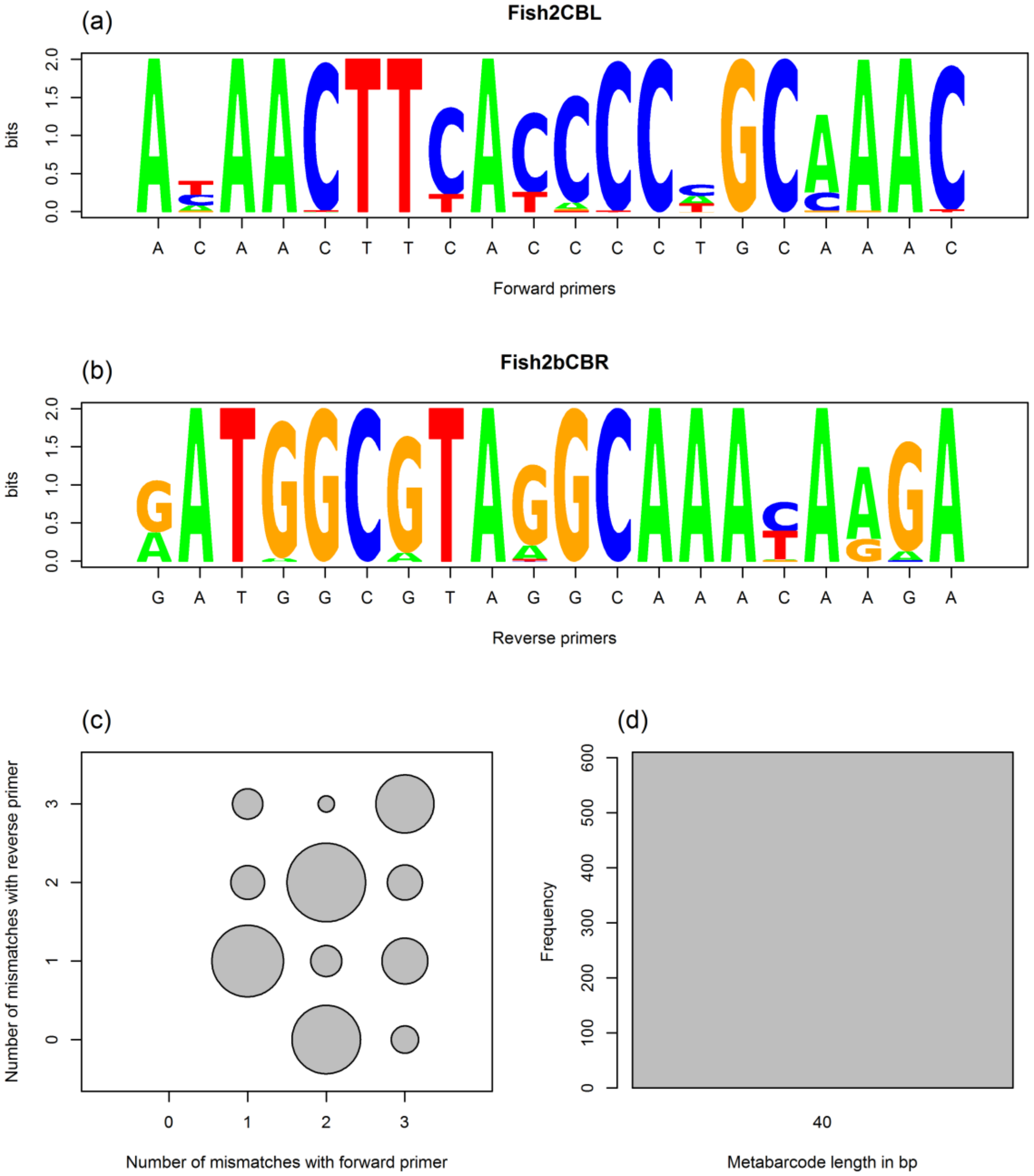


Figure S2.2 *In silico* test results of the Cytb_02 (Fish2CBL and Fish2bCBR) primer pair. Sequence logo for (a) the forward primer Fish2CBL and (b) the reverse primer Fish2bCBR; the taller the letter, the more conserved the corresponding nucleotide position. (c) Mismatch analysis of primers for the species in the curated database and (d) Length distribution of the metabarcode.

S2.5.3 Cytb_03 *in silico* test results

Forward primer:

Fish2degCBL	ACAACTTCACCCCTGCRAAY	Thomsen et al. (2012a)

Reverse primer:

Fish2CBR	GATGGCGTAGGCAAATAGGA	Thomsen et al. (2012a)

Recommended annealing temperature (T_A): 50 °C

Min. length: 40 bp **Mean length:** 40 bp **Max. length:** 40 bp

Coverage in Species level (*Bc*): 44.78% (30/67)

Specificity in Species level (*Bs*): 86.67% (26/30)

Unresolved species pairs: Ameiurus (*A. nebulosus* and *A. melas*), Hypophthalmichthys (*Hypophthalmichthys molitrix* and *H. nobilis*)

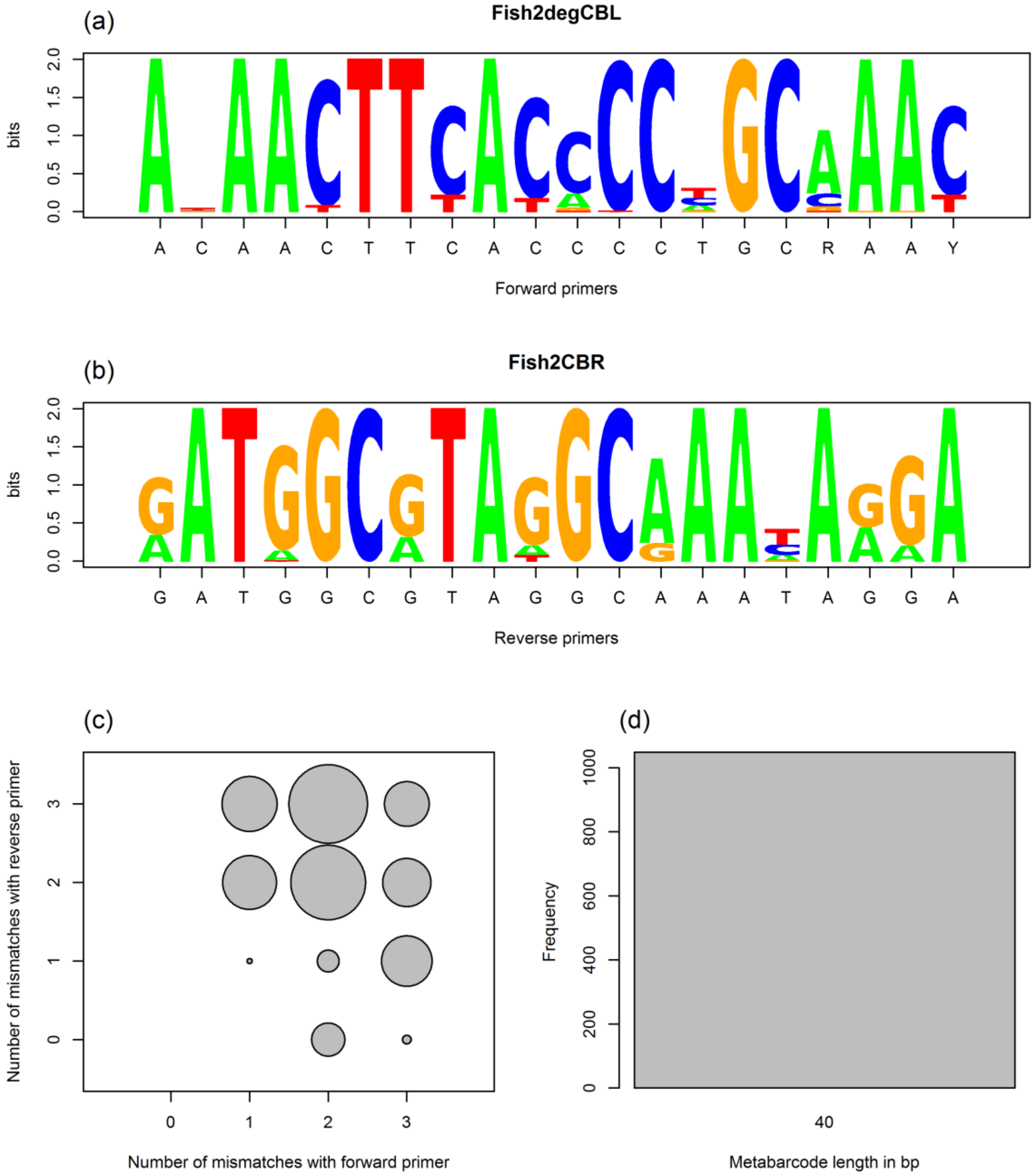


Figure S2.3 *In silico* test results of the Cytb_03 (Fish2degCBL and Fish2CBR) primer pair. Sequence logo for (a) the forward primer Fish2degCBL and (b) the reverse primer Fish2CBR; the taller the letter, the more conserved the corresponding nucleotide position. (c) Mismatch analysis of primers for the species in the curated database and (d) Length distribution of the metabarcode.

S2.5.4 12S_01 *in silico* test results

Forward primer:

Tele02_F	AAACTCGTGCCAGCCACC	Taberlet et al. (2018)

Reverse primer:

Tele02_R	GGGTATCTAATCCCAGTTTG	Taberlet et al. (2018)

Recommended annealing temperature (T_A): 54 °C

Min. length: 163 bp **Mean length:** 168 bp **Max. length:** 201 bp

Coverage in Species level (*Bc*): 95.38% (62/65)

Potentially not amplified species: *Alburnoides bipunctatus*§, *Alosa fallax*§, *Ambloplites rupestris*§

Specificity in Species level (*Bs*): 95.16% (59/62)

Unresolved species pairs: Coregonus (*Coregonus lavaretus*, *C. autumnalis*, *C. oxyrinchus*)

Notes: § This species cannot be *in silico* amplified because the sequences of this species in the curated 12S database are missing either any of primer binding site or both of them.

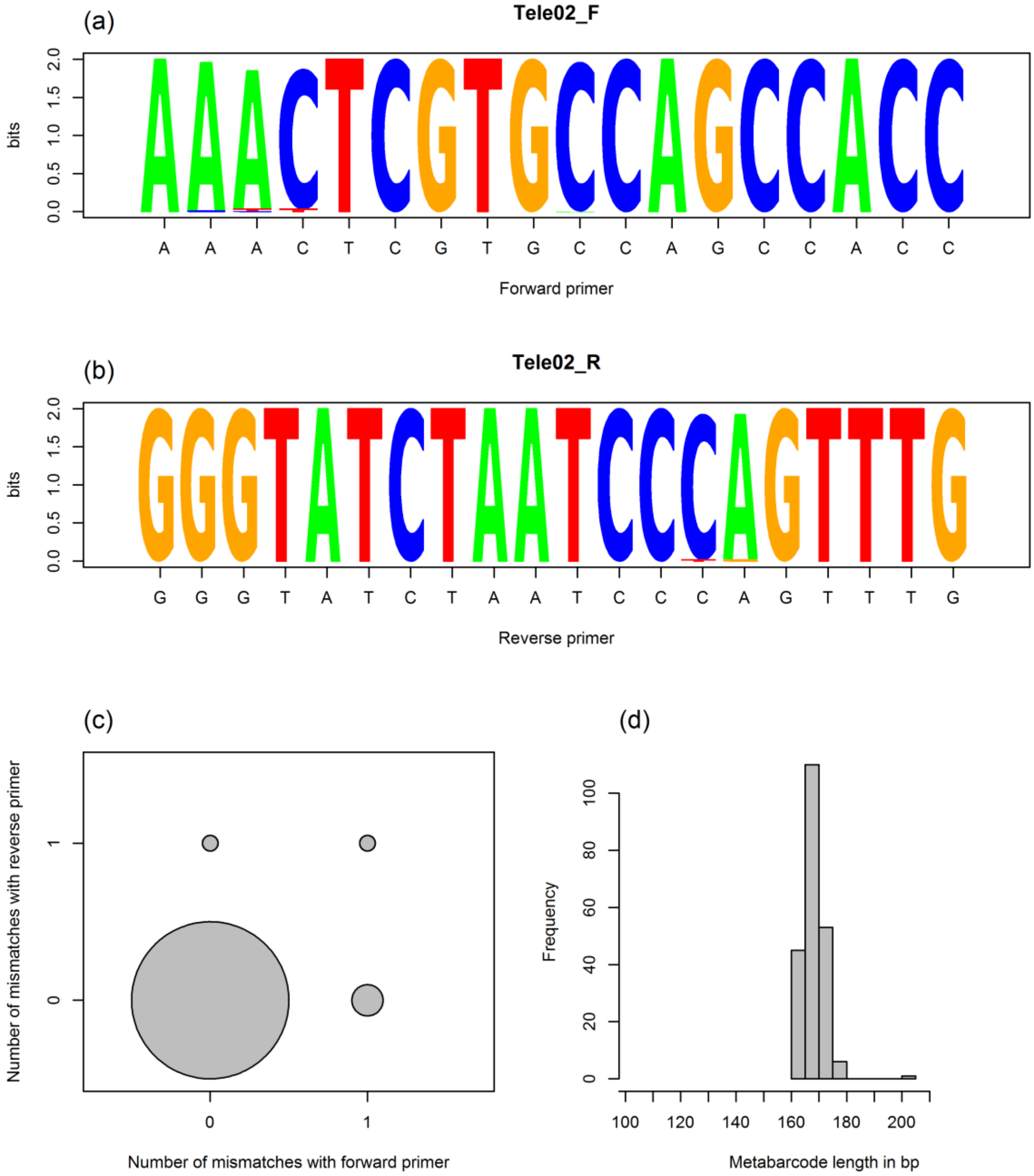


Figure S2.4 *In silico* test results of the 12S_01 (Tele02_F and Tele02_R) primer pair. Sequence logo for (a) the forward primer Tele02_F and (b) the reverse primer Tele02_R; the taller the letter, the more conserved the corresponding nucleotide position. (c) Mismatch analysis of primers for the species in the curated database and (d) Length distribution of the metabarcode.

S2.5.5 12S_02 *in silico* test results

Forward primer:

MiFish-U_Fa GCCGGTAAAACTCGTGCCAGC This study

Reverse primer:

MiFish-U_R CATAGTGGGGTATCTAATCCCAGTTTG Miya et al. (2015)

Recommended annealing temperature (T_A): 61 °C

Min. length: 168 bp **Mean length:** 172 bp **Max. length:** 205 bp

Coverage in Species level (*Bc*): 83.07% (54/65)

Potentially not amplified species: *Alburnoides bipunctatus*[§], *Alosa fallax*[§], *Ambloplites rupestris*[§], *Cobitis taenia*[§], *Misgurnus bipartitus*[§], *Misgurnus fossilis*[§], *Neogobius fluviatilis*[§], *Neogobius melanostomus*[‡], *Osmerus eperlanus*[§], *Platichthys flesus*[§], *Silurus glanis*[§]

Specificity (*Bs*) in Species level: 94.44% (51/54)

Unresolved species pairs: Coregonus (*C. lavaretus*, *C. autumnalis* and *C. oxyrinchus*)

Notes: § This species cannot be *in silico* amplified because the sequences of this species in the curated 12S database are missing part of the orward primer binding site. ‡ This species cannot be *in silico* amplified because there is one insertion at the forward primer binding site.

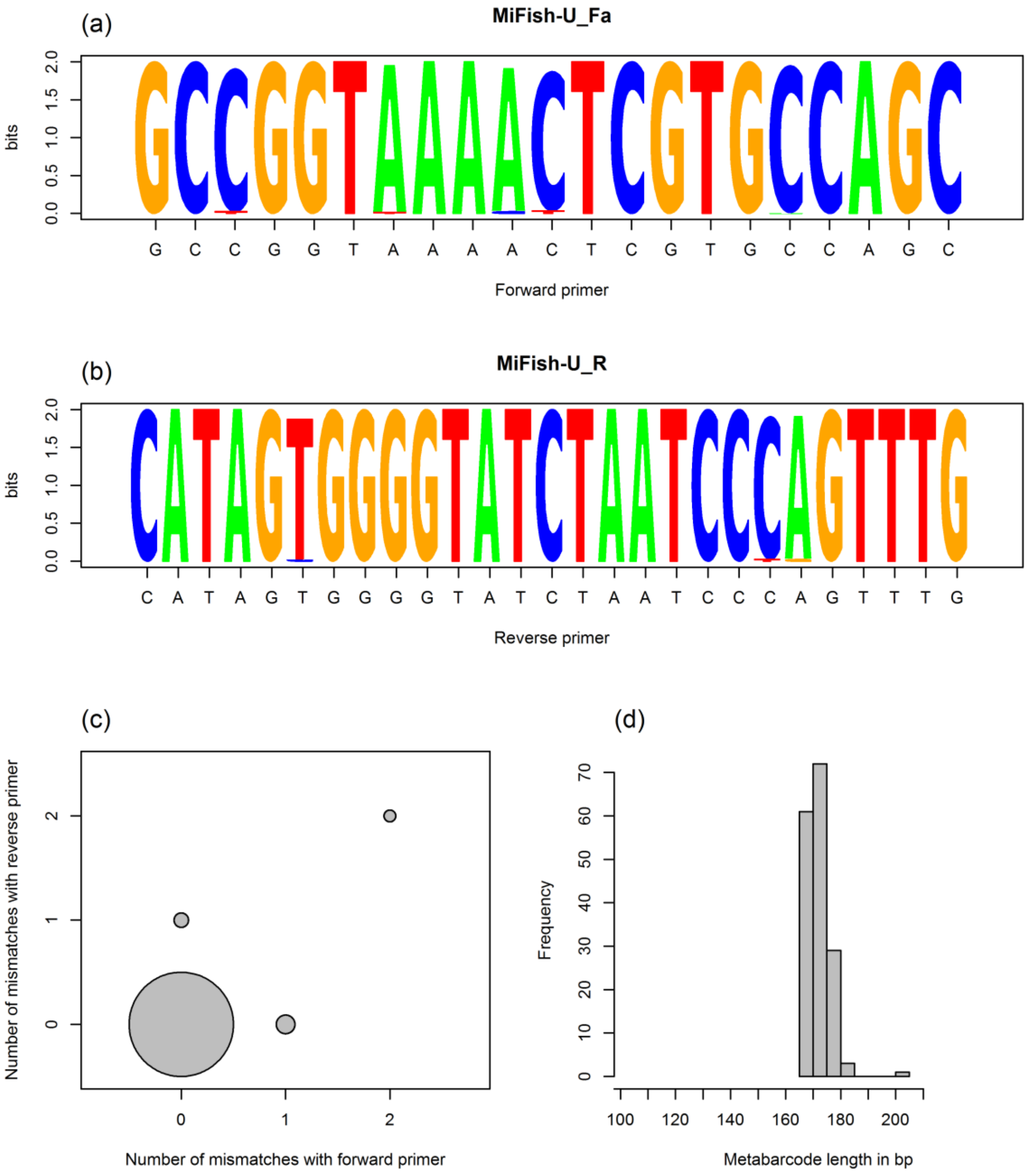


Figure S2.5 *In silico* test results of the 12S_02 (MiFish-U_Fa and MiFish-U_R) primer pair. Sequence logo for (a) the forward primer MiFish-U_Fa and (b) the reverse primer MiFish-U_R; the taller the letter, the more conserved the corresponding nucleotide position. (c) Mismatch analysis of primers for the species in the curated database and (d) Length distribution of the metabarcode.

S2.5.6 12S_03 *in silico* test results

Forward primer:

12S_V5_F ACTGGGATTAGATACCCC Riaz et al. (2011)

Reverse primer:

12S_V5_R TAGAACAGGCTCCTCTAG Riaz et al. (2011)

Recommended annealing temperature (T_A): 58 °C

Min. length: 89 bp **Mean length:** 106 bp **Max. length:** 108 bp

Coverage in Species level (*Bc*): 95.38% (62/65)

Potentially not amplified species: *Alburnoides bipunctatus*§, *Alosa fallax*§, *Ambloplites rupestris*§

Specificity in Species level (*Bs*): 92.18% (57/62)

Unresolved species pairs: Coregonus (*Coregonus lavaretus*, *C. autumnalis* and *C. albula*); Percidae (*Sander lucioperca* and *Perca fluviatilis*)

Notes: § This species cannot be *in silico* amplified because the sequences of this species in the curated 12S database are missing either any of primer binding site or both of them.

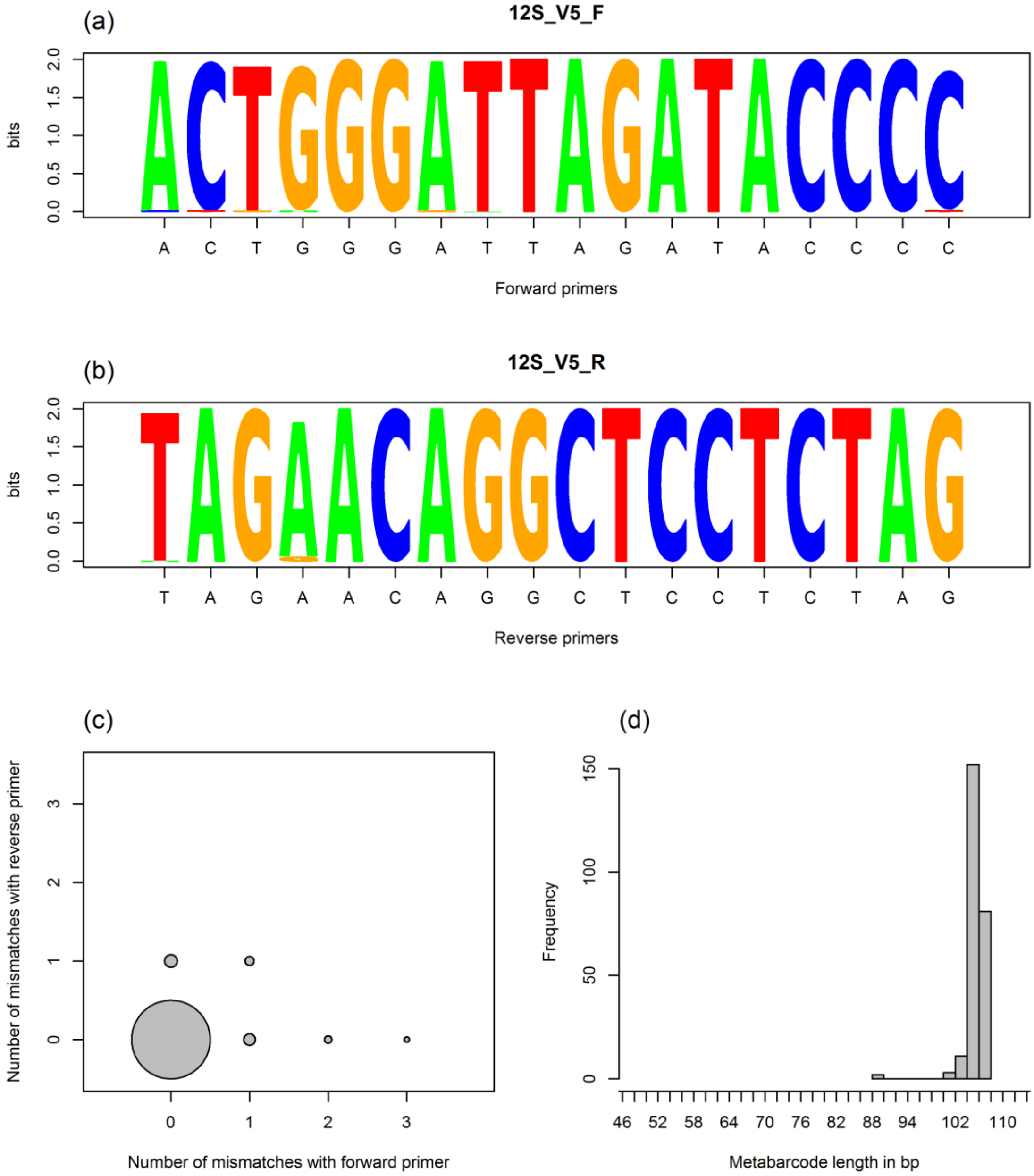


Figure S2.6 *In silico* test results of the 12S_03 (12S_V5_F and 12S_V5_R) primer pair. Sequence logo for (a) the forward primer 12S_V5_F and (b) the reverse primer 12S_V5_R; the taller the letter, the more conserved the corresponding nucleotide position. (c) Mismatch analysis of primers for the species in the curated database and (d) Length distribution of the metabarcode.

S2.5.7 12S_04 *in silico* test results

Forward primer:

12S_V5_F	ACTGGGATTAGATACCCC	Riaz et al. (2011)

Reverse primer:

12S_V5_R2	CTACACCTCGACCTGACG	this study

Recommended annealing temperature (T_A): 55 °C

Min. length: 223 bp **Mean length:** 238 bp **Max. length:** 241 bp

Coverage in Species level (*Bc*): 95.38% (62/65)

Potentially not amplified species: *Alburnoides bipunctatus*§, *Alosa fallax*§, *Ambloplites rupestris*§

Specificity in Species level (*Bs*): 95.16% (59/62)

Unresolved species pairs: Coregonus (*Coregonus lavaretus*, *C. autumnalis* and *C. albula*)

Notes: § This species cannot be *in silico* amplified because the sequences of this species in the curated 12S database are missing either any of primer binding site or both of them.

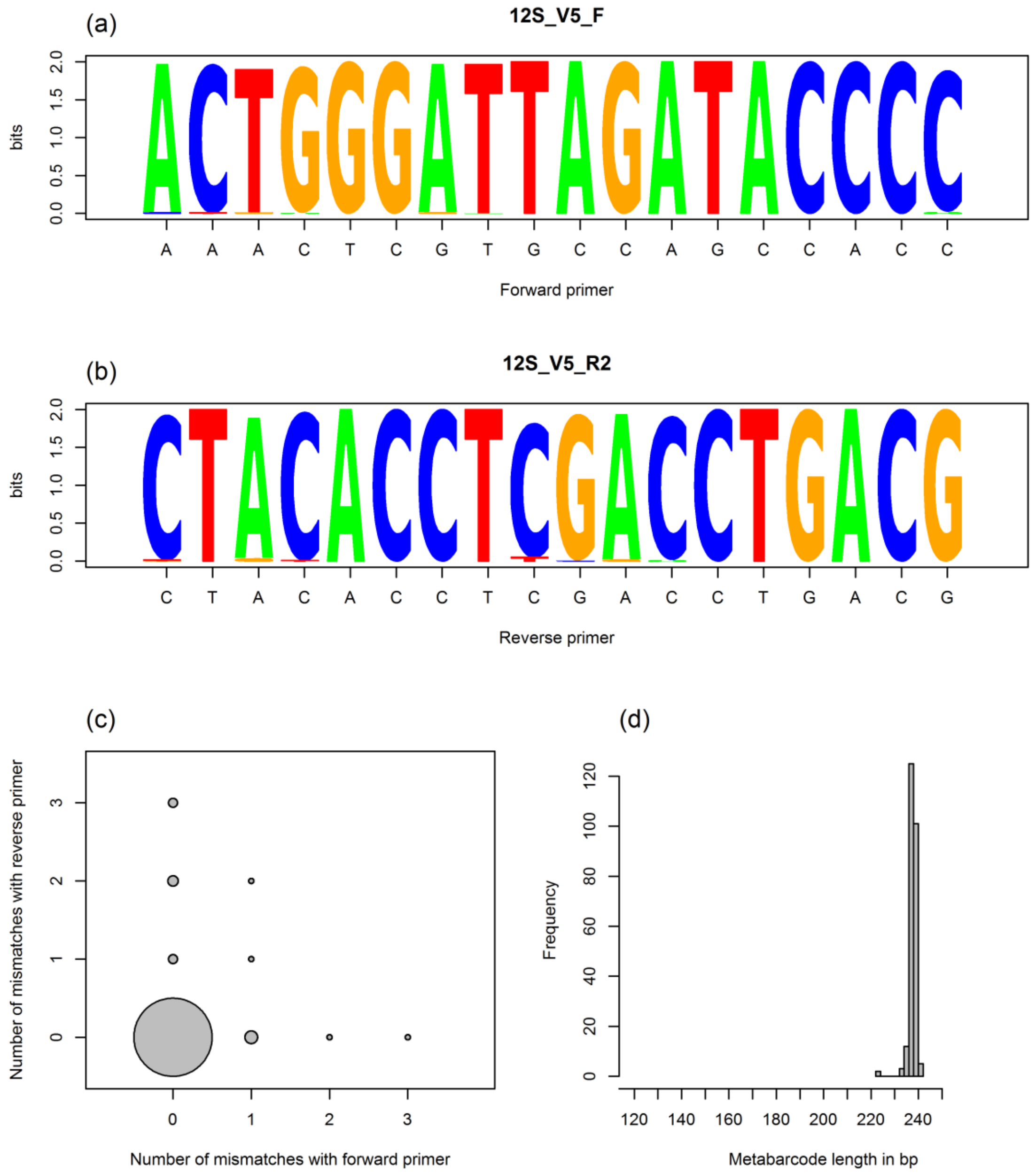


Figure S2.7 *In silico* test results of the 12S_04 (12S_V5_F and 12S_V5_R2) primer pair. Sequence logo for (a) the forward primer 12S_V5_F and (b) the reverse primer 12S_V5_R2; the taller the letter, the more conserved the corresponding nucleotide position. (c) Mismatch analysis of primers for the species in the curated database and (d) Length distribution of the metabarcode.

S2.5.8 12S_05 *in silico* test results

Forward primer:

12S_V5_F	ACTGGGATTAGATACCCC	Riaz et al. (2011)

Reverse primer:

Ac12s_R	GAGAGTGACGGGCGGTGT	Evans et al. (2016)

Recommended annealing temperature (T_A): 54 °C

Min. length: 375 bp **Mean length:** 396 bp **Max. length:** 398 bp

Coverage in Species level (*Bc*): 95.38% (62/65)

Potentially not amplified species: *Alburnoides bipunctatus*§, *Alosa fallax*§, *Ambloplites rupestris*§

Specificity in Species level (*Bs*): 96.77% (60/62)

Unresolved species pairs: Coregonus (*Coregonus lavaretus* and *C. albula*)

Notes: § This species cannot be *in silico* amplified because the sequences of this species in the curated 12S database are missing either any of primer binding site or both of them.

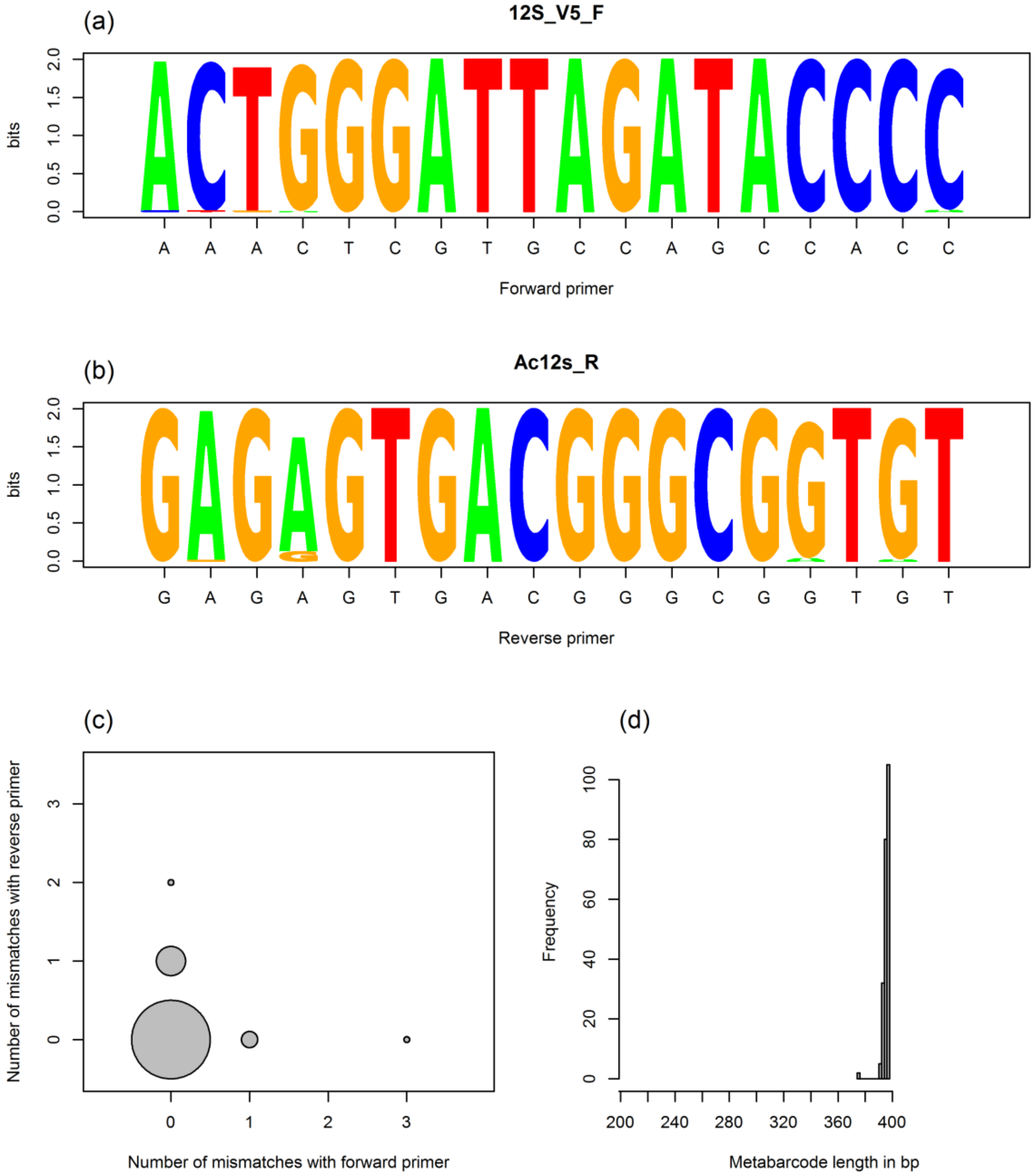


Figure S2.8 *In silico* test results of the 12S_05 (12S_V5_F and Ac12s_R) primer pair. Sequence logo for (a) the forward primer 12S_V5_F and (b) the reverse primer Ac12s_R; the taller the letter, the more conserved the corresponding nucleotide position. (c) Mismatch analysis of primers for the species in the curated database and (d) Length distribution of the metabarcode.

S2.5.9 12S_06 *in silico* test results

Forward primer:

Teleo_F	ACACCGCCCGTCACTCT	Valentini et al. (2016)

Reverse primer:

Teleo_R	CTTCCGGTACACTTACCATG	Valentini et al. (2016)

Recommended annealing temperature (T_A): 55 °C

Min. length: 60 bp **Mean length:** 63 bp **Max. length:** 67 bp

Coverage in Species level (*Bc*): 93.84% (61/65)

Potentially not amplified species: *Alburnoides bipunctatus*§, *Ambloplites rupestris*§, *Lampetra fluviatilis*, *Petromyzon marinus*

Specificity in Species level (*Bs*): 85.24% (52/61)

Unresolved species pairs: Ameiurus (*Ameiurus nebulosus* and *A. melas*), Coregonus (*Coregonus lavaretus*, *C. autumnalis* and *C. oxyrinchus*), Cyprinidae (*H. molitrix* and *Ctenopharyngodon idella*), Leuciscus (*Leuciscus idus* and *L. leuciscus*)

Notes: § This species cannot be *in silico* amplified because the sequences of this species in the curated 12S database are missing either any of primer binding site or both of them.

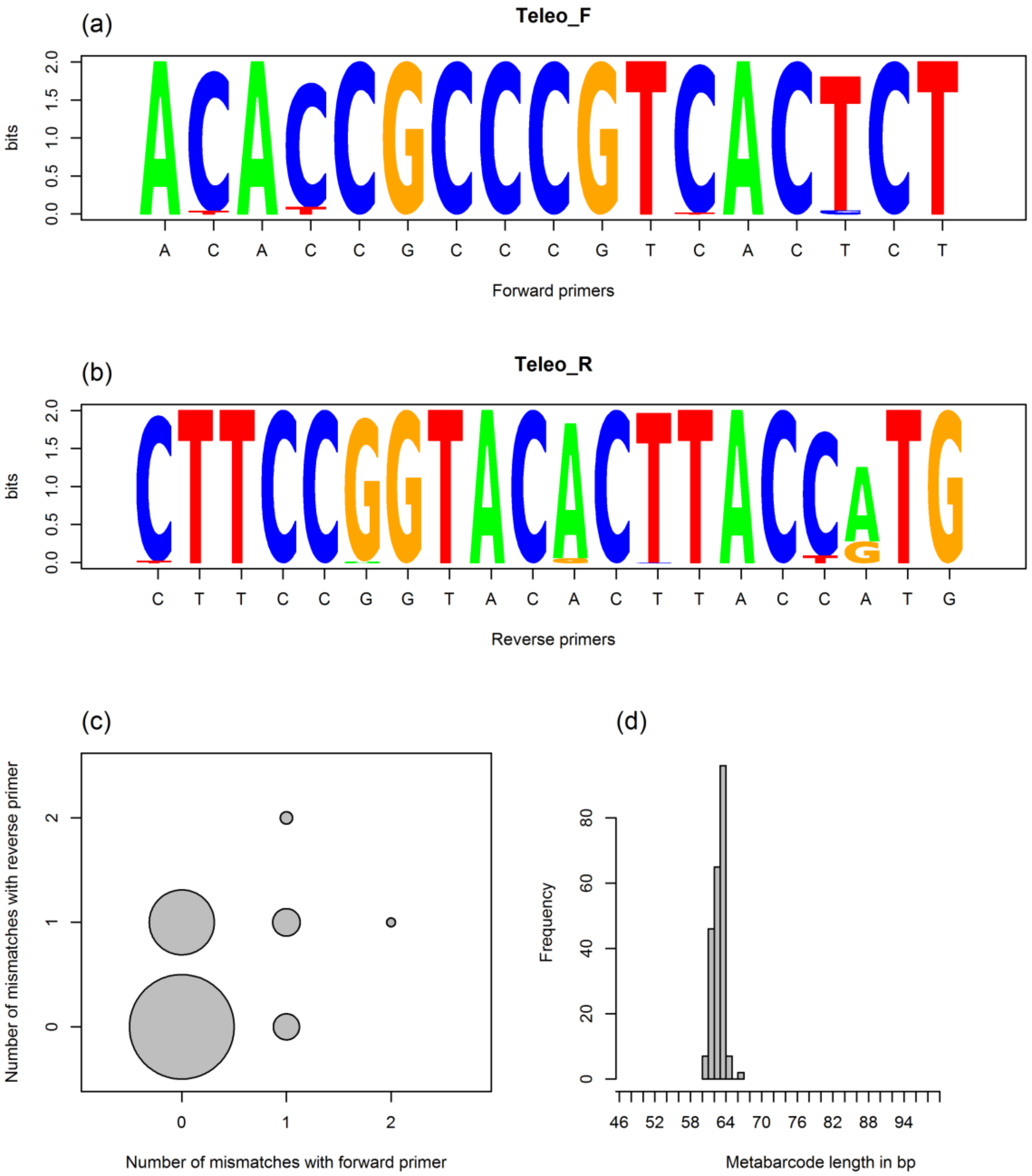


Figure S2.9 *In silico* test results of the 12S_06 (Teleo_F and Teleo_R) primer pair. Sequence logo for (a) the forward primer Teleo_F and (b) the reverse primer Teleo_R; the taller the letter, the more conserved the corresponding nucleotide position. (c) Mismatch analysis of primers for the species in the curated database and (d) Length distribution of the metabarcode.

Chapter 3 The effect of filtration method on the efficiency of environmental DNA capture and quantification via metabarcoding[2]

Abstract

Environmental DNA (eDNA) density is low in environmental samples, and a capture method, such as filtration, is often required to concentrate eDNA for downstream analyses. In this study, six treatments, with differing filter types and pore sizes for eDNA capture, were compared for their efficiency and accuracy to assess fish community structure with known fish abundance and biomass via eDNA metabarcoding. Our results show that different filters (except for 20 μm large-pore filters) are broadly consistent in their DNA capture ability. The 0.45 μm filters perform the best in terms of total DNA yield, the probability of species detection, repeatability within pond and consistency between ponds. However, the performance of 0.45 μm filters is only marginally better than for 0.8 μm filters, while filtration time is significantly longer. Given this trade-off, the 0.8 μm filter is the optimal pore size of the membrane filter for turbid, eutrophic and high fish density ponds analysed here. The 0.45 μm Sterivex enclosed filters perform reasonably well and are suitable in situations where on-site filtration is required. Finally, pre-filters are applied only if absolutely essential for reducing the filtration time or increasing the throughput volume of the capture filters. In summary, this study found encouraging similarity in the results obtained from different

[2] This chapter has been published as ***Li, J.***, Lawson Handley, L.J., Read, D.S. & Hänfling, B. (2018) The effect of filtration method on the efficiency of environmental DNA capture and quantification via metabarcoding. *Molecular Ecology Resources*, 18,1102-1114. https://doi.org/10.1111/1755-0998.12899

filtration methods, but the optimal pore size of filter or filter type might strongly depend on the water type under study.

3.1 Introduction

The analysis of environmental DNA (eDNA) is a non-invasive genetic method to detect the presence of organisms, including cryptic taxa, that takes advantage of intracellular or extra-organismal DNA in the environment (Lawson Handley 2015; Thomsen & Willerslev 2015; Goldberg et al. 2016). Generally, eDNA density is low in environmental samples, and a capture method is therefore required to concentrate eDNA for downstream analyses. The two main approaches to capture eDNA in aquatic environments are precipitation and filtration.

Capturing eDNA through precipitation entails adding ethanol or isopropanol with sodium acetate to water samples (Dejean et al. 2011; Foote et al. 2012; Doi et al. 2017b). Samples can be preserved quickly and easily in the field using such an approach, but it is only feasible for small volumes of water (<30 mL), which could reduce the probability of detection, particularly of rare species (Deiner et al. 2015; Eichmiller et al. 2016). Therefore, most recent studies have used filtration-based methods, which can process larger volumes of typically 250 mL to 5 L, or even up to 45 L (Civade et al. 2016). Previous studies have used a wide range of filter types (e.g., different membrane materials and pore sizes) and approaches (e.g., on-site or in the laboratory) to filtration. On-site filtration followed by immediate preservation theoretically enhances DNA integrity and is critical for some remote field surveys where access to laboratory facilities is not available. Enclosed filters such Sterivex units (Millipore) or Nalgene analytical test filter funnels (Thermo Fisher Scientific), in combination with a portable peristaltic or hand-driven pump are popular protocols for the capture of eDNA in the field (Keskin 2014; Bergman et al. 2016; Wilcox et al. 2016; Spens et al. 2017). However, a larger number of water samples can be filtered simultaneously in a laboratory setting, which reduces the processing time. Four main types of membrane filter (so-called "open filters") are commonly used in the laboratory set-ups of freshwater studies: (1) 0.45 μm cellulose nitrate (CN) filters (e.g., Goldberg et al. 2011; Pilliod et al. 2013), (2) 0.45 μm nylon filters (e.g., Thomsen et al. 2012a), (3) 0.7 or 1.5 μm glass fibre (GF) filters (e.g., Wilcox et al. 2013; Miya et al. 2015), and (4) 1.2 μm polycarbonate (PC) filters (e.g., Egan et al. 2015).

The suitability of various pore sizes of the filter to capture eDNA may be heavily influenced by the heterogeneous nature of aquatic ecosystems. Suspended particulate matter (SPM, e.g., organic matter and sediment) can quickly block 0.2 or 0.45 μm filters (Minamoto et al. 2016; Shaw et al. 2016b), which will severely prolong filtration time and potentially increase concentration of PCR inhibitors (Tsai & Olson 1992; McKee et al. 2015). For highly turbid water such as ponds or tropical freshwater ecosystems, even 3 μm PC filters are easily blocked (Minamoto et al. 2016; Robson et al. 2016). Most previous studies that have investigated the impact of different types and pore sizes of filter on DNA quantity, have focussed on individual target species using real-time quantitative PCR (qPCR) (e.g., Eichmiller et al. 2016; Lacoursière-Roussel et al. 2016b; Minamoto et al. 2016; Robson et al. 2016).

Recently, eDNA-based metabarcoding using High-Throughput Sequencing (HTS) has emerged as a powerful tool to monitor entire aquatic communities (e.g., Deiner et al. 2016; Hänfling et al. 2016; Port et al. 2016; Valentini et al. 2016). To my knowledge, few previous studies have investigated if and how the choice of filtration method impacts on estimates of fish community composition. The preliminary results of Miya et al. (2016) showed that the number of detected fish species is significantly higher when using enclosed 0.45 µm polyvinylidene difluoride (PVDF) filters compared to 0.7 µm GF filters, although different filtration systems and extraction methods were used in each case. Djurhuus et al. (2017) found that different filter membrane materials (0.2 µm PC, CN, polyethersulfone "PES", and PVDF) and extraction methods do not affect estimates of species richness and community composition across multiple trophic levels. Majaneva et al. (2018) indicated that 0.45 µm mixed cellulose ester (MCE) filters (described as CN filters in the study) represent the community composition of metazoan more consistently than 0.2 µm PES filters, while the effect of using 12 µm filters as pre-filters remains ambiguous.

The aim of the present study is to further investigate the impact of different filters on eDNA capture and community diversity estimation through eDNA metabarcoding. Specifically, this study compare different pore sizes of the membrane filter, different types of filter ("open filters" and "enclosed filters"), and the impact of pre-filtration. I evaluate the effect on filtration time, total eDNA recovered, the probability of species detection, repeatability, and the relationship between read counts and known fish abundance or biomass in four fish ponds with differing assemblages.

3.2 Materials and methods

3.2.1 Study site and water sampling

This study was carried out at four artificially stocked ponds (E1–E4) with turbid and eutrophic, and high fish density condition at the National Coarse Fish Rearing Unit (Nottingham, UK), run by the UK Environment Agency. The size of each pond is 5100 m^2 (60 m × 85 m), and the depth is 1–1.5 m. Generally, these ponds are used to rear approximately one-year-old common British coarse fish from June to January before they are used in stocking programmes for conservation purposes or recreational fishing.

All fish were measured and weighed before stocking in the ponds on 15th June 2015 and after harvesting on 18th January 2016. Fish abundance and biomass at the time of water sampling in August 2015 were estimated, assuming that death and growth curves of these fish are linear (Appendix S3.1 Figure S3.1 & Figure S3.2). The fish stock information in August 2015 is shown in Table 3.1.

Water sampling was carried out on 6th August 2015. The dissolved oxygen (DO) concentration was similar between ponds (7.9 ± 0.8 mg L^{-1}). For each pond, 12 water samples were collected at evenly distributed points around the shore. A 1 L sterile bottle was used to collect water at each point just below the surface, and then the water was pooled into a 12.5 L sterile water container. After inverting and shaking the collection container, the water was then subsampled with 25 Gosselin 500 mL sterile plastic bottles. All samples were stored in cool boxes, transferred to the eDNA laboratory at the University of Hull (UoH) within 2 hrs and refrigerated until filtration.

Table 3.1 Fish stock information on four experiment ponds at the National Coarse Fish Rearing Unit.

Pond	**Species**			**August 2015**	
	Scientific name	Common name	Code	Abundance	Biomass (kg)
E1	*Rutilus rutilus*	Roach	ROA	33515	199.7
E1	*Barbus barbus*	Barbel	BAR	9695	118.8
E1	*Squalius cephalus*	Chub	CHU	14943	445.2
E1	*Abramis brama*	Bream	BRE	500	7.1
E1	*Tinca tinca*	Tench	TEN	944	10.9
E1	*Carassius carassius*	Crucian carp	CAR	489	10.2
E2	*Rutilus rutilus*	Roach	ROA	4730	52.4
E2	*Leuciscus leuciscus*	Dace	DAC	34729	287.0
E2	*Barbus barbus*	Barbel	BAR	9691	295.6
E2	*Abramis brama*	Bream	BRE	487	4.7
E2	*Carassius carassius*	Crucian carp	CAR	4910	86.8
E3	*Squalius cephalus*	Chub	CHU	18967	542.6
E3	*Rutilus rutilus*	Roach	ROA	30156	321.2
E3	*Carassius carassius*	Crucian carp	CAR	3474	58.6
E3	*Tinca tinca*	Tench	TEN	4773	58.2
E4	*Leuciscus leuciscus*	Dace	DAC	29322	248.0
E4	*Barbus barbus*	Barbel	BAR	9508	268.7
E4	*Scardinius erythrophthalmus*	Rudd	RUD	8334	71.1
E4	*Abramis brama*	Bream	BRE	4962	52.6
E4	*Carassius carassius*	Crucian carp	CAR	199	17.6
E4	*Tinca tinca*	Tench	TEN	4763	43.5

Notes: Abundance represents number of individuals. Full scientific, common names and three letter codes used in figures are given.

3.2.2 eDNA capture treatments

Six filtration-based eDNA capture treatments were used for each pond. These treatments were: (1) “0.45MCE”: 0.45 µm mixed cellulose acetate and nitrate (also known as MCE) filters, 47 mm diameter (Whatman); (2) “0.8MCE”: 0.8 µm MCE filters, 47 mm diameter (Whatman); (3) “1.2MCE”: 1.2 µm MCE filters, 50 mm diameter (Whatman); (4) “0.45Sterivex”: 0.45 µm Sterivex-HV PVDF units (Millipore); (5) “PF_0.45MCE”: 0.45 µm MCE filters, 47 mm diameter (Whatman) after pre-filtration with 20-µm qualitative cellulose filters, Grade 4 (Whatman); and (6) “PF”: the pre-filters used in the “PF_0.45MCE” treatment. Each treatment was replicated five times, filtering 300 mL water each time, resulting in a total of 120 replicates. These treatments were used to measure three different effects: pore sizes (0.45MCE, 0.8MCE and 1.2MCE), filter types (0.45MCE and 0.45Sterivex) and pre-filtration (0.45MCE and PF_0.45MCE) (Figure 3.1).

To reduce cross-contamination, the samples from individual ponds were filtered separately in order of pond E1 to E4. For each replicate (apart from the “0.45Sterivex” treatment), 300 mL water was filtered using Nalgene filtration units (Thermo Fisher Scientific) in combination with a vacuum pump (15–20 in. Hg; Pall Corporation). For each pond, the same filtration unit was used for all five replicates of the same capture treatment. The filtration units were cleaned with 10% v/v commercial bleach solution and 5% v/v microsol detergent (Anachem, UK), and then rinsed thoroughly with deionised water after each filtration to prevent cross-contamination. Filtration blanks (*N* = 5) with 300 mL deionised water were run before the first filtration and after every wash run in order to test for possible contamination at the filtration stage. For the “0.45Sterivex” treatment, 300 mL water was directly filtered with 0.45 µm Sterivex units in combination with a vacuum pump (15–20 in. Hg; Pall Corporation). All samples were filtered within 24 hrs of the collection in a dedicated eDNA filtration laboratory at UoH.

After filtration, all membrane filters were placed into 50 mm sterile petri dishes sealed with parafilm, while Sterivex units were closed with inlet and outlet caps. All samples were stored in a freezer at –20^{o}C until DNA extraction. DNA extraction was carried out using the PowerWater (Sterivex) DNA Isolation Kits (MoBio Laboratories Inc., now Qiagen) following the manufacturer's protocol. Total DNA concentration was quantified using a NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific) after extraction.

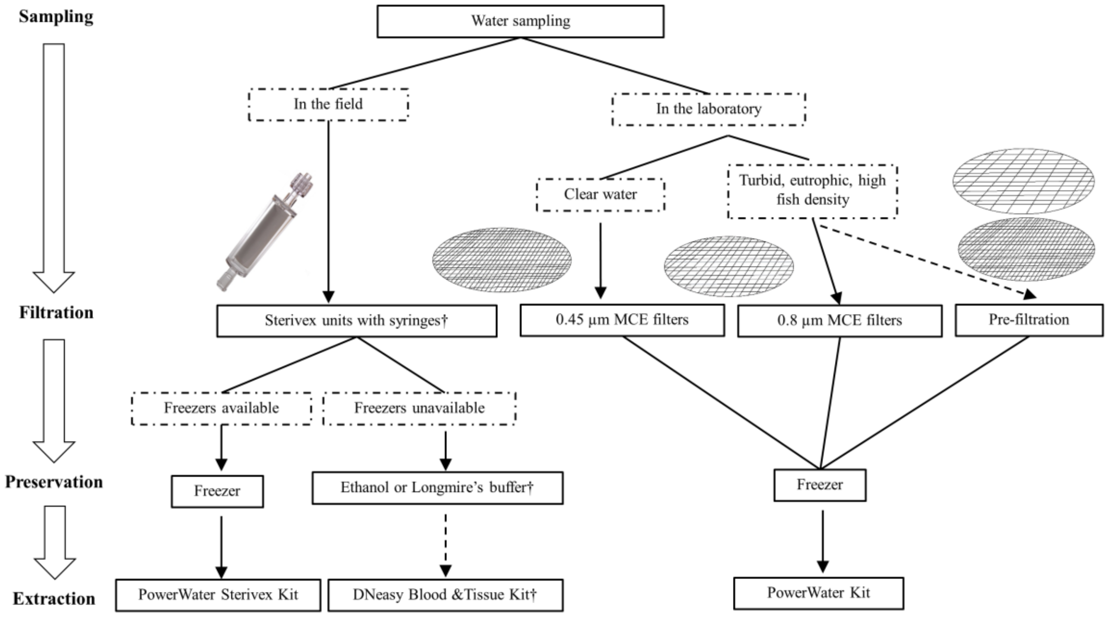


Figure 3.1 Flow chart illustrating the selection of eDNA capture, preservation and extraction methods based on the filtration equipment and aquatic ecosystems of study. "MCE": mixed cellulose acetate and nitrate. *Notes*: Pre-filters are applied only if it substantially reducing the filtration time or increasing the throughput volume of the capture filters. "†" refers to this method was recommended by Spens et al. (2017).

3.2.3 Library preparation and sequencing

Extracted DNA samples were PCR-amplified targeting a 106-bp vertebrate-specific fragment of the mitochondrial 12S rRNA region (Riaz et al. 2011) following a one-step library preparation protocol (Kozich et al. 2013) with amplification primers that include PCR primers, indices and flow cell adapters. Previous studies showed that this fragment has a low false negative rate in both marine mesocosm and coastal ecosystem eDNA metabarcoding studies of bony fishes (Kelly et al. 2014; Port et al. 2016). We also previously tested this fragment *in vitro* on 22 common freshwater fish species and 10 mock communities (see more detail in Chapter 2) and *in situ* on three deep lakes in the English Lake District, and demonstrated their suitability for eDNA metabarcoding of UK lake fish communities (Hänfling et al. 2016).

All PCRs were set up in a PCR workstation in our dedicated eDNA laboratory to minimise the risk of contamination. All samples (*N* = 120) together with five filtration and extraction controls, five no-template PCR controls and five positive PCR controls (the Eastern Happy, *Astatotilapia calliptera,* a cichlid from Lake Malawi, which is not present in the UK) were included in the Illumina MiSeq library construction and sequencing (*N* = 135). The library preparation protocol for this study was described in Chapter 2 Section 2.2.3.1. The final library concentration was quantified by QUBIT v3.0 using the dsDNA HS Assay Kit (Thermo Fisher Scientific) and qPCR using the NEBNext® Library Quant Kit (New England Biolabs). The library was adjusted to 2 nM and denatured following the Illumina MiSeq library denaturation and dilution guide. Because of the low fish diversity in the ponds, the final 10 pM denatured library was mixed with 30% PhiX control to improve the diversity of the library. The library was sequenced on an Illumina MiSeq platform using the MiSeq reagent kit v2 (2 × 250 cycles) at UoH. The custom sequencing and index primers were added to the appropriate wells of the MiSeq reagent cartridge as described by Kozich et al. (2013).

3.2.4 Data analysis

3.2.4.1 Bioinformatics analysis

Raw read data from Illumina MiSeq sequencing have been submitted to NCBI (BioProject: PRJNA414952; BioSample accession: SAMN07811461–SAMN07811580; Sequence Read Archive accessions: SRR6189420–SRR6189539). Bioinformatics analysis was implemented following a custom reproducible metabarcoding pipeline (metaBEAT v0.97.8) (see more detail in Appendix S2.1 Section S2.1.1; Hänfling et al. 2016) with a custom-made 12S rRNA reference database as described in Chapter 2. The maximum likelihood phylogenetic tree of the all 12S rRNA sequences from the custom reference database is shown in Appendix S3.1 Figure S3.3. Sequences for which the best BLAST hit had a bit score below 80 or had less than 100% identity to any sequence in the curated database were considered non-target sequences. To assure full reproducibility of our bioinformatics analysis, the custom reference database and the Jupyter notebook for data processing have been deposited in an additional dedicated GitHub repository (https://github.com/HullUni-bioinformatics/Li_et_al_2018_eDNA_filtration).

3.2.4.2 Criteria for reducing false positives and quality control

Filtered data were summarised into the number of sequence reads per species (hereon referred to as read counts) for downstream analyses (Appendix S3.2). Two criteria were applied to reduce the possibility of false positives. (1) The low-frequency noise threshold (proportion of positive species read counts of all read counts in the real sample) was set to filter some high-quality annotated reads passing the previous filtering steps that have high-confidence BLAST matches but may be inaccurate due to potential low-level contamination during the library construction process (De Barba et al. 2014; Hänfling et al. 2016; Port et al. 2016). The low-frequency noise threshold was set to 0.001 in this study; therefore, all taxonomic assignments with the frequency below this threshold were omitted from further downstream analysis. (2) After the low-frequency noise threshold was applied, remaining taxonomic assignments of taxa that were not stocked in the ponds (i.e., brown trout *Salmo trutta*, bleak *Alburnus alburnus*, and Gudgeon *Gobio gobio*) were also treated as false positives and excluded.

Samples were excluded from the analysis because they performed poorly in terms of PCR and sequencing depth due to low DNA concentrations. Two samples (T3-1-3 and T2-2-3) show extremely low levels of DNA concentration and failed PCR. One sample (T4-1-3) has only slightly reduced DNA concentration but consistently produced poor results during PCR which resulted in no read count assigned to fish (Figure 3.2; Appendix S3.1 Figure S3.4).

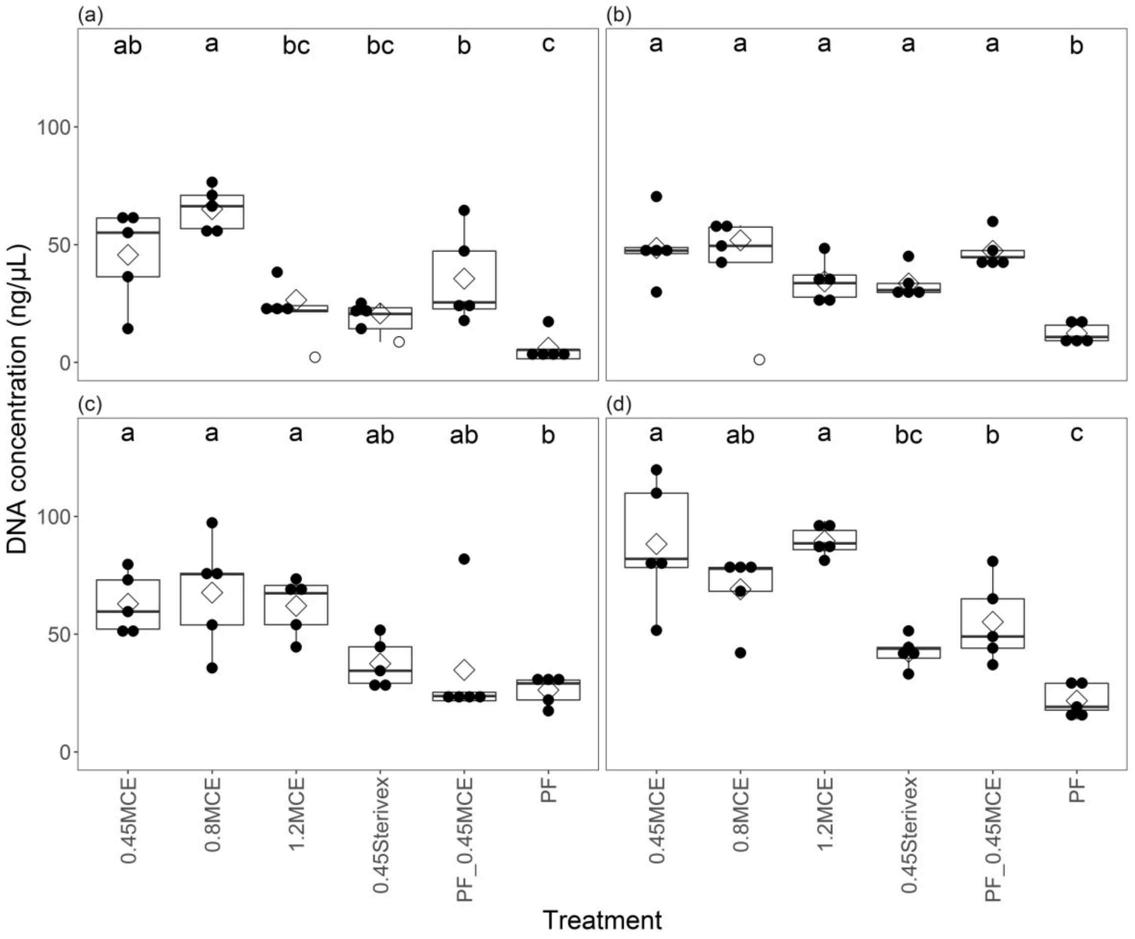


Figure 3.2 DNA yield recovered from six eDNA capture treatments from four ponds (a–d correspond to ponds E1–E4 respectively). Five replicates under each treatment. Treatments that differ significantly ($p < 0.05$) are indicated by the different letters in boxplots. "0.45MCE": 0.45 µm mixed cellulose acetate and nitrate (MCE) filters; "0.8MCE": 0.8 µm MCE filters; "1.2MCE": 1.2 µm MCE filters; "0.45Sterivex": 0.45 µm Sterivex-HV enclosed units; "PF_0.45MCE": 0.45 µm MCE filters after 20 µm qualitative cellulose pre-filters, and "PF": 20 µm qualitative cellulose pre-filters. *Notes:* "Diamonds ◊" show average values and the white dots represent outliers, identified in "Data analysis" (Section 3.2.4), are excluded downstream analysis.

3.2.4.3 Similarity and statistical analyses

All similarity and statistical analyses were performed in R v3.3.2 (R_Core_Team 2016), and graphs were plotted using GGPLOT2 v2.2.1 (Wickham & Chang 2016).

To better quantify the heterogeneity between filtration replicates, the Horn similarity index was calculated based on species relative abundance using SPADER v0.1.1 (Chao et al. 2016) with the function *SimilarityMult*. To investigate effects of different capture treatments on fish communities, non-metric multidimensional scaling (NMDS) allied with analysis of similarities (ANOSIM) were performed using the abundance-based Bray-Curtis dissimilarity index with the function *metaMDS* and *anosim* respectively in VEGAN v2.4-4 (Oksanen et al. 2017). The treatment with high repeatability should have high mean Horn index and low variation in NMDS ordination. The ANOSIM statistic R is based on the difference of mean ranks between treatments and within treatments.

Two-way analysis of variance (ANOVA) was conducted to test the interaction between four ponds and six treatments for filtration time, total DNA yield, the probability of species detection, Horn index, and the correlation coefficient between read counts and abundance or biomass after square-root or Tukey's ladder of powers transformation. Kruskal-Wallis one-way ANOVA with Dunn's test was conducted to test differences between the capture treatments for filtration time and Horn index. ANOVA with Tukey's test was conducted to test differences between the capture treatments for total DNA yield. The significance of linear correlations between read counts and abundance or biomass was evaluated by calculating the Pearson's product-moment correlation coefficient.

The full R script is available on the GitHub repository (https://github.com/HullUni-bioinformatics/Li_et_al_2018_eDNA_filtration/tree/master/R_script).

3.3 Results

3.3.1 Filtration time

The filtration time across all treatments and ponds varied from 3 to 120 min (Figure 3.3). There were significant effects of “treatment”, “pond”, the “interaction” between ponds and treatments across the entire dataset (Table 3.2, Global), and when comparing different treatments under specific aims (Table 3.2). The average filtration time differed considerably among the four ponds under the same filtration treatment, suggesting that SPM content varied among ponds (Appendix 3.1 Table S3.1). In relation to the specific comparisons: the filtration time decreased on average by 19.88 ± 14.17 min when the pore size increased from 0.45 to 0.8 μm and by 5.68 ± 5.98 min when the pore size increased from 0.8 to 1.2 μm. Overall, filtration time significantly decreased with increasing pore size, but the pattern was complex since significant interactions between treatments and ponds were observed (Table 3.2, Pore sizes). Individual *post hoc* tests showed that not all pairwise comparisons among pore sizes were significant (e.g., pond E4, Figure 3.3d). Filtration time was on average 18.00 ± 6.48 min longer using the “0.45Sterivex” compared to the “0.45MCE”. This pattern was also seen in three out of the four ponds when looked at individually, but none of the *post hoc* tests within ponds was significant (Figure 3.3). Across the four ponds, it was possible to filter 300 mL water in around 4 min using pre-filters themselves (Figure 3.3; Appendix S3.1 Table S3.1). Filtration time decreased on average by 27.00 ± 13.87 min when comparing the 0.45 μm filters after pre-filtration (“PF_0.45MCE”) to those without pre-filtration (“0.45MCE”); and this significant trend was observed in ponds E1 and E3 (Figure 3.3a, c).

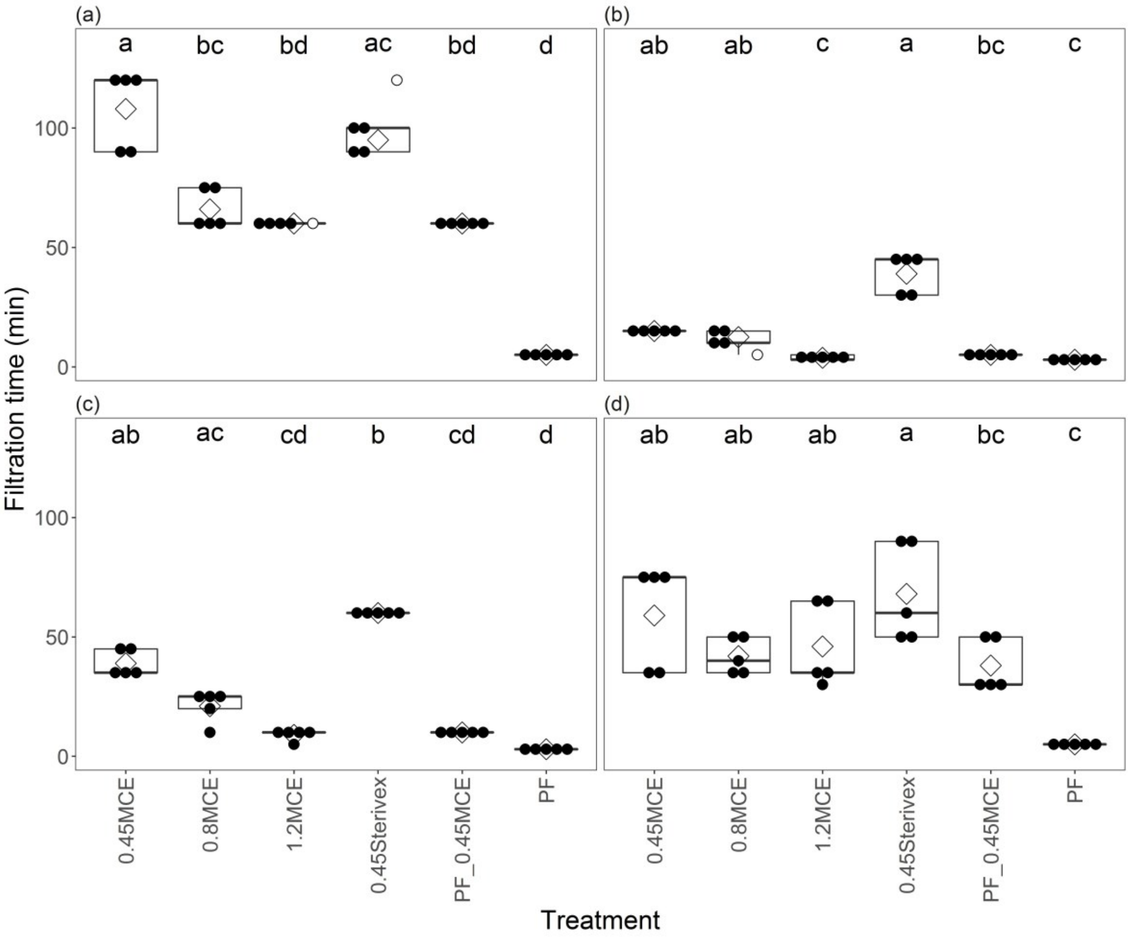


Figure 3.3 Filtration time of six eDNA capture treatments from four ponds (a–d correspond to ponds E1–E4 respectively). Five replicates under each treatment. Treatments that differ significantly ($p < 0.05$) are indicated by the different letters in boxplots. Abbreviations of treatments are the same as in Figure 3.2. *Notes*: "Diamonds ◊" show average values and the white dots represent outliers, identified in "Data analysis" (Section 3.2.4), are excluded downstream analysis.

Table 3.2 Two-way analysis of variance (ANOVA) results for filtration time, total DNA yield, the species detection probability, Horn index, correlation with abundance, and correlation with biomass using six eDNA capture treatments across four ponds (E1–E4).

Evaluation criterion	**Group**	**Treatment**	**Pond**	**Interaction**
Filtration time (min)	Global	F(5, 93) = 234.96***	F(3, 93) = 288.44***	F(15, 93) = 14.35***
	Pore sizes	F(2, 46) = 47.88***	F(3, 46) = 173.90***	F(6, 46) = 4.31**
	Filter types	F(1, 31) = 12.43**	F(3, 31) = 61.92***	F(3, 31) = 5.11**
	Pre-filtration	F(1, 32) = 123.11***	F(3, 32) = 169.41***	F(3, 32) = 4.12*
Total DNA yield (ng µL^{-1})	Global	F(5, 93) = 42.07***	F(3, 93) = 24.06***	F(15, 93) = 2.96***
	Pore sizes	F(2, 46) = 2.82; *p* = 0.07	F(3, 46) = 17.61***	F(6, 46) = 3.46**
	Filter types	F(1, 31) = 34.00***	F(3, 31) = 8.63***	F(3, 31) = 1.09; *p* = 0.36
	Pre-filtration	F(1, 32) = 8.57**	F(3, 32) = 4.49**	F(3, 32) = 1.43; *p* = 0.25
Probability of species detection	Global	F(5, 93) = 4.80***	F(3, 93) = 94.28***	F(15, 93) = 1.48; *p* = 0.13
	Pore sizes	F(2, 46) = 1.89; *p* = 0.16	F(3, 46) = 48.79***	F(6, 46) = 1.13; *p* = 0.36
	Filter types	F(1, 31) = 4.90*	F(3, 31) = 28.27***	F(3, 31) = 2.39; *p* = 0.09
	Pre-filtration	F(1, 32) = 0.65; *p* = 0.43	F(3, 32) = 32.54***	F(3, 32) = 2.85; *p* = 0.05
Horn index	Global	F(5, 204) = 14.09***	F(3, 204) = 34.67***	F(15, 204) = 6.55***
	Pore sizes	F(2, 100) = 10.33***	F(3, 100) = 30.29***	F(6, 100) = 9.31***
	Filter types	F(1, 68) = 53.63***	F(3, 68) = 5.18**	F(3, 68) = 4.29**
	Pre-filtration	F(1, 72) = 34.96***	F(3, 72) = 24.86***	F(3, 72) = 24.29**

Correlation with abundance	Global	F(5, 93) = 1.58; $p = 0.17$	F(3, 93) = 4.48*	F(15, 93) = 1.05; $p = 0.41$
	Pore sizes	F(2, 46) = 3.22*	F(3, 46) = 3.73*	F(6, 46) = 1.94; $p = 0.09$
	Filter types	F(1, 31) = 0.05; $p = 0.83$	F(3, 31) = 1.70; $p = 0.19$	F(3, 31) = 0.58; $p = 0.63$
	Pre-filtration	F(1, 32) = 0.0025; $p = 0.96$	F(3, 32) = 5.79**	F(3, 32) = 0.69; $p = 0.56$
Correlation with biomass	Global	F(5, 93) = 2.30; $p = 0.051$	F(3, 93) = 8.85***	F(15, 93) = 1.51; $p = 0.11$
	Pore sizes	F(2, 46) = 5.80**	F(3, 46) = 12.31***	F(6, 46) = 2.61*
	Filter types	F(1, 31) = 0.005; $p = 0.95$	F(3, 31) = 2.93*	F(3, 31) = 0.81; $p = 0.50$
	Pre-filtration	F(1, 32) = 0.44; $p = 0.51$	F(3, 32) = 7.53***	F(3, 32) = 0.21; $p = 0.89$

Notes. The compared treatments in three different groups are: pore sizes (0.45MCE, 0.8MCE and 1.2MCE), filter types (0.45MCE and 0.45Sterivex) and pre-filtration (0.45MCE and PF_0.45MCE). Replicates identified as outliers are excluded. Significant codes: *** 0.001; ** 0.01; * 0.05.

3.3.2 DNA yield

The DNA concentration across all treatments and ponds ranged from 1.15 to 119.70 ng μL^{-1} (Figure 3.2). There were significant effects of "treatment", "pond", the "interaction" between ponds and treatments across the entire dataset (Table 3.2, Global). In relation to the specific comparisons: there was no significant effect of different pore sizes of the filter (Table 3.2, Pore sizes, p = 0.07). Comparing the "0.45Sterivex" and the "0.45MCE", there were significant effects of "treatment" and "pond" (Table 3.2, Filter types). Individual *post hoc* tests showed that there was no significant difference between using the "0.45Sterivex" and the "0.45MCE" treatments from ponds E1 to E3, but the total DNA yield recovered from the "0.45Sterivex" was significantly lower than the "0.45MCE" in pond E4 (Figure 3.2d). The average DNA yield recovered from the pre-filters themselves ("PF") was the lowest of the six filtration treatments (Appendix S3.1 Table S3.1, 16.65 ± 9.85 ng/μL). After pre-filtration, the "PF_0.45MCE" still recovered 73.27 ± 10.56% total eDNA; hence only 26.73 ± 10.56% of the total eDNA remained on the 20 μm pre-filters. There were significant effects of "treatment" and "pond" between the "0.45MCE" and the "PF_0.45MCE" (Table 3.2, Pre-filtration). Individual *post hoc* tests showed that the total DNA yield recovered from the "0.45MCE" was significantly higher than the "PF_0.45MCE" in pond E4 only (Figure 3.2d).

3.3.3 Probability of species detection

All eight stocked species (bream *Abramis brama*, barbel *Barbus barbus*, crucian carp *Carassius carassius*, chub *Squalius cephalus*, dace *Leuciscus leuciscus*, roach *Rutilus rutilus*, rudd *Scardinius erythrophthalmus*, and tench *Tinca tinca*) were detected in this study (Figure 3.4). The rarest species in ponds E1 and E2 was bream. This species was not detected in pond E2 with any treatment, but it was detected with “0.45Sterivex” in pond E1. Roach were not detected using the pre-filters (“PF”) in pond E2 (Figure 3.4). In ponds E3 and E4, all stocked species were detected by all of the treatments (Figure 3.4c, d). There were significant effects of “treatment” and “pond” across the entire dataset, but there was no significant difference in “interaction” between ponds and treatments (Table 3.2, Global). In relation to the specific comparisons: there was no significant difference when comparing different filter pore sizes (Table 3.2, Pore sizes, $p = 0.16$), and filtration with and without pre-filters (Table 3.2, Pre-filtration, $p = 0.43$). The Sterivex units (“0.45Sterivex”) performed slightly better than the “0.45MCE” in terms of probability of species detection (Table 3.2, Filter types, $p < 0.05$). The average probability of species detection was the lowest using the pre-filters themselves (“PF”) of the six filtration treatments (Appendix S3.1 Table S3.1, 0.64 ± 0.27).

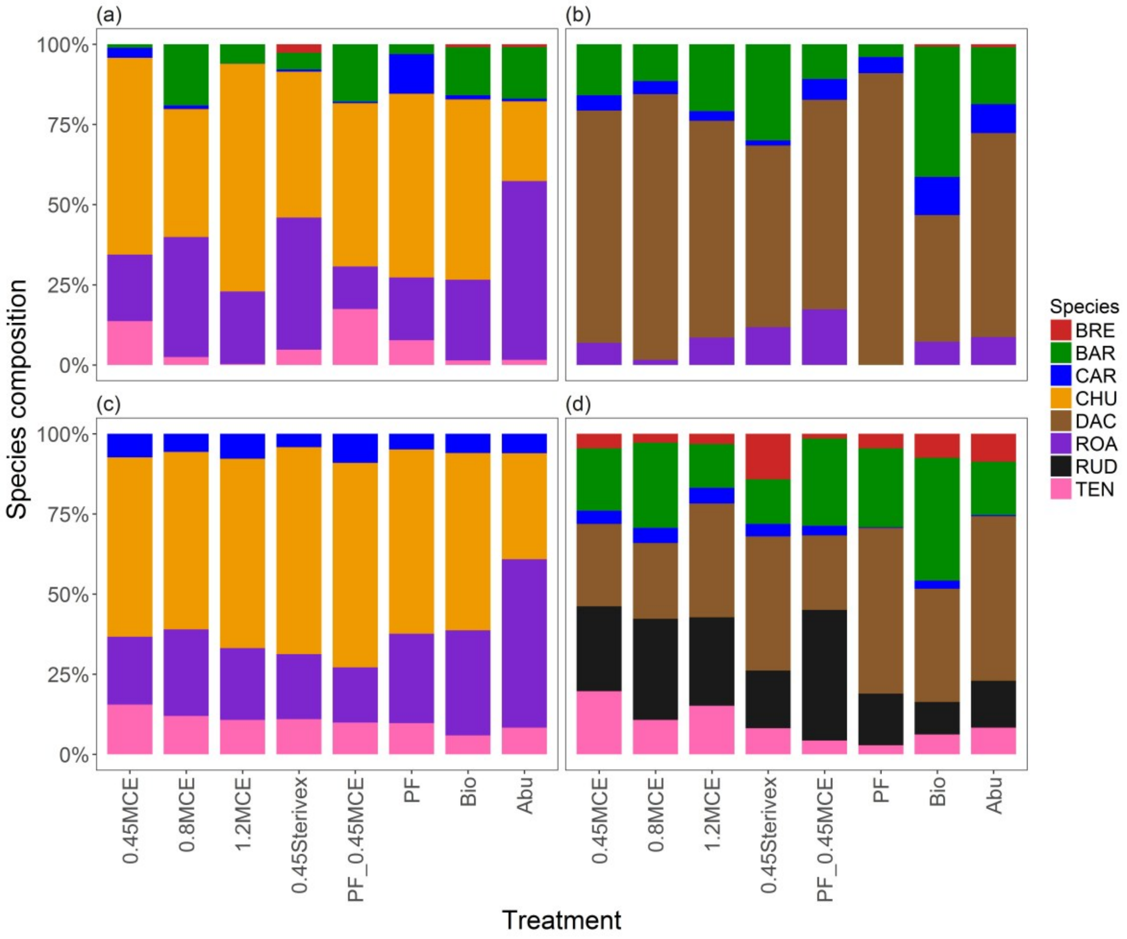


Figure 3.4 Species composition of averaged read counts (number of replicates = 5) using six eDNA capture treatments from four ponds (a–d correspond to ponds E1–E4 respectively). Species three letter codes are given in Table 3.1 and abbreviations of treatments are the same as in Figure 3.2. “Bio” and “Abu” refer to fish biomass and abundance density respectively, calculated based on Table 3.1. *Notes*: Replicates identified as outliers are excluded.

3.3.4 Variation between filtration replicates

Overall, there was considerable variation in species composition among individual filtration replicates within ponds (Figure 3.5a1, b1, c1, d1; Appendix S3.1 Figure S3.5). In terms of Horn index (similarity between replicates), there were significant effects of “treatment”, “pond”, the “interaction” between ponds and treatments across the entire dataset (Table 3.2, Global), and when comparing different treatments under specific aims (Table 3.2). The NMDS showed a high degree of overlap between the six capture treatments across four ponds (Figure 3.5a2, b2, c2, d2) indicating that different filtration treatments yielded broadly similar community composition estimates. Notable exceptions to this pattern were the pre-filters ("PF") and in some ponds (e.g., ponds E1 & E2) "PF_0.45MCE", where individual replicates were more widely scattered and often outside the ellipses of other treatments. In the ANOSIM test, the average values of the *R* statistic in global tests with all treatments were low (Appendix S3.1 Table S3.2, 0.15 ± 0.03), which showed that there was no obvious difference between treatments; and the *p* values suggesting that the variation was attributed to filtration replicates instead of treatments (Appendix S3.1 Table S3.2, $p = 0.03 \pm 0.02$).

In relation to the specific comparisons: overall, Horn index significantly decreased with increasing pore size, but the pattern was complex since significant interactions between treatments and ponds were observed (Table 3.2, Pore sizes). Individual *post hoc* tests showed that not all pairwise comparisons among pore sizes were significant (e.g., pond E2, Figure 3.5b1). The NMDS analysis showed that there was only clear discrimination between the "0.45MCE" and the "0.8MCE" in pond E1 (Figure 3.5a2; Appendix S3.1 Table S3.2, ANOSIM: $R = 0.52$, $p = 0.01$). There was greater variation among the "0.45Sterivex" replicates compared to the "0.45MCE" replicates (Figure 3.5). The community similarity of the "0.45Sterivex" was significantly lower than the "0.45MCE" across four ponds (Table 3.2, Filter types; Figure 3.5a1, b1, c1, d1). The NMDS ordination showed that significant difference was observed between the "0.45Sterivex" and the "0.45MCE" in ponds E3 (Figure 3.5c2; Appendix S3.1 Table S3.2, ANOSIM: $R = 0.64$, $p = 0.02$) and E4 (Figure 3.5d2; Appendix S3.1 Table S3.2, ANOSIM: $R = 0.30$, $p = 0.02$). Greater variance between replicates was observed for the pre-filters ("PF") themselves compared to other treatments (Figure 3.5). Repeatability was similar for the 0.45 µm filters when using pre-filters (“PF_0.45MCE”) and without using pre-filters (“0.45MCE”), except in pond E1 where the Horn index was significantly lower for “PF_0.45MCE” than “0.45MCE” (Figure 3.5a1). The NMDS ordination showed that there was no significant difference between the “PF_0.45MCE” and the “0.45MCE” across four ponds (Figure 3.5a2, b2, c2, d2; Appendix S3.1 Table S3.2, ANOSIM: $R = 0.07 \pm 0.06$, $p = 0.26 \pm 0.12$).

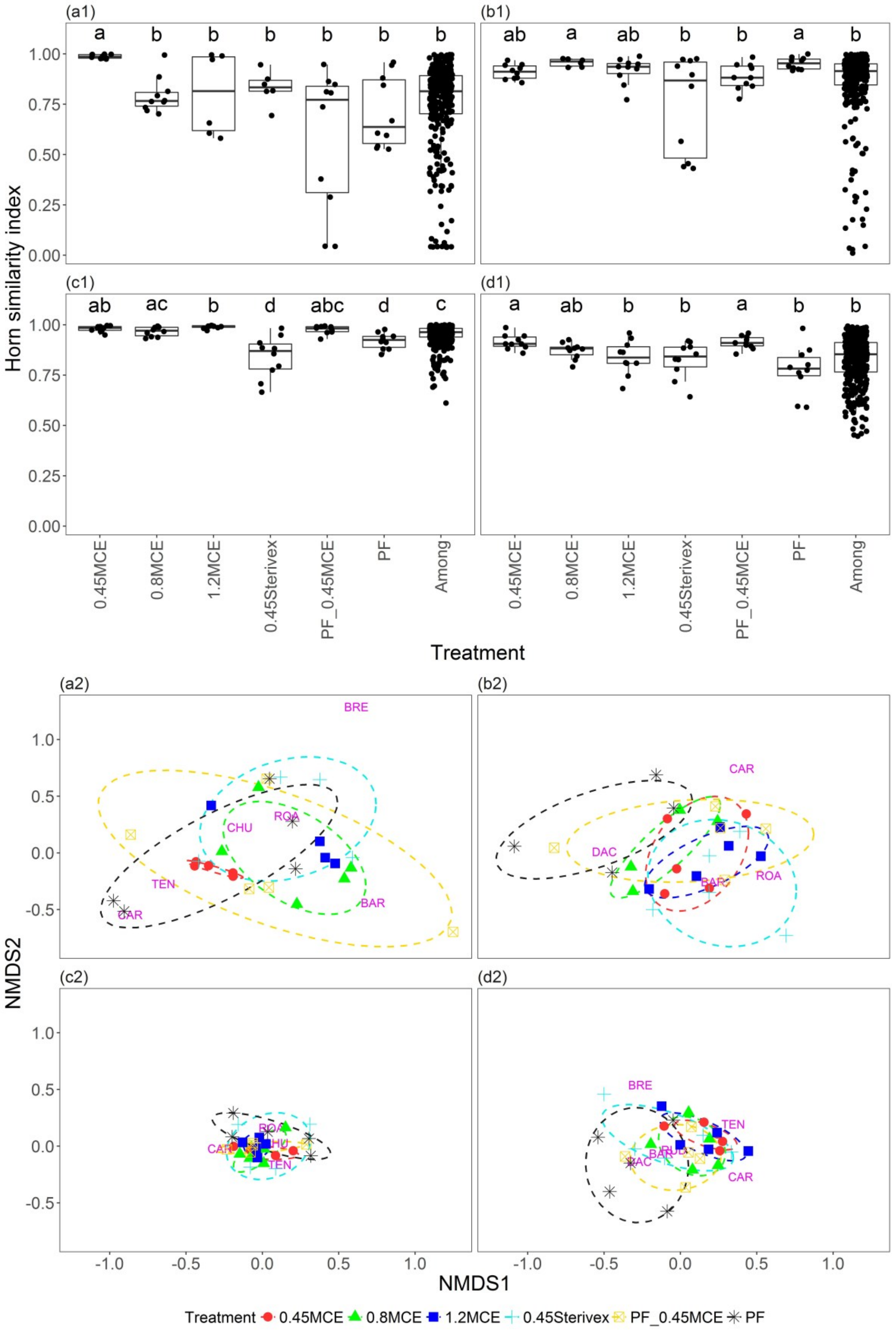

(a1)
a b b b b b b
(b1)
ab a ab b b a b
(c1)
ab ac b d abc d c
(d1)
a ab b b a b b
Horn similarity index
1.00
0.75
0.50
0.25
0.00
0.45MCE
0.8MCE
1.2MCE
0.45Sterivex
PF_0.45MCE
PF
Among
Treatment
(a2)
BRE
CHU
ROA
TEN
CAR
BAR
(b2)
CAR
DAC
ROA
BAR
(c2)
ROA
CAR
CHU
TEN
(d2)
BRE
TEN
DAC
BAR
CAR
NMDS2
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
NMDS1
Treatment
● 0.45MCE
▲ 0.8MCE
■ 1.2MCE
+ 0.45Sterivex
⊠ PF_0.45MCE
✳ PF

Figure 3.5 Pairwise Horn similarity index (a1–d1) and non-metric multidimensional scaling (NMDS) (a2–d2) based on six eDNA capture treatments from four ponds (a–d correspond to ponds E1-E4 respectively). "Among" refers to all filtration replicates among treatments within pond (A1–D1). Treatments that differ significantly (p < 0.05) are indicated by the different letters in boxplots (a1–d1). The ellipse indicates the 50% similarity level within each capture treatment (a2–d2). Species three letter codes are given in Table 3.1 and abbreviations of treatments are the same as in Figure 3.2. *Notes:* Five replicates under each treatment and replicates identified as outliers are excluded.

3.3.5 Correlations between read counts and fish abundance or biomass

There were consistent, positive correlations between average read counts of five replicates and fish abundance or biomass across the six treatments and four ponds (Figure 3.6; Appendix S3.1 Figure S3.6). There was no significant effect of "treatment", or "interaction" between ponds and treatments, on correlations between read counts and abundance or biomass across the entire dataset (Table 3.2, Global). In relation to the specific comparisons: overall, there were significant effects of different pore sizes of the filter (Table 3.2, Pore sizes). Individual *post hoc* tests showed that a significant difference in correlations between read counts and abundance or biomass was only observed between "0.45MCE" and "1.2MCE" treatments, and the 1.2 μm MCE filters performed better than 0.45 μm MCE filters. There was no significant effect on correlations between read counts and abundance or biomass between "0.45Sterivex" and "0.45MCE" treatments (Table 3.2, Filter types), and filtration with and without pre-filtration (Table 3.2, Pre-filtration).

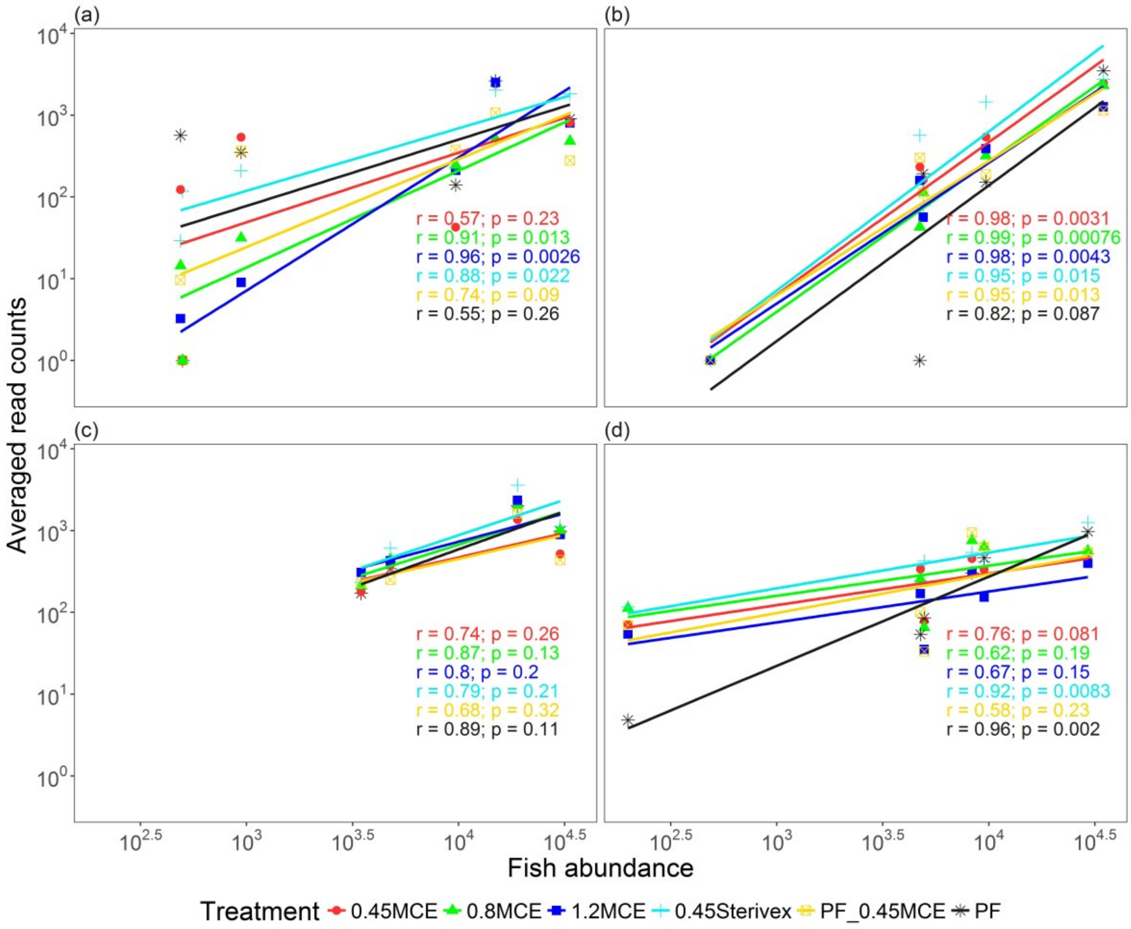


Figure 3.6 Correlations between averaged read counts (number of replicates = 5) and fish abundance using six eDNA capture treatments from four ponds (a–d correspond to ponds E1–E4 respectively). Abbreviations of treatments are the same as in Figure 3.2. *Notes:* Replicates identified as outliers are excluded.

3.4 Discussion

3.4.1 Optimal pore size of membrane filter

Turner et al. (2014) previously determined that aqueous eDNA particles from common carp *Cyprinus carpio* range between < 0.2 and > 180 μm and therefore recommended 0.2 μm pore size filters for optimal capture of common carp eDNA. In a pilot study, this pore size of filter were observed led to clogging quickly; therefore three pore sizes (0.45, 0.8 and 1.2 μm) of the membrane filter were compared.

The study demonstrated that the pore size of filter has a considerable impact on filtration time. When changing from 0.45 to 0.8 µm filters, on average, 36% filtration time is saved, whereas only 15% filtration time is saved increasing pore size from 0.8 to 1.2 µm. This result supports previous studies (Turner et al. 2014; Eichmiller et al. 2016; Minamoto et al. 2016) indicating that the smaller pore size of filters is more likely to clog and increase filtration time. However, different pore sizes do not affect the amount of total eDNA recovered and the probability of species detection. The similarity among filtration replicates decreases with increasing pore size, and the repeatability among filtration replicates using the 0.45 µm MCE filters is the highest compared to the other pore sizes of the filter. This in turns indicates that stochastic sampling effects could be minimised by using a smaller pore size of the filters. The finding is supported by other metabarcoding studies demonstrated that the performance of 0.45 µm filters in representing the community composition is more consistent (Miya et al. 2016; Majaneva et al. 2018). After pooling that data from all five replicates consistently positive relationships are found between read counts and fish abundance or biomass, although correlations are not always statistically significant. The 0.8 and 1.2 µm MCE filters perform better than 0.45 µm MCE filters in terms of correlations between read counts and fish abundance or biomass. In contrast, Eichmiller et al. (2016) found that different pore sizes (0.2, 0.6, 1.0 and 5.0 µm) of PC filter affect the slope of the common carp biomass/eDNA copies relationship; and 0.2–0.6 µm filters are optimal for biomass quantification in the laboratory. Turner et al. (2014) showed that PC filters have relatively uniform sized pores, in contrast, the MCE filters are less uniform and more likely to retain particles by entrapment. The structural difference between PC filters and MCE filters could explain why our results are different from Eichmiller et al. (2016). Previous studies have also demonstrated that filter materials can also drastically affect the recovery of eDNA (Liang & Keeley 2013; Renshaw et al. 2015; Hinlo et al. 2017). The other potential reason for difference between studies could be that previous studies were based on target species detection via qPCR assays, comparing absolute DNA concentrations across samples, as opposed to metabarcoding of the whole community comparing relative sequencing read counts in the current study. In support of this, Djurhuus et al. (2017) found that different filter materials do not result in different richness and community composition based on metabarcoding.

The 0.45 μm MCE filters perform the best among the six filtration treatments in terms of DNA yield, repeatability within pond and consistency between ponds. However, filtration time is significantly longer for the 0.45 μm MCE filters than the 0.8 μm MCE filters. The correlations between read counts and fish abundance or biomass recovered by the 0.8 μm MCE filters are slightly better than those of the 0.45 μm MCE filters even though there is no significant difference between the treatments. Therefore, the 0.8 μm MCE filters appear to provide a reasonable balance between filtration time and quantification efficacy in this study and may be optimal in turbid, eutrophic, high fish density water bodies, whereas 0.45 μm MCE filters may be more suitable to clearer waters (Figure 3.1).

3.4.2 Performance of enclosed (Sterivex) filters

Previous studies showed that filtration using enclosed Sterivex units is an effective protocol for capturing target species DNA with qPCR assays (Keskin 2014; Bergman et al. 2016; Spens et al. 2017). To our knowledge, Spens et al. (2017) is the only published study comparing Sterivex units with membrane filters using qPCR. Here, I directly compared the performance of MCE filters and Sterivex units of the same pore size via metabarcoding.

On average, filtration time using the Sterivex units increases by 18 min per sample compared to using 0.45 µm MCE filters. This difference is not due to vacuum pumps as the same pump was used for both filter types. However, Spens et al. (2017) observed that 1 L clear lake water can be filtered through 0.22 µm Sterivex units in around 10 min using 50 mL syringes comparing to 0.45 µm MCE filters (described as CN filters in the study) in 15–30 min using a vacuum pump. To minimise filtration time, the Sterivex units are recommended to use together with prepacked sterile syringes in situations where on-site filtration is required (Figure 3.1). With respect to DNA yield, the 0.45 µm Sterivex filters recover slightly less DNA than the 0.45 µm MCE filters. The Horn index and NMDS ordination showed there is a greater variation among the 0.45 µm Sterivex replicates compared to the 0.45 µm MCE replicates. However, the correlations between read counts and fish biomass or abundance are not significantly different between the treatments when all data were pooled. Therefore, 0.45 µm Sterivex units can be considered an efficient eDNA capture method for metabarcoding.

3.4.3 Efficiency and impact of pre-filtration

The water from Calverton fish ponds is turbid and eutrophic, with high levels of algae. Our pilot study showed that a small amount of water (i.e., 250 mL) could be filtered through 1.2 µm filters before clogging. This is considerably less than previous metabarcoding studies in less eutrophic lakes, in which at least 1 L water was filtered (Hänfling et al. 2016; Port et al. 2016) and reduced sample volumes could potentially impact rare species detection. Pre-filtration could potentially help to prevent clogging, substantially reduce filtration time, and reduce the capture of unwanted SPM and PCR inhibitors. I therefore investigated the impact of pre-filtration by comparing results from

0.45 µm MCE filters with and without passing through 20 µm pre-filters, as well as the analysing pre-filters themselves.

Across the four ponds, it is possible to filter 300 mL water in around 4 min using the pre-filters themselves. The pre-filtering step reduces the filtration time through the 0.45 µm MCE filters by approximately 50%, resulting in a considerable overall time saving per sample. This could be an important consideration when eutrophic habitat or water with high sediment content is sampled. After pre-filtration, 73.27% total eDNA is recovered on the 0.45 µm MCE filters (with a corresponding 26.73% total eDNA remained on pre-filters). Pre-filtration followed by capture onto 0.45 µm MCE filters did not result in a significantly different probability of species detection, repeatability between filtration replicates, and correlations between read counts and fish biomass or abundance when compared to other treatments. However, Majaneva et al. (2018) demonstrated that pre-filtration (12 µm pre-filters with 0.45 µm filters), could potentially reduce the number of detected metazoan taxa, although it recovers higher diversity index values and more consistent community composition.

In terms of the pre-filters themselves, the overall probability of species detection (0.64 ± 0.27) is lower than other membrane filters, and greater variance between replicates is observed compared to other treatments. Similar results were found by Robson et al. (2016), who showed that 2 L water samples can be filtered in less than 3 min using 20 µm filters, but a 0.57 probability of single-species detection is achieved compared to 1.00 probability using 3 µm PC filters.

The results indicate that pre-filtration with 20 µm filters could prevent SPM from clogging finer filters without affecting metabarcoding results but that the pre-filters themselves are not suitable for metabarcoding due to the potential of reduced total DNA yield, the probability of species detection and repeatability. Despite the advantages of pre-filtration demonstrated here, it should be noted that there is a drawback of pre-filtration in terms of more handling, which could increase the opportunity for contamination (Turner et al. 2014). Thus, I recommend pre-filters are applied only if absolutely essential for reducing the filtration time or increasing the throughput volume of the capture filters (Figure 3.1).

3.5 Conclusion

This study demonstrates that the DNA yield, the probability of species detection, and correlations between abundance/biomass and read counts are encouragingly comparable between different filter types (0.45 μm MCE filters and 0.45 μm Sterivex units) and pore sizes (0.45, 0.8 and 1.2 μm). Therefore, eDNA metabarcoding results seem quite robust to the choice of the filtration method when a sufficient number of replicates is carried out. It is worth noting that the suitability of various pore sizes of the filter to capture eDNA is likely to be heavily influenced by the heterogeneous nature of water bodies. For turbid, eutrophic, high fish density ponds, such as those studied here, 0.8 μm MCE filters provide the optimal trade-off between rapid filtration time and the probability of species detection, but smaller pore sizes of the filter may be more suitable for clearer, low species density conditions. Further study of the impact of heterogeneity (in terms of SPM, biochemical oxygen demand “BOD”, chemical oxygen demand “COD”, dissolved oxygen “DO”, pH, and watercolour etc.) between water bodies on eDNA capture is required. Finally, this study report high variation among filtration replicates, which is consistent with Lanzén et al. (2017) who indicated that technical replicates of DNA extraction can improve diversity and compositional dissimilarity. Spatial heterogeneity of eDNA within water bodies has also been reported in several studies (e.g. Jerde et al. 2011; Pilliod et al. 2013; Civade et al. 2016; Hänfling et al.

2016). Future studies, for example incorporating species occupancy models for imperfect species detection (Pilliod et al. 2013; Schmidt et al. 2013; Hänfling et al. 2016; Valentini et al. 2016), are needed to further investigate the multiple opportunities for heterogeneity encountered in eDNA studies.

3.6 Supporting Information

Appendix S3.1 Supplementary tables and figures

Table S3.1 Summary of filtration time, total DNA yield, the species detection probability, Horn index, correlation with abundance, and correlation with biomass using six eDNA capture treatments across four ponds (E1–E4).

Evaluation criterion	**Treatment**	**E1**	**E2**	**E3**	**E4**	**Summary**
Filtration time (min)	0.45MCE	108.00 ± 16.43	15.00 ± 0.00	39.00 ± 5.48	59.00 ± 21.91	55.25 ± 37.36
	0.8MCE	66.00 ± 8.22	12.50 ± 2.89	21.00 ± 6.52	42.00 ± 7.58	36.58 ± 21.99
	1.2MCE	60.00 ± 0.00	3.80 ± 1.10	9.00 ± 2.24	46.00 ± 17.46	28.11 ± 25.49
	0.45Sterivex	95.00 ± 5.77	39.00 ± 8.22	60.00 ± 0.00	68.00 ± 20.49	63.95 ± 22.58
	PF_0.45MCE	60.00 ± 0.00	5.00 ± 0.00	10.00 ± 0.00	38.00 ± 10.95	28.25 ± 23.36
	PF	5.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	5.00 ± 0.00	4.00 ± 1.03
Total DNA yield (ng μL^{-1})	0.45MCE	45.69 ± 20.33	48.54 ± 14.41	62.93 ± 12.89	88.27 ± 27.10	61.36 ± 24.89
	0.8MCE	65.07 ± 9.22	51.85 ± 7.41	67.60 ± 23.54	69.06 ± 15.75	64.00 ± 15.82
	1.2MCE	26.45 ± 7.95	34.38 ± 9.14	61.98 ± 12.28	89.54 ± 6.64	54.49 ± 26.76
	0.45Sterivex	20.76 ± 4.73	33.56 ± 6.63	37.47 ± 10.38	42.44 ± 6.72	34.23 ± 10.47
	PF_0.45MCE	35.55 ± 19.76	47.35 ± 7.43	34.77 ± 26.35	55.17 ± 17.67	43.21 ± 19.55
	PF	6.19 ± 6.48	12.40 ± 4.61	26.27 ± 6.29	21.76 ± 7.06	16.65 ± 9.85
Species detection probability	0.45MCE	0.77 ± 0.09	0.64 ± 0.09	1.00 ± 0.00	0.93 ± 0.09	0.83 ± 0.16
	0.8MCE	0.64 ± 0.08	0.60 ± 0.16	1.00 ± 0.00	0.90 ± 0.09	0.79 ± 0.19
	1.2MCE	0.67 ± 0.13	0.72 ± 0.11	1.00 ± 0.00	0.93 ± 0.09	0.84 ± 0.17
	0.45Sterivex	0.50 ± 0.14	0.60 ± 0.20	0.95 ± 0.11	0.93 ± 0.09	0.76 ± 0.24
	PF_0.45MCE	0.53 ± 0.22	0.68 ± 0.18	1.00 ± 0.00	0.97 ± 0.08	0.79 ± 0.24

	PF	0.50 ± 0.12	0.40 ± 0.20	1.00 ± 0.00	0.80 ± 0.14	0.64 ± 0.27
Horn index	0.45MCE	0.99 ± 0.01	0.91 ± 0.04	0.98 ± 0.01	0.91 ± 0.04	0.95 ± 0.04
	0.8MCE	0.79 ± 0.09	0.96 ± 0.02	0.97 ± 0.02	0.87 ± 0.04	0.89 ± 0.09
	1.2MCE	0.80 ± 0.20	0.92 ± 0.06	0.99 ± 0.01	0.84 ± 0.08	0.90 ± 0.12
	0.45Sterivex	0.83 ± 0.08	0.75 ± 0.24	0.84 ± 0.10	0.83 ± 0.09	0.81 ± 0.15
	PF_0.45MCE	0.58 ± 0.35	0.89 ± 0.06	0.98 ± 0.02	0.91 ± 0.03	0.84 ± 0.23
	PF	0.71 ± 0.18	0.95 ± 0.03	0.92 ± 0.04	0.78 ± 0.12	0.84 ± 0.15
Correlation with abundance	0.45MCE	Cor = 0.57, *p* = 0.23	Cor = 0.98**	Cor = 0.74, *p* = 0.26	Cor = 0.76, *p* = 0.08	Cor = 0.69**
	0.8MCE	Cor = 0.91*	Cor = 0.99**	Cor = 0.87, *p* = 0.13	Cor = 0.62, *p* = 0.19	Cor = 0.80**
	1.2MCE	Cor = 0.96**	Cor = 0.98**	Cor = 0.80, *p* = 0.20	Cor = 0.67, *p* = 0.15	Cor = 0.84**
	0.45Sterivex	Cor = 0.88*	Cor = 0.95*	Cor = 0.79, *p* = 0.21	Cor = 0.92*	Cor = 0.78**
	*P*F_0.45MCE	Cor = 0.74, *p* = 0.09	Cor = 0.95*	Cor = 0.68, *p* = 0.32	Cor = 0.58, *p* = 0.23	Cor = 0.74**
	*P*F	Cor = 0.55, *p* = 0.26	Cor = 0.82, *p* = 0.09	Cor = 0.89, *p* = 0.11	Cor = 0.96**	Cor = 0.69**
Correlation with biomass	0.45MCE	Cor = 0.62, *p* = 0.19	Cor = 0.96*	Cor = 0.88, *p* = 0.12	Cor = 0.68, *p* = 0.14	Cor = 0.75**
	0.8MCE	Cor = 0.91*	Cor = 0.97*	Cor = 0.94, *p* = 0.06	Cor = 0.70, *p* = 0.12	Cor = 0.88**
	1.2MCE	Cor = 0.99**	Cor = 0.96*	Cor = 0.95, *p* = 0.05	Cor = 0.61, *p* = 0.20	Cor = 0.91**
	0.45Sterivex	Cor = 0.85*	Cor = 0.94*	Cor = 0.90, *p* = 0.10	Cor = 0.82, *p* = 0.05	Cor = 0.82**
	PF_0.45MCE	Cor = 0.76, *p* = 0.08	Cor = 0.91*	Cor = 0.90, *p* = 0.10	Cor = 0.70, *p* = 0.12	Cor = 0.81**
	PF	Cor = 0.61, *p* = 0.19	Cor = 0.81, *p* = 0.10	Cor = 0.96*	Cor = 0.93*	Cor = 0.71**

Notes: Abbreviations of treatments are the same as in Figure 3.2. Five replicates under each treatment and replicates identified as outliers are excluded. Significant codes: *** 0.001; ** 0.01; * 0.05.

Table S3.2 Analysis of similarities (ANOSIM) pairwise comparisons of fish community structures obtained using six eDNA capture treatments from four ponds (E1–E4).

Pairwise comparisons	**E1**		**E2**		**E3**		**E4**		**Mean ± *SD***	
	R	***p***	***R***	***p***	***R***	***p***	***R***	***p***	***R***	***p***
Global test	0.17	0.02	0.19	0.01	0.11	0.05	0.12	0.02	0.15 ± 0.03	0.03 ± 0.02
0.45MCE, 0.8MCE	0.52	0.01	-0.09	0.76	0.04	0.23	0.06	0.37	0.13 ± 0.23	0.34 ± 0.32
0.45MCE, 1.2MCE	0.31	0.07	0.20	0.14	0.24	0.09	0.00	0.43	0.19 ± 0.11	0.18 ± 0.17
0.8MCE, 1.2MCE	-0.03	0.56	0.24	0.07	0.04	0.28	0.26	0.06	0.13 ± 0.13	0.24 ± 0.23
0.45MCE, 0.45Sterivex	0.14	0.21	0.09	0.23	0.64	0.02	0.30	0.02	0.29 ± 0.22	0.12 ± 0.11
0.45MCE, PF_0.45MCE	0.08	0.14	0.15	0.19	-0.02	0.42	0.06	0.31	0.07 ± 0.06	0.26 ± 0.12

Note: Abbreviations of treatments are the same as in Figure 3.2. R values were derived from Bray-Curtis dissimilarity matrices. Five replicates under each treatment and replicates identified as outliers are excluded.

Table S3.3 Nested ANOVA fiting the generalised linear model (GLM) results for filtration time, total DNA yield, the species detection probability, Horn index, correlation with abundance, and correlation with biomass using six eDNA capture treatments across four ponds (E1–E4).

Evaluation criterion	**Random effects**		**Fixed effects**			
	Groups	**Variance**	**Treatment**	**Estimate**	**t-value**	***p*-value**
Filtration time (min)	Treatment : Pond	0.58 ± 0.76	T1_0.45	5.74 ± 0.82	6.97	0.00136**
			T2_0.8	4.73 ± 0.56	-1.79	0.09
	Pond	2.08 ± 1.44	T3_1.2	4.08 ± 0.56	-2.96	0.00969**
			T4_Sterivex	6.47 ± 0.56	1.31	0.21
	Residual	0.21 ± 0.46	T5_PF_0.45	4.07 ± 0.56	-2.97	0.00958**
			TP5_PF	1.85 ± 0.56	-6.93	4.82e-06***
Total DNA yield (ng µL^{-1})	Treatment : Pond	0.43 ± 0.65	T1_0.45	7.67 ± 0.60	12.85	3.68e-06***
			T2_0.8	7.91 ± 0.57	0.42	0.68
	Pond	0.79 ± 0.89	T3_1.2	7.08 ± 0.57	-1.05	0.31
			T4_Sterivex	5.72 ± 0.57	-3.43	0.0037**
	Residual	1.06 ± 1.03	T5_PF_0.45	6.41 ± 0.56	-2.23	0.04*
			TP5_PF	3.86 ± 0.56	-6.76	6.95e-06***
Species detection probability	Treatment : Pond	0.002 ± 0.04	T1_0.45	0.80 ± 0.12	6.37	0.00495**
			T2_0.8	0.74 ± 0.05	-1.11	0.28
	Pond	0.06 ± 0.24	T3_1.2	0.79 ± 0.05	-0.10	0.92
			T4_Sterivex	0.70 ± 0.05	-1.84	0.08
	Residual	0.02 ± 0.13	T5_PF_0.45	0.75 ± 0.05	-0.78	0.45
			TP5_PF	0.62 ± 0.05	-3.40	0.00403**
Horn index	Treatment : Pond	0.02 ± 0.16	T1_0.45	0.73 ± 0.11	6.59	7.2e-05***
			T2_0.8	0.56 ± 0.12	-1.40	0.18
	Pond	0.02 ± 0.14	T3_1.2	0.60 ± 0.12	-1.11	0.28
			T4_Sterivex	0.37 ± 0.12	-3.00	0.009**

	Residual	0.04 ± 0.21	T5_PF_0.45	0.53 ± 0.12	-1.69	0.11
			TP5_PF	0.46 ± 0.12	-2.25	0.04*
Correlation with abundance	Treatment : Pond	0.0004 ± 0.02	T1_0.45	0.38 ± 0.08	4.91	0.00052***
			T2_0.8	0.51 ± 0.09	1.4	0.18
	Pond	0.0088 ± 0.09	T3_1.2	0.59 ± 0.09	2.3	0.04
			T4_Sterivex	0.42 ± 0.09	0.47	0.65
	Residual	0.0753 ± 0.27	T5_PF_0.45	0.39 ± 0.09	0.12	0.9
			TP5_PF	0.44 ± 0.09	0.62	0.54
Correlation with biomass	Treatment : Pond	0.006 ± 0.08	T1_0.45	0.40 ± 0.09	4.47	0.00156**
			T2_0.8	0.52 ± 0.10	1.22	0.24
	Pond	0.015 ± 0.12	T3_1.2	0.64 ± 0.10	2.44	0.03*
			T4_Sterivex	0.43 ± 0.10	0.22	0.83
	Residual	0.060 ± 0.24	T5_PF_0.45	0.45 ± 0.10	0.51	0.62
			TP5_PF	0.46 ± 0.10	0.54	0.6

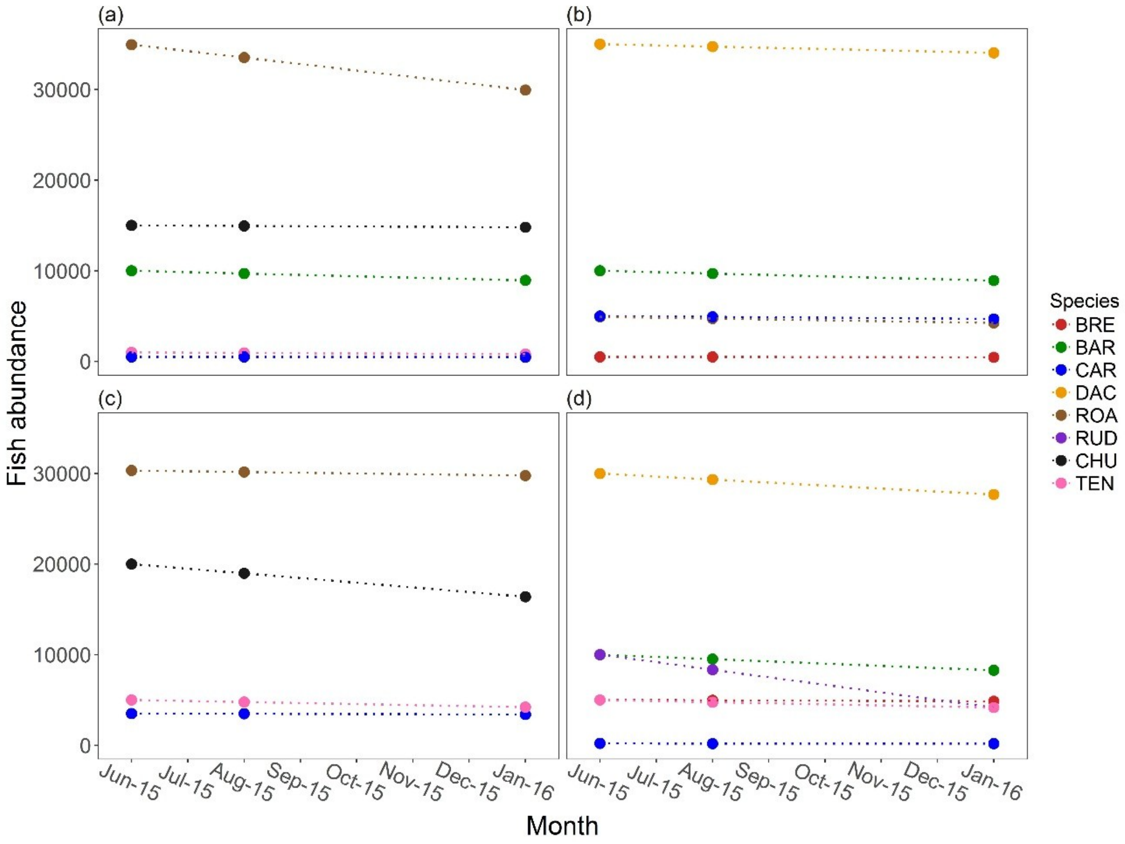


Figure S3.1 Death curves in the four Calverton fish ponds (a–d correspond to ponds E1–E4 respectively) from June 2015 to January 2016.

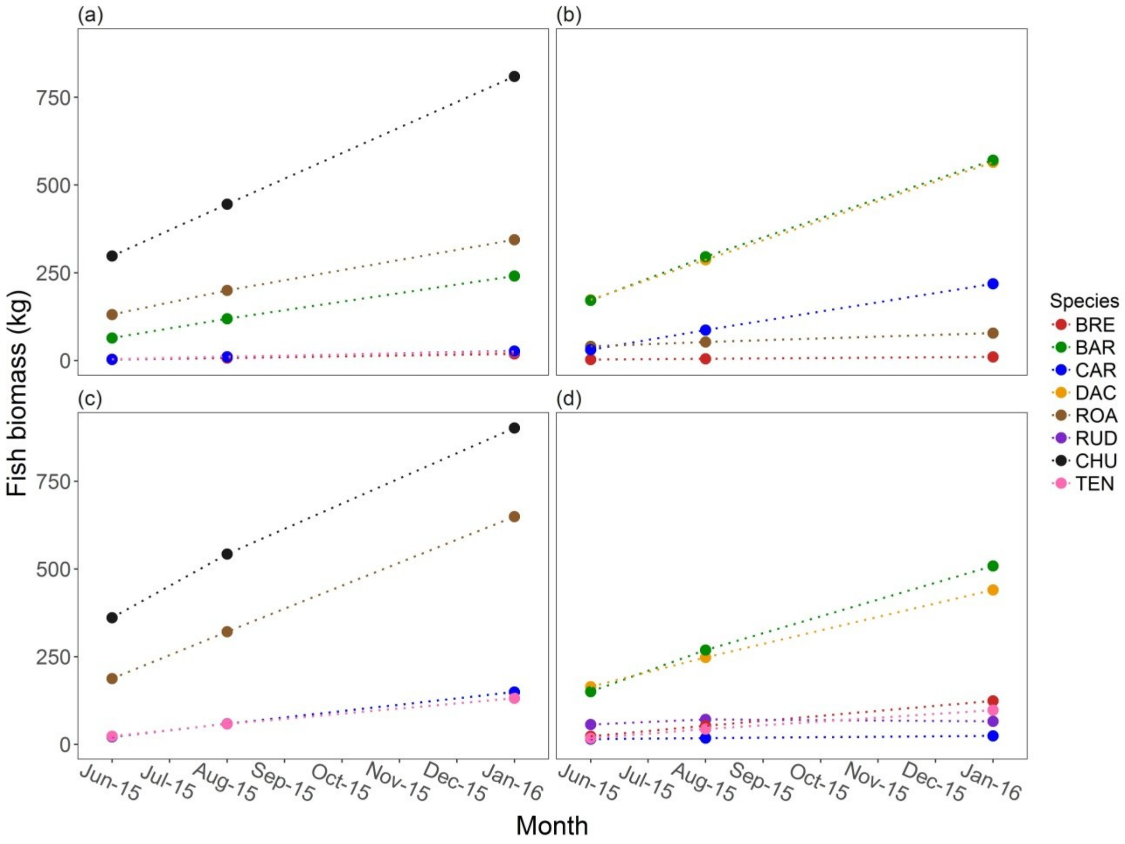


Figure S3.2 Growth curves in the four Calverton fish ponds (a–d correspond to ponds E1–E4 respectively) from June 2015 to January 2016.

Figure S3.3 Maximum likelihood phylogenetic tree of the all 12S sequences from the custom reference database (.png; supplied in a separate file).

The file can be viewed or downloaded use the link as below:

https://github.com/HullUni-

bioinformatics/Li_et_al_2018_eDNA_filtration/blob/master/Figure_S3.png

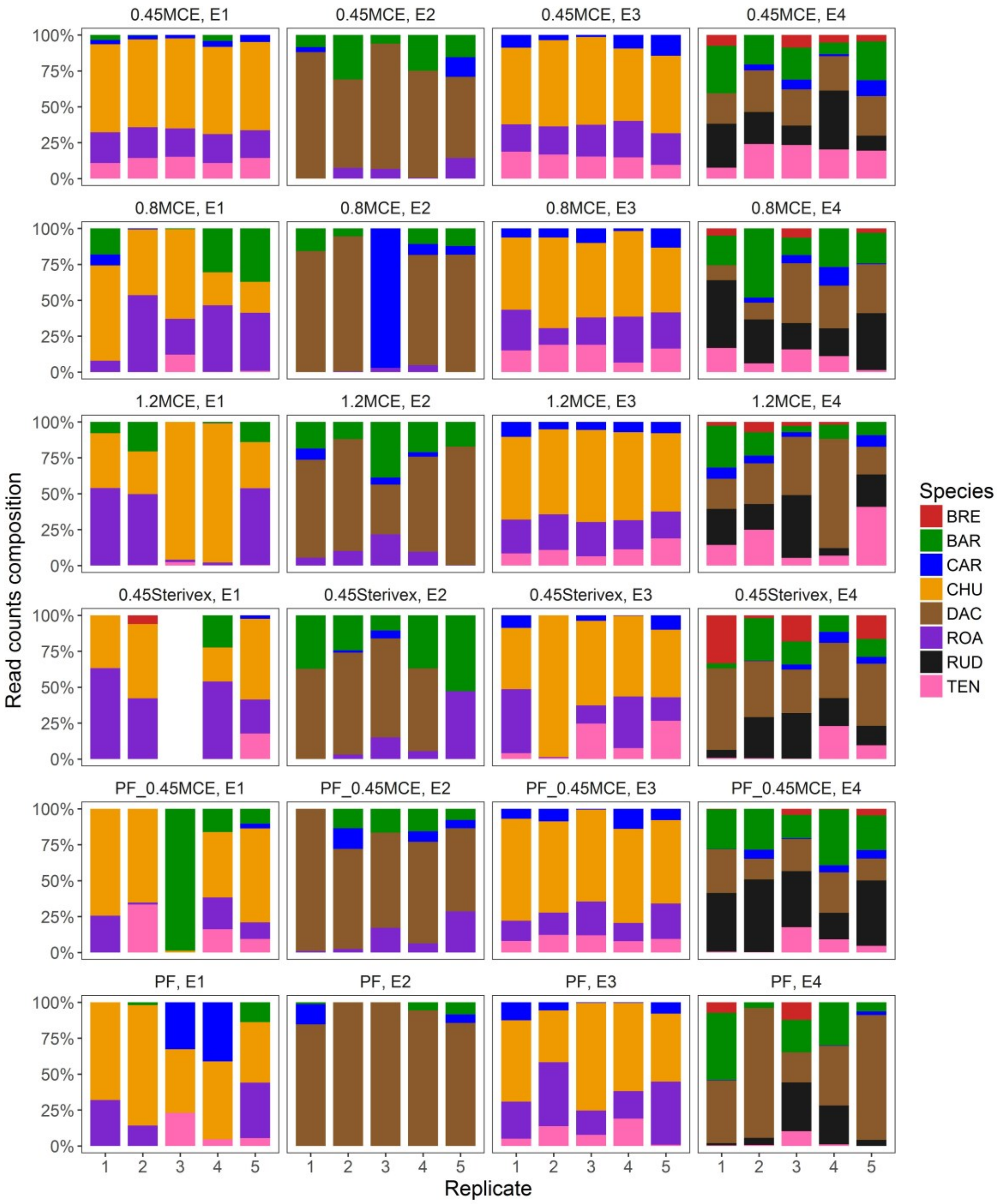


Figure S3.4 Fish composition of read counts under each replicate using six eDNA capture treatments from four fish ponds (E1–E4). Abbreviations of treatments are the same as in Figure 3.2.

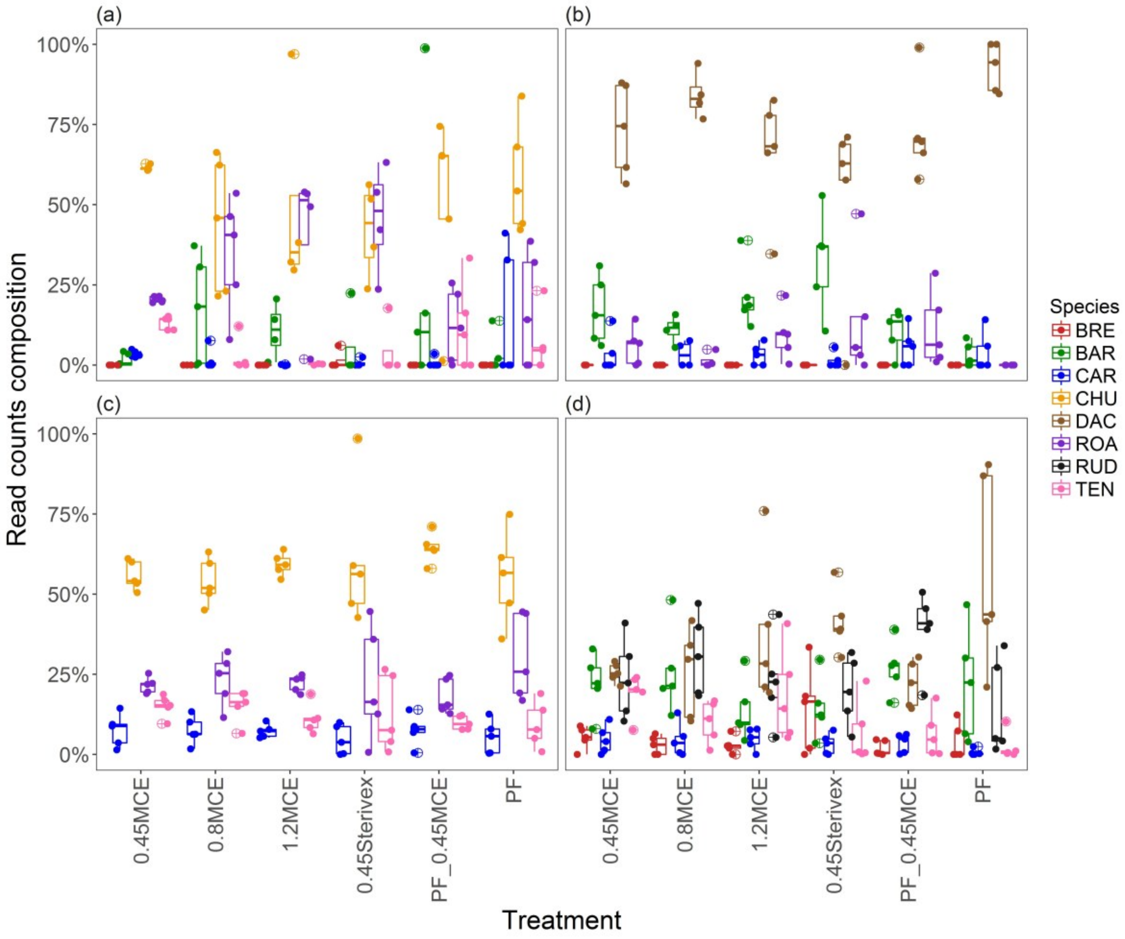


Figure S3.5 Boxplot of species composition under each replicate using six eDNA capture treatments from four fish ponds (a–d correspond to ponds E1–E4 respectively). Species three letter codes are given in Table 3.1 and abbreviations of treatments are the same as in Figure 3.2. *Notes*: Replicates identified as outliers are excluded.

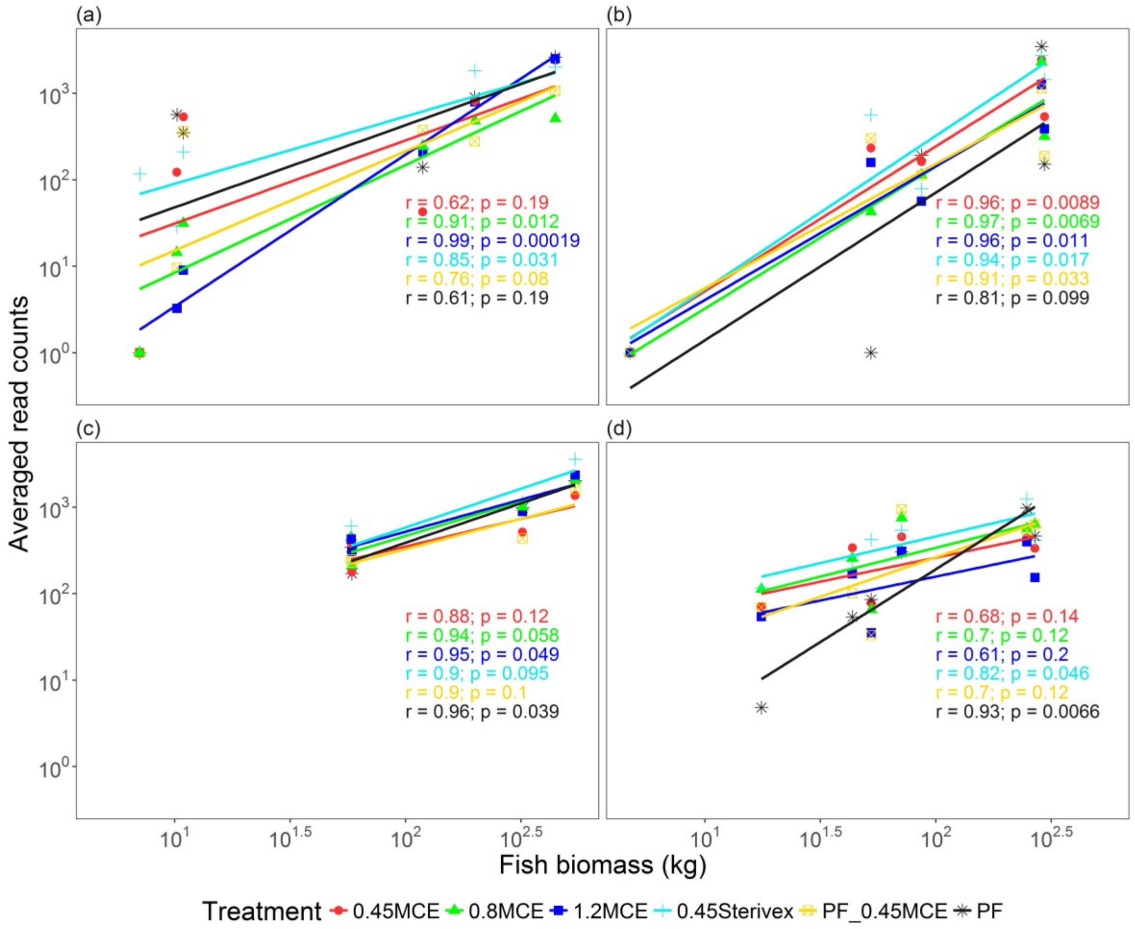


Figure S3.6 Correlations between averaged read counts (number of replicates = 5) and fish biomass using six eDNA capture treatments from four ponds (a–d correspond to ponds E1–E4 respectively). Abbreviations of treatments are the same as in Figure 3.2. *Notes*: Replicates identified as outliers are excluded.

Appendix S3.2 Read counts of OTUs data was used for the R script (.csv; supplied in a separate file)

The file can be viewed or downloaded use the link as below:

https://github.com/HullUni-bioinformatics/Li_et_al_2018_eDNA_filtration/blob/master/Appendix_S1.csv

Chapter 4 Limited dispersion and quick degradation of environmental DNA in fish ponds inferred by metabarcoding[3]

Abstract

In this study, environmental DNA (eDNA) metabarcoding is applied to explore the spatial and temporal distribution of eDNA in two ponds following introduction and removal of two rare fish species. When two rare species were introduced and kept at a fixed location in the ponds, eDNA concentration (i.e., proportional read counts abundance) of the introduced species typically peaks after two days. Thereafter, it gradually declined and stabilised after six days. These findings are supported by the highest community dissimilarity of different sampling positions is observed on the second day after introduction, which then gradually decreased over time. On the sixth day, there was no longer a significant difference in community dissimilarity between sampling days. The introduced species were no longer detected at any sampling positions 48 hrs after removal from the ponds. The eDNA signal and detection probability of the introduced species were strongest near the keepnets, resulting in the highest community variance of different sampling events at this position. Thereafter, the eDNA signal significantly decreased with increasing distance, although the signal increased slightly again at 85 m position away from keepnets. Collectively, these findings reveal that eDNA distribution in lentic ecosystems is highly localised in space and time, which adding to the growing weight of evidence that eDNA signal provides a good approximation of the presence and distribution of species in ponds. Moreover,

[3] This chapter will be published as ***Li, J.***, Lawson Handley, L.J., Harper, L.R., Brys, R., Watson, H.V. & Hänfling, B. Limited dispersion and quick degradation of environmental DNA in fish ponds inferred by metabarcoding.

eDNA metabarcoding is a powerful tool for detection of rare species alongside more abundant species due to use of generic PCR primers and can enable monitoring of spatial and temporal community variance.

4.1 Introduction

Environmental DNA (eDNA) analysis has emerged as a powerful tool in biological conservation for rapid and effective biodiversity assessment. This tool relies on the detection of genetic material that organisms leave behind in their environment (Taberlet et al. 2012a; Thomsen & Willerslev 2015). An important application of this method is discovery, surveillance and monitoring of invasive, rare, or threatened species, especially in environments where organisms or communities are difficult to observe, such as aquatic environments (reviewed in Rees et al. 2014; Lawson Handley 2015; Barnes & Turner 2016; Deiner et al. 2017a). Several studies have found positive relationships between eDNA concentration and organism density in aquatic ecosystems (e.g., Takahara et al. 2012; Pilliod et al. 2013; Li et al. 2018a). However, in freshwater ecosystems, the detection probability of eDNA is highly dependent on its characteristics, including the origin (physiological sources), state (physical forms), transport (physical movement), and fate (degradation) of eDNA molecules (reviewed in Barnes & Turner 2016). Consequently, understanding of eDNA characteristics is crucial to improve eDNA sampling designs and ensure the accuracy and reliability of eDNA biodiversity assessments (Goldberg et al. 2018).

Organisms shed DNA into their environment as sloughed tissues (e.g., faeces, urine, moulting, mucus or gametes) and whole cells, which then break down and release DNA (reviewed in Lawson Handley 2015; Thomsen & Willerslev 2015). Studies have demonstrated that eDNA production rates can be highly variable among species in aquatic ecosystems (Goldberg et al. 2011; Thomsen et al. 2012b; Sassoubre et al. 2016), and several factors can influence the amount of genetic material released by organisms into water, including biomass, life stage, breeding, and feeding behaviour (Maruyama et al. 2014; Pilliod et al. 2014; Klymus et al. 2015; Tillotson et al. 2018).

Once released into the environment, eDNA is transported away from organisms and begins to degrade. To better understand the distribution of eDNA in relation to species distribution, investigations have begun to examine how this complex DNA signal is transported horizontally (i.e., downstream) and vertically (i.e., settling) in aquatic environments. In lotic ecosystems, including rivers and streams, eDNA studies on horizontal transport produced variable results, where eDNA is transported metres to kilometres depending on stream discharge (Deiner & Altermatt 2014; Pilliod et al. 2014; Jane et al. 2015; Jerde et al. 2016). In contrast to lotic ecosystems, the natural hydrology of lentic ecosystems, such as lakes and ponds, may be less complex. In still water, eDNA has been shown to accumulate nearby to target organisms, with detection rate and eDNA concentration dropping off dramatically less than a few metres from the target organisms (Takahara et al. 2012; Eichmiller et al. 2014; Dunker et al. 2016). Additionally, eDNA detection may provide a more contemporary picture of species distribution, as transport is less important in lentic ecosystems. This may allow for greater settling of eDNA in sediment at the location where DNA shedding took place. Indeed, eDNA concentration of targeted fish is higher in sediment than in surface water of lentic systems (Eichmiller et al. 2014; Turner et al. 2015). Therefore, sedimentary eDNA can also result in false positive detections and affect inferences made regarding the current presence of a species.

eDNA degradation can also reduce the detectability of species over time. The rate of degradation in water can range from hours to weeks, depending on the ecosystem, target species, and eDNA capture method in question (Dejean et al. 2011; Takahara et al. 2012; Thomsen et al. 2012a; Thomsen et al. 2012b; Goldberg et al. 2013; Balasingham et al. 2017; Baker et al. 2018). Additionally, environmental conditions (e.g., chlorophyll α, natural inhibitors, microbial activity, biochemical oxygen demand “BOD”, temperature, pH, and ultraviolet B “UV-B” radiation) play an integral role in eDNA degradation rates (Barnes et al. 2014; Pilliod et al. 2014; Strickler et al. 2015; Lance et al. 2017; Stoeckle et al. 2017; Seymour et al. 2018).

The complex nature of eDNA has led to a new branch of eDNA research that aims to disentangle the factors influencing its ecology, such as the distribution of eDNA across both spatial and temporal scales (Spear et al. 2015; Wilcox et al. 2016; Tillotson et al. 2018), and a mechanistic understanding of eDNA ecology in relation to transport, retention (i.e., deposition or capture by sediment) and subsequent resuspension (Jane et al. 2015; Jerde et al. 2016; Shogren et al. 2016; Shogren et al. 2017).

The majority of the aforementioned studies have targeted single species using real-time quantitative PCR (qPCR) or droplet digital PCR (ddPCR) to investigate eDNA characteristics. Recently, eDNA metabarcoding, which combines PCR amplification with High-Throughput Sequencing (HTS), has emerged as a powerful, efficient, and economical tool for biodiversity assessment and monitoring of entire aquatic communities (e.g., Deiner et al. 2016; Hänfling et al. 2016; Port et al. 2016; Valentini et al. 2016). This tool removes the need to select target organisms *a priori* with the use of generic PCR primers that amplify multiple taxa, thus facilitating detection of invasive or threatened species when conducting holistic biodiversity assessment and routine freshwater monitoring (Thomsen & Willerslev 2015). Encouragingly, Harper et al. (2018) demonstrated that great crested newt *Triturus cristatus* detection via metabarcoding with no threshold is equivalent to qPCR with a stringent detection threshold. eDNA metabarcoding has also been applied to large-scale investigations of spatial or temporal variation in marine and freshwater communities, with some studies indicating that communities can be distinguished from 100 m to 2 km due to stream discharge or tidal patterns (Civade et al. 2016; Port et al. 2016; O'Donnell et al. 2017; Kelly et al. 2018; Li et al. 2018b).

In this study, I capitalise on the diagnostic power of eDNA metabarcoding. I explore the spatial and temporal distribution of fish communities in two aquaculture ponds and evaluate the detection sensitivity of this tool for low-density species alongside highly abundant species. Two primary objectives are investigated. Firstly, the shedding and decay rates of eDNA in fish ponds are explored, following the introduction and removal of two rare species at a fixed location. Secondly, this study examines the spatial distribution of fish communities after rare species introduction and removal. I expect that eDNA would be shed and diffused away from its source (the rare and introduced species), and this increased movement of eDNA particles would homogenise β-diversity in terms of community similarity, thus eroding the distance-decay relationship of eDNA. Theoretically, the eDNA signal of introduced species will increase until plateau after placing the keepnets into the ponds. After removal the keepnets, the eDNA signal of introduced species will decrease until vanish. With the sampling distance increasing to the keepnets, the eDNA signal will decrease (Figure 4.1). The results of this research are critical to understand the characteristics of eDNA in ponds including production, degradation and transport, and to inform effective sampling strategies.

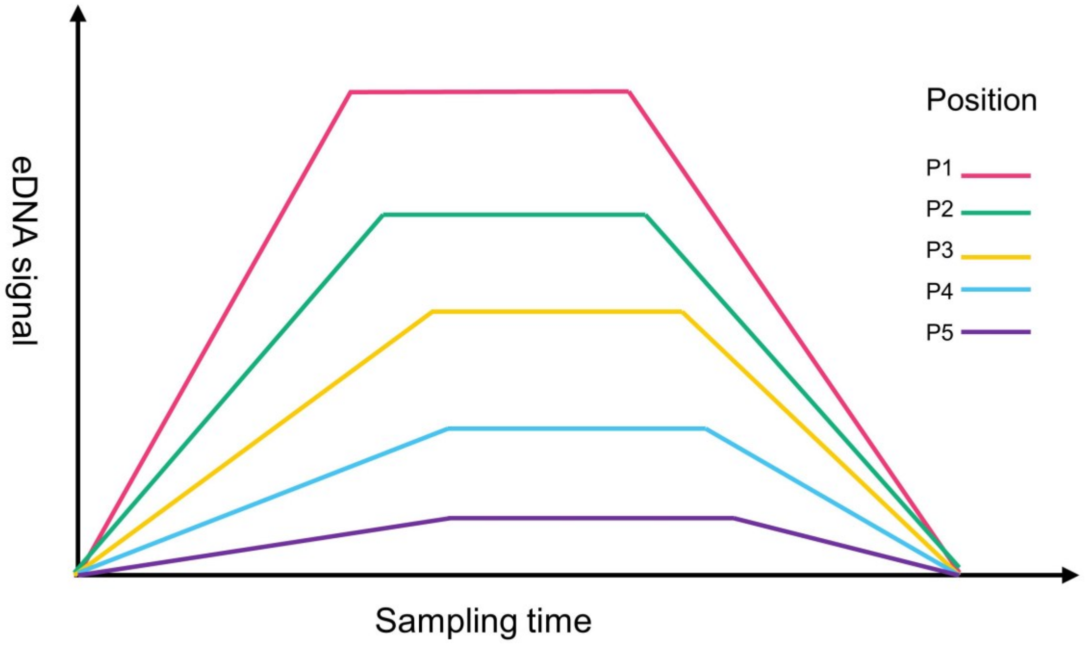


Figure 4.1 Expected temporal and spatial pattern of eDNA signal of the introduced species in keepnets following the introduction and removal at a fixed location. The linear distance of each sampling position to keepnets with the introduced species is 0 m (P1), 28 m (P2), 56 m (P3), 85 m (P4), and 104 m (P5).

4.2 Materials and methods

4.2.1 Study site and water sampling

This study was carried out at two artificially stocked ponds with a high fish density and in a turbid and eutrophic condition. The two ponds (E1 and E4) are located at the National Coarse Fish Rearing Unit (NCFRU, Calverton, Nottingham, UK), run by the UK Environment Agency. The ponds are groundwater fed with no inflow from surface water bodies. The dimension of each pond is approximately 60 m × 85 m, with an average depth of 1.5 m. In each pond, there were two feeding devices with timers that release food hourly, and two automatic aerators near the feeding devices to increase the dissolved oxygen (DO) profile. The automatic aerators also created flowing conditions for fish to feed in and to help build the right kind of muscle which they will need for life in the wild (Figure 4.2). Generally, these ponds are used to rear approximately one-year-old common British coarse fish before they are used in stocking programmes for conservation purposes or recreational fishing.

The experiment was conducted from 19th September to 3rd October 2016. DO and temperature were monitored daily in each pond during the entire sampling period. DO concentration and temperature were 8.4 ± 1.3 mg L^{-1} and 15.6 ± 1.4 °C in pond E1, and 7.1 ± 1.4 mg L^{-1} and 16.0 ± 1.3 °C in pond E4. Stocked fish in both ponds were measured and weighed before stocking on 16th June 2016 and after harvesting on 18th November 2016. Fish abundance and biomass at time of water sampling in September 2016 were estimated, assuming that the death and growth curves of these fish are linear (Appendix S4.1 Figure S4.1 & Figure S4.2). The fish stock information in September 2016 is shown in Table 4.1. On 19th September at 15:00 (hereafter referred to as “D0”), an hour prior to introduction of additional fish species, one 2 L water sample was taken just below the pond surface using sterile Gosselin™ HDPE plastic bottles (Fisher Scientific) at each of the five sampling positions (hereafter referred to as “P1–P5”) spread over 104 m, to confirm fish community composition and check for potential contamination from aberrant species. Briefly, four sampling positions (P1–P4) were distributed equidistant on the same shoreline of the pond, whereas P5 was on the catercorner of P1 (Figure 4.2). After sampling on D0, four new keepnets containing 25 individuals each of the introduced species were placed in P1 of each pond (Figure 4.2). In pond E1, the introduced species were chub *Squalius cephalus* (26.0 ± 1.8 g) and rudd *Scardinius erythrophthalmus* (21.8 ± 1.5 g), whereas rudd (22.4 ± 1.6 g) and dace *Leuciscus leuciscus* (19.8 ± 1.5 g) were introduced to pond E4. After fish introduction, five 2 L water samples were collected at 10:00 on days 2, 4, 6 and 8 (hereafter referred to as “D2–D8”, introductory stage) at each position (P1–P5) in each pond. On D8, the keepnets with introduced species were removed after water sampling on that day was completed. No fish died in the keepnets. After removal of the keepnets, water samples were collected in the same manner on days 10, 12 and 14 (hereafter referred to as “D10–D14”, removal stage) in order to estimate eDNA decay of the introduced species once removed from the pond. In each pond, forty samples were taken over the course of the experiment (80 samples in total). The introduced species were weighed after removal from ponds, and then released back into indoor tanks at the NCFRU. All animal research was approved by the University of Hull’s Faculty of Science Ethics Committee (Approval #U093).

Table 4.1 Fish stock information on two experimental ponds at the National Coarse Fish Rearing Unit.

Pond	**Species**			**September 2016**	
	Scientific name	Common name	Code	Abundance	Biomass (kg)
E1	*Barbus barbus*	Barbel	BAR	7245	267.99
E1	*Abramis brama*	Bream	BRE	6449	152.33
E1	*Carassius carassius*	Crucian carp	CAR	2309	80.44
E1	*Squalius cephalus*†	Chub	CHU	50	1.30
E1	*Leuciscus leuciscus*	Dace	DAC	18544	123.96
E1	*Rutilus rutilus*	Roach	ROA	3452	44.64
E1	*Scardinius erythrophthalmus*†	Rudd	RUD	50	1.09
E1	*Tinca tinca*	Tench	TEN	3605	59.09
E4	*Barbus barbus*	Barbel	BAR	4230	165.07
E4	*Abramis brama*	Bream	BRE	1130	32.33
E4	*Carassius carassius*	Crucian carp	CAR	1766	79.25
E4	*Squalius cephalus*	Chub	CHU	16395	492.01
E4	*Leuciscus leuciscus*†	Dace	DAC	50	0.99
E4	*Rutilus rutilus*	Roach	ROA	24732	355.53
E4	*Scardinius erythrophthalmus*†	Rudd	RUD	50	1.12
E4	*Tinca tinca*	Tench	TEN	645	9.28

Notes: "†" indicates the rare species introduced to each pond for the purposes of this study. Abundance represents number of individuals. Full scientific, common names and three letter codes used in figures and tables are given.

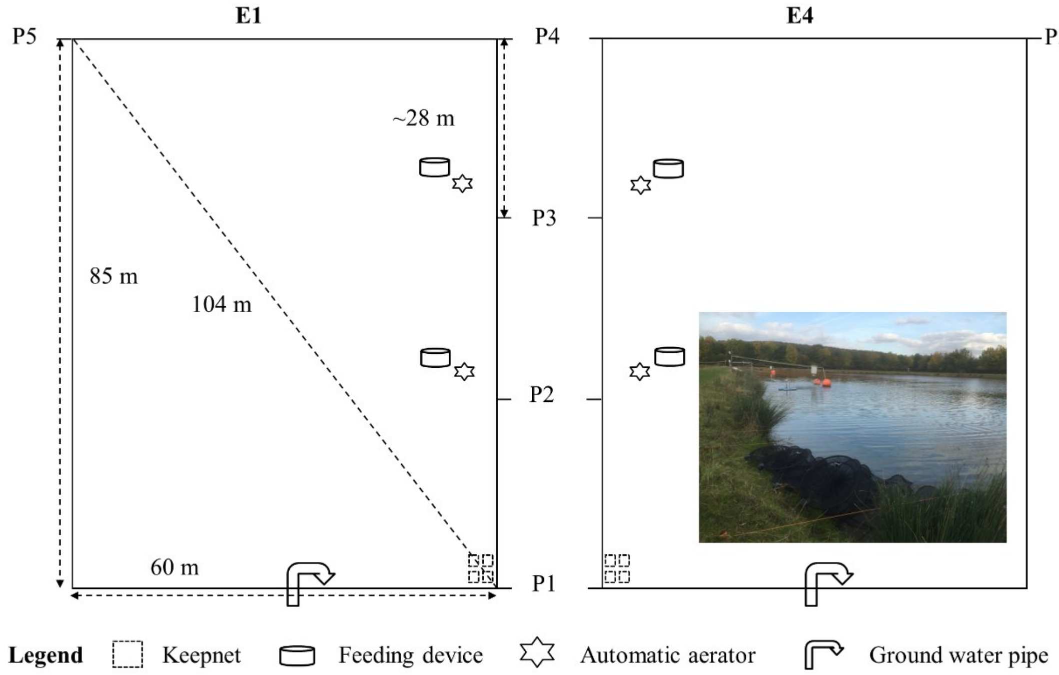


Figure 4.2 Schematic of sampling strategy at the National Coarse Fish Rearing Unit. The linear distance of each sampling position to keepnets with the introduced species is 0 m (P1), 28 m (P2), 56 m (P3), 85 m (P4), and 104 m (P5).

4.2.2 eDNA capture and extraction

After each sampling event, all water samples were filtered immediately in a laboratory at NCFRU that was decontaminated before filtration by bleaching (10% v/v commercial bleach) floors and surfaces. Three filtration replicates (300 mL) were subsampled from each 2 L water sample collected at every sampling position. All filtration replicates were filtered through sterile 0.8 µm mixed cellulose acetate and nitrate (MCE) filters, 47 mm diameter (Whatman) using Nalgene filtration units in combination with a vacuum pump (15–20 in. Hg; Pall Corporation). Our previous study demonstrated that 0.8 µm is the optimal membrane filter pore size for turbid, eutrophic, and high fish density ponds, and achieves a good balance between rapid filtration time and the probability of species detection via metabarcoding (see more detail in Chapter 3; Li et al. 2018a).

To reduce cross-contamination, samples from the same pond were filtered in the same batch and in order of collection from P1 to P5. The same filtration unit was used for all three filtration replicates of each sample. The filtration units were soaked in 10% v/v commercial bleach solution 10 mins and 5% v/v microsol detergent (Anachem, UK) 5 mins, and then rinsed thoroughly with deionised water after each round of filtration to prevent cross-contamination. One filtration blank (300 mL deionised water) was processed for each pond on every day of filtration to monitor contamination risk. After filtration, all membrane filters were placed into 50 mm sterile petri dishes (Fisher Scientific) using sterile tweezers, sealed with Parafilm® (Bemis Company, Inc.), and stored at –20°C until DNA extraction. DNA extraction was carried out using the PowerWater® DNA Isolation Kit (MoBio Laboratories Inc., now QIAGEN) following the manufacturer's protocol.

4.2.3 Library preparation and sequencing

Extracted DNA samples were amplified with vertebrate-specific primers (Riaz et al. 2011) that target a 106-bp fragment of the mitochondrial 12S rRNA region in fish, using a two-step PCR protocol for library preparation that implements a nested tagging approach (Kitson et al. 2019). Previous eDNA metabarcoding studies of marine mesocosms and coastal ecosystems showed that this fragment has a low false negative rate for bony fishes (Kelly et al. 2014; Port et al. 2016). We also previously tested this fragment *in situ* on three deep lakes in the Lake District, England, where metabarcoding results were compared to long-term data from established survey methods (Hänfling et al. 2016), and at NCFRU to investigate the impact of different filters on eDNA capture and quantification (see more detail in Chapter 3; Li et al. 2018a). Taken together, our previous findings demonstrated that this 106-bp fragment is highly suitable for eDNA metabarcoding of UK freshwater fish communities.

In the two-step library preparation protocol, the first PCR reactions were set up in a UV and bleach sterilised laminar flow hood in our dedicated eDNA laboratory at the University of Hull to minimise contamination risk. All filtration replicates (*N* = 240), together with 16 filtration and extraction blanks, 16 no-template controls (NTCs), and 16 single-template positive controls (STCs) were included in library construction (*N* = 288) for sequencing on an Illumina MiSeq. For the STCs, we used genomic DNA (0.08 ng uL^{-1}) of the Eastern happy (*Astatotilapia calliptera*), a cichlid from Lake Malawi that is not present in natural waters in UK.

The first PCR reaction was carried out in 25 μL volumes containing: 12.5 μL of 2 × MyTaq HS Red Mix (Bioline), 0.5 μM of each tagged primer, 2.5 μL of template DNA, and 7.5 μL of molecular grade water. Eight-strip PCR tubes with individually attached lids and mineral oil (Sigma-Aldrich) were used to reduce cross-contamination between samples. After PCR preparation, reaction tubes were brought to our PCR room for amplification, where all post-PCR work was carried out. Thermal cycling parameters were as follows: 98 °C for 5 min, 35 cycles of 98 °C for 10 sec, 58 °C for 20 sec, and 72 °C for 30 sec, followed by a final elongation step at 72 °C for 7 min. Three PCR technical replicates were performed for each sample, then pooled to minimise bias in individual PCRs. The indexed first PCR products of each sample were then pooled according to sampling event and pond, and 100 μL of pooled products cleaned using the Mag-Bind® RXNPure Plus Kit (Omega Bio-tek) using a dual bead-based size selection protocol (Bronner et al. 2014). Ratios used for size selection were 0.9× and 0.15× magnetic beads to PCR product.

The second PCR reactions were carried out in 50 µL volumes containing: 25 µL of 2 × MyTaq HS Red Mix (Bioline), 1.0 µM of each tagged primer, 5 µL of template DNA and 15 µL of molecular grade water. Reactions without template DNA were prepared in our dedicated eDNA laboratory, and first PCR products added later in the PCR room. Thermal cycling parameters were as follows: initial denaturation at 95 °C for 3 min, followed by 10 cycles of 98 °C for 20 s, and 72 °C 1 min, with a final extension of 72 °C for 5 min. The second PCR products (50 µL) were cleaned using the Mag-Bind® RXNPure Plus Kit (Omega Bio-tek) according to a dual bead-based size selection protocol (Bronner et al. 2014). Ratios used for size selection were 0.7× and 0.15× magnetic beads to PCR product. The cleaned second PCR products were normalised according to sample number and concentration across sampling events and ponds based on the Qubit™ 3.0 fluorometer results using a Qubit™ dsDNA HS Assay Kit (Invitrogen), then pooled. The final library concentration was quantified by qPCR using the NEBNext® Library Quant Kit (New England Biolabs). The pooled, quantified library was adjusted to 4 nM and denatured following the Illumina MiSeq library denaturation and dilution guide. To improve clustering during initial sequencing, the denatured library (13 pM) was mixed with 10% PhiX genomic control. The library was sequenced on an Illumina MiSeq platform using the MiSeq reagent kit v2 (2 × 250 cycles) at the University of Hull.

4.2.4 Data analysis

4.2.4.1 Bioinformatics analysis

Raw read data from the Illumina MiSeq have been submitted to NCBI (BioProject: PRJNA486650; BioSample accession: SAMN09859568–SAMN09859583; Sequence Read Archive accessions: SRR7716776–SRR7716791). Bioinformatics analysis was implemented using a custom, reproducible pipeline for metabarcoding data (metaBEAT v0.97.10) (see more detail in Appendix S2.1 Section S2.1.1; Hänfling et al. 2016) with a custom reference database described in Chapter 2. Sequences for which the best BLAST hit had a bit score below 80 or had less than 100% identity to any sequence in the curated database were considered non-target sequences. To assure full reproducibility of our bioinformatics analysis, the custom 12S reference database and the Jupyter notebook for data processing have been deposited in a dedicated GitHub repository (https://github.com/HullUni-bioinformatics/Li_et_al_2019_eDNA_dynamic). The Jupyter notebook also performs demultiplexing of the indexed barcodes added in the first PCR reactions.

4.2.4.2 Criteria for reducing false positives and quality control

Filtered data were summarised as the number of sequence reads per species (hereon referred to as read counts) for downstream analyses (Appendix S4.2). After bioinformatics analysis, the low-frequency noise threshold (proportion of STC species read counts in the real sample) was set to 0.002 to filter out high-quality annotated reads that passed the previous filtering steps and had high-confidence BLAST matches, but may have resulted from contamination during the library construction process or sequencing (De Barba et al. 2014; Hänfling et al. 2016; Port et al. 2016).

4.2.4.3 Statistical analyses

All statistical analyses were performed in R v3.5.0 (R_Core_Team 2018), and graphs were plotted using GGPLOT2 v2.2.1 (Wickham & Chang 2016). The sequence read counts of different filtration replicates (*N* = 3) were averaged to provide a single read count for each sampling position unless otherwise specified. The fish community of each sampling position was standardised to proportional abundance (i.e., number of read counts per species relative to total number of read counts in that sample) using the "total" method with the function *decostand* in VEGAN v2.4-4 (Oksanen et al. 2017). To evaluate spatial and temporal species turnover between eDNA communities, the observed variation in distance measured as Bray-Curtis dissimilarity among sampling events and positions were apportioned using permutational multivariate analysis of variance (PERMANOVA) with the function *adonis* in VEGAN v2.4-4 (Oksanen et al. 2017). To determine the relationship between β-diversity in Bray-Curtis distance matrices of different sampling days (D0–D14) and the geographic distance matrix of different sampling positions (P1–P5), the Mantel correlations were performed with the function *mantel.rtest* of ADE4 v1.7-11 (Stéphane et al. 2018). To examine temporal and spatial variance in fish communities after the introduction and removal of introduced species, pairwise Bray-Curtis dissimilarities were calculated using the function *vegdist* in VEGAN v2.4-4 (Oksanen et al. 2017), and Kruskal-Wallis one-way ANOVA with Dunn's test using Bonferroni adjustment conducted to test for differences in Bray-Curtis dissimilarity between different sampling days and positions. The statistical significance level of this study is set at 0.05. The full R script is available on the GitHub repository (https://github.com/HullUni-bioinformatics/Li_et_al_2019_eDNA_dynamic/tree/master/R_script).

4.3 Results

The library generated 16.99 million reads with 13.21 million reads passing filter including 10.94% PhiX control. Following quality filtering and removal of chimeric sequences, the average read count per sample (excluding controls) was 14,441. After

BLAST searches for taxonomic assignment, 51.50% ± 10.87% reads in each sample were assigned to fish.

4.3.1 Species detection in the background communities

All stocked species were detected over the course of the experiment in ponds E1 and E4. In pond E1, the stocked species were common bream *Abramis brama*, barbel *Barbus barbus*, crucian carp *Carassius carassius*, dace, roach *Rutilus rutilus* and tench *Tinca tinca*. In pond E4, the stocked species were common bream, barbel, crucian carp, chub, roach, and tench (Figure 4.3). Moreover, apart from tench in pond E4, stocked species were detected across all sampling positions (Figure 4.3; Appendix S4.1 Table S4.1). Tench was the rarest stocked species in pond E4 (proportional individual and biomass was 1.32% and 0.82%, respectively; Figure 4.3b; Table 3.1) which may explain imperfect species detection.

4.3.2 Spatio-temporal detection of introduced species

The introduced species were not detected in samples taken prior to species introduction (i.e., D0), or in process controls (filtration, extraction and NTCs). Therefore, the introduced species were not present in the environment or as laboratory contaminants before the experiment began. After introduction of rudd and chub into pond E1, rudd were detected across the entire period the species were present (D2–D8), whereas chub were not recovered on D6 in pond E1. In pond E4, both the introduced species, rudd and dace, were identified across the entire period the species were present (Figure 4.3). In terms of sampling position, the eDNA signal of the introduced species was strongest close to the keepnets (P1) and decreased with increasing distance from this location (Figure 4.3). In pond E1, both introduced species were detected until P4

(85 m from the keepnets), but not at the catercorner of the keepnets (P5, 104 m away from the keepnets). In contrast, in pond E4, both introduced species could be detected at P5 on D6 (Figure 4.4). The detection probability of the introduced species at P1 across both ponds (Appendix S4.1 Table S4.2, 0.88 ± 0.13) was significantly higher than other sampling positions during the entire period the species were present (Appendix S4.1 Table S4.2, ANOVA, p consistently < 0.05). Moreover, eDNA concentration (i.e., proportional read counts abundance) of introduced species was highest on D2 at the original source (P1) in both ponds (Figure 4.4a, f). Thereafter, eDNA concentration decreased gradually and reached equilibrium (i.e., the production rate equal to degradation rate) on D6, with a slight increase on D8 (Figure 4.4a, f). There was also some variation in eDNA concentration among species that was unrelated to fish density. For instance, the eDNA concentration of rudd was higher than chub in pond E1 but lower than dace in pond E4 (Figure 4.4), even though the biomass of rudd was lower than chub in pond E1 and higher than dace in pond E4 (Table 4.1). Notably, after the introduced species had been removed for 48 hrs (D8–D10), they were no longer detectable at any position in both ponds (Figure 4.3 & Figure 4.4).

Compared to the expected temporal and spatial pattern of eDNA signal of the introduced species in keepnets following the introduction and removal at a fixed location (Figure 4.1), the escalating and plateau stage of the introduced species was not observed following the introduction of rare species in this study instead of fluctuation between D2 and D8. Moreover, the signal of the introduced species disappeared quickly instead of gradually reduced (Figure 4.4).

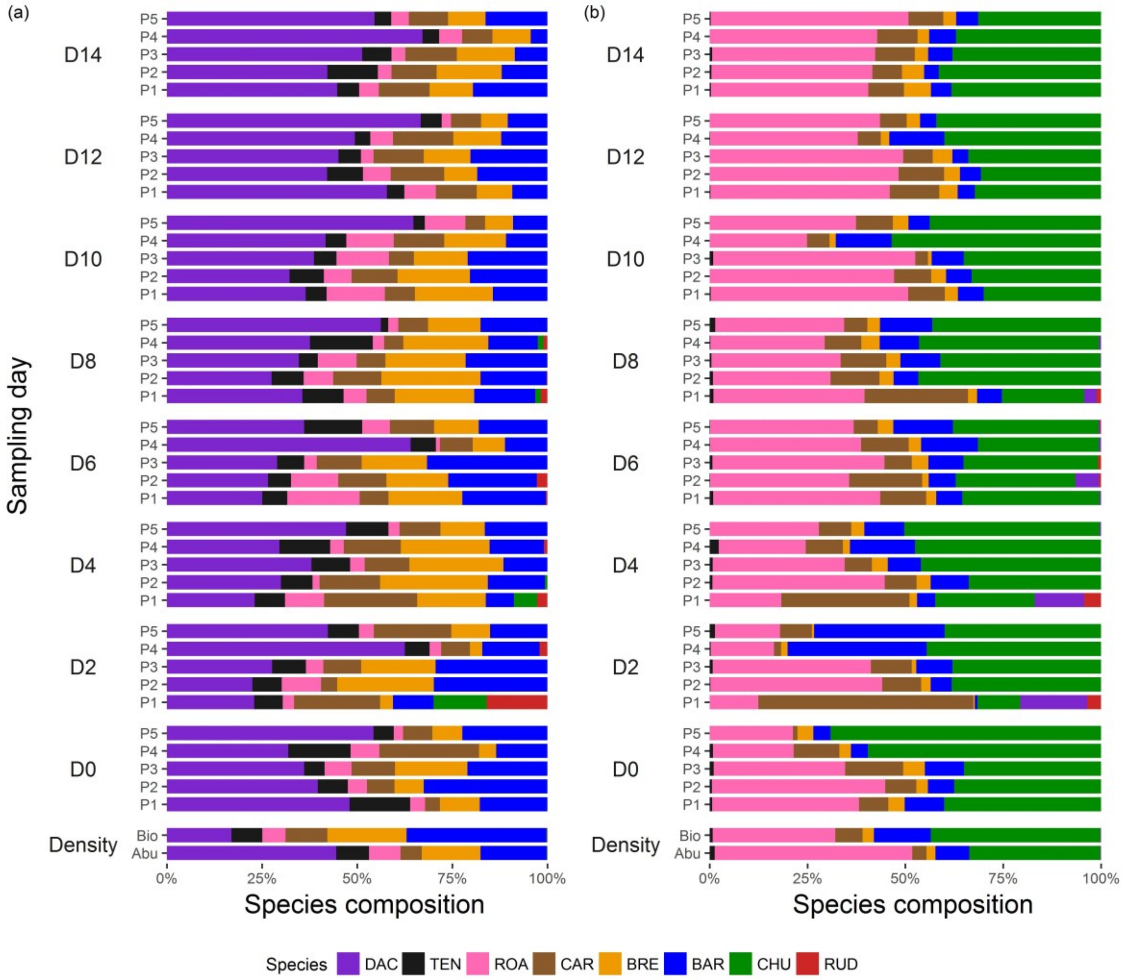


Figure 4.3 Species composition of averaged read counts (number of replicates = 3) for five sampling positions over 14 days in ponds (a) E1 and (b) E4. “Bio” and “Abu” refers to fish biomass and abundance density respectively, calculated based on Table 4.1. Species three letter codes correspond to species given in Table 4.1. After control samples were taken on D0, the rare species were introduced and samples were taken on days 2, 4, 6, 8, 10, 12, and 14 (D2–D14) from the five sampling positions (P1–P5). The introduced species were removed on D8 after sampling. The linear distance of each sampling position to keepnets of introduced species is 0 m (P1), 28 m (P2), 56 m (P3), 85 m (P4), and 104 m (P5).

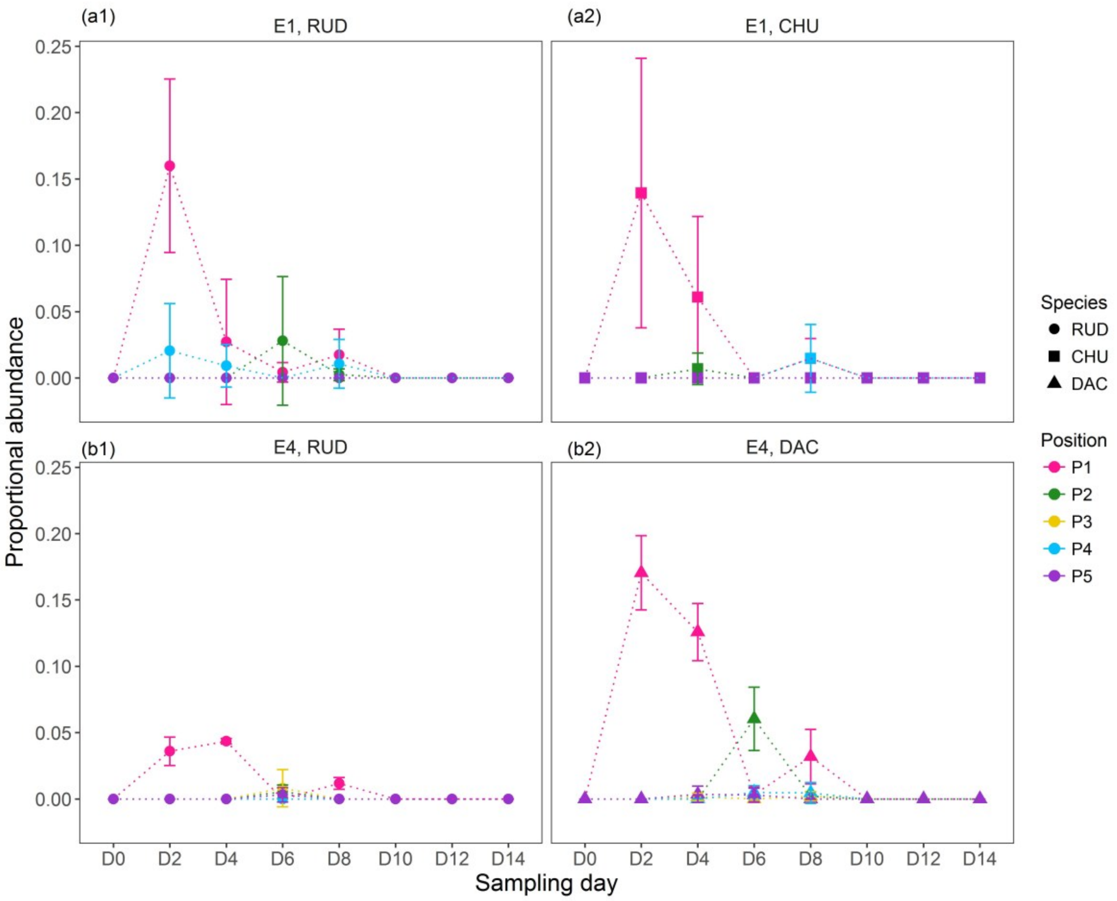
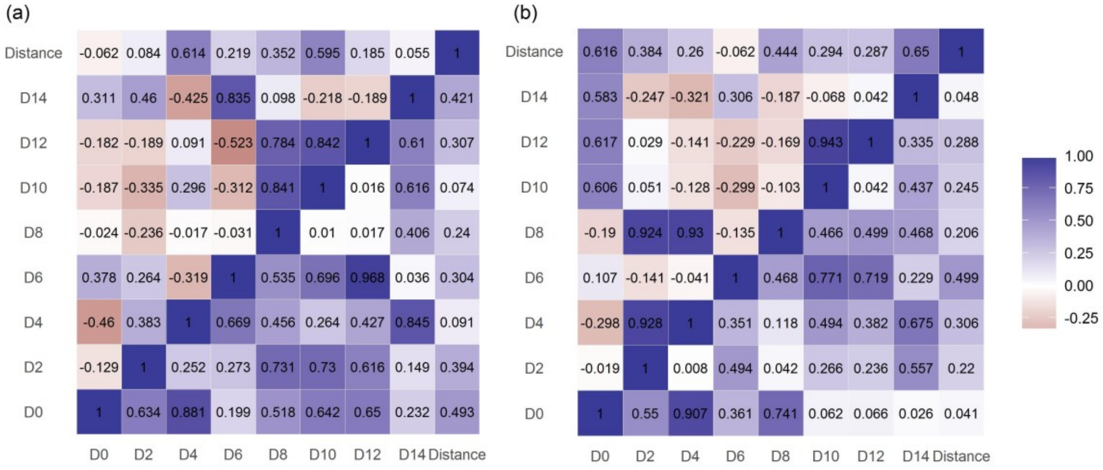


Figure 4.4 Temporal change over 14 days (D0–D14) in averaged proportional abundance of introducted species across five sampling positions (P1–P5) in ponds (a1-a2) E1 and (b1-b2) E4. The standard error bars represent three filtration replicates per sample. Species three letter codes correspond to species given in Table 4.1. The different sampling stages and linear distance between sampling positions are described in Figure 4.3.

4.3.3 Community variance in Bray-Curtis dissimilarity

On the whole, sampling day and position had significant effects on community variance, using Bray-Curtis dissimilarity for ponds E1 (PERMANOVA; sampling days df = 7, R^2 = 0.296, p = 0.002; positions df = 4, R^2 = 0.235, p = 0.002) and E4 (PERMANOVA; sampling days df = 7, R^2 = 0.241, p = 0.013; positions df = 4, R^2 = 0.271, p = 0.001). Specifically, the estimates of community dissimilarity for different sampling positions between different sampling days were not correlated with geographic distance, except D0 in pond E4 (Figure 4.5b, p = 0.041, Mantel's r = 0.616). Moreover, there were significant correlations of community dissimilarity between D8, D10 and D12, D6 and D14 in pond E1. Significant correlations of community dissimilarity were observed between D0 and D14, D2 and D4, D2 and D8, D10 and D12 in pond E4. All the r statistics and p-values as determined by the Mantel test are shown in Figure 4.5.

Overall, fish communities varied in Bray-Curtis dissimilarity before introduction on D0, introduction from D2 to D8, and removal from D10 to D14 (Appendix S4.1 Figure S4.3). The Bray-Curtis dissimilarity of the removal stage was significantly lower than the introductory stage in both ponds E1 and E4 (Appendix S4.1 Figure S4.3; Dunn's test: E1 z = 3.71, $p < 0.05$; E4 z = 2.98, $p < 0.05$). In pond E4, community dissimilarity of the removal stage was also significantly lower than before the introduction of species (Appendix S4.1 Figure S4.3b; Dunn's test: z = 2.45, $p < 0.05$). More specifically, after the introduction of rare species, the highest community dissimilarities of different sampling positions were observed on D2 and decreased over time in both ponds. There was no significant difference between sampling days during D4–D14 and D6–D14 in ponds E1 and E4, respectively (Figure 4.6a1, b1). In terms of sampling position, the highest community variances of different sampling days occurred close to the keepnets (P1) in both ponds (Figure 4.6a2, b2), and the community dissimilarity significantly declined with increasing distance from P1 to P3. However, communities were more dissimilar at P4 compared to P3, with a significant increase in Bray-Curtis dissimilarity values (Figure 4.6a2, b2; Dunn's test: E1 z = 2.92, p < 0.05; E4 z = 2.95, p < 0.05). In pond E1, there was a significant reduction in community dissimilarity at P5 compared to P4 (Figure 4.6a2; Dunn's test: z = 2.83, $p < 0.05$), whereas in pond E4, there was no significant difference in community dissimilarity between P4 and P5 (Figure 4.6b2).



(a)

Distance	-0.062	0.084	0.614	0.219	0.352	0.595	0.185	0.055	1
D14	0.311	0.46	-0.425	0.835	0.098	-0.218	-0.189	1	0.421
D12	-0.182	-0.189	0.091	-0.523	0.784	0.842	1	0.61	0.307
D10	-0.187	-0.335	0.296	-0.312	0.841	1	0.016	0.616	0.074
D8	-0.024	-0.236	-0.017	-0.031	1	0.01	0.017	0.406	0.24
D6	0.378	0.264	-0.319	1	0.535	0.696	0.968	0.036	0.304
D4	-0.46	0.383	1	0.669	0.456	0.264	0.427	0.845	0.091
D2	-0.129	1	0.252	0.273	0.731	0.73	0.616	0.149	0.394
D0	1	0.634	0.881	0.199	0.518	0.642	0.65	0.232	0.493
	D0	D2	D4	D6	D8	D10	D12	D14	Distance

(b)

Distance	0.616	0.384	0.26	-0.062	0.444	0.294	0.287	0.65	1
D14	0.583	-0.247	-0.321	0.306	-0.187	-0.068	0.042	1	0.048
D12	0.617	0.029	-0.141	-0.229	-0.169	0.943	1	0.335	0.288
D10	0.606	0.051	-0.128	-0.299	-0.103	1	0.042	0.437	0.245
D8	-0.19	0.924	0.93	-0.135	1	0.466	0.499	0.468	0.206
D6	0.107	-0.141	-0.041	1	0.468	0.771	0.719	0.229	0.499
D4	-0.298	0.928	1	0.351	0.118	0.494	0.382	0.675	0.306
D2	-0.019	1	0.008	0.494	0.042	0.266	0.236	0.557	0.22
D0	1	0.55	0.907	0.361	0.741	0.062	0.066	0.026	0.041
	D0	D2	D4	D6	D8	D10	D12	D14	Distance

Figure 4.5 Heatmap of community correlation as determined by the Mantel test between Bray-Curtis distance matrices of different sampling days (D0–D14) and the geographic distance matrix of different sampling positions (P1–P5) in ponds (a) E1 and (b) E4. "Distance" refers to the distance matrix based on the linear distance between different sampling positions. The upper triangular and lower triangular is Mantel *r* statistics and *p*-values, respectively. The different sampling stages and linear distance between sampling positions are described in Figure 4.3.

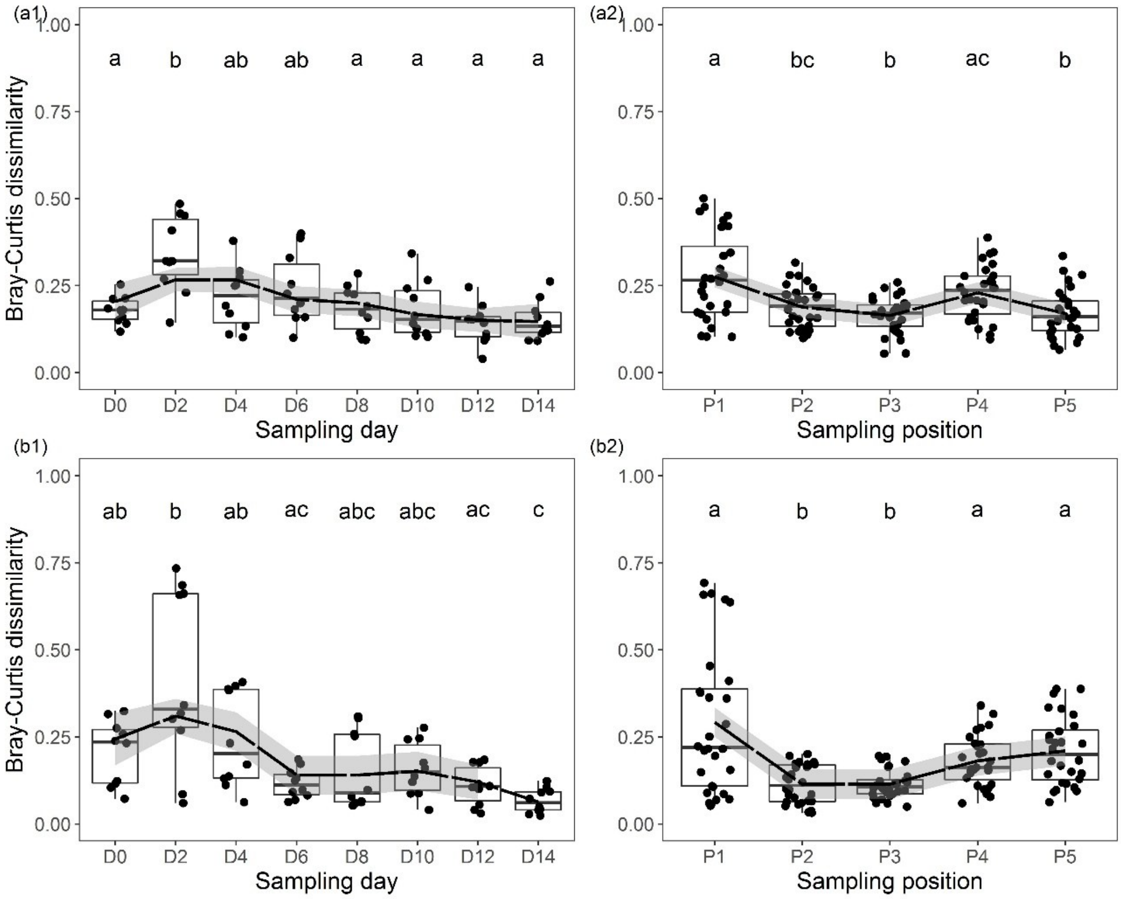


Figure 4.6 Temporal change (D0–D14) in community dissimilarity of the five sampling positions (P1–P5) in ponds (a1) E1 and (b1) E4, where each point represents the Bray-Curtis dissimilarity of two different sampling positions on the same sampling day. Spatial change (P1–P5) in community dissimilarity of the eight sampling days (D0–D14) in ponds (a2) E1 and (b2) E4, where each point represents the Bray-Curtis dissimilarity of two different sampling days at the same sampling position. Sampling days or positions that differ significantly ($p < 0.05$) from one another are indicated with different letters in each boxplot. Dashed lines represent the fit of non-linear regressions, and grey shaded areas denote the 95% confidence interval as calculated using the standard error. The different sampling stages and linear distance between sampling positions are described in Figure 4.3.

4.4 Discussion

Spatial heterogeneity of eDNA distribution has been reported in lentic ecosystems (Takahara et al. 2012; Eichmiller et al. 2014; Hänfling et al. 2016; Lawson Handley et al. 2019). Therefore, an understanding of the spatial heterogeneity of eDNA distribution is critical to the design of effective sampling protocols for accurate species detection and abundance estimates in lentic ecosystems, especially in order to detect rare or invasive species. To my knowledge, this study is the first that uses metabarcoding to investigate the spatial and temporal community variances in ponds to understand eDNA characteristics in these systems, including production, degradation and transport following the introduction and removal of rare species.

4.4.1 eDNA production

The eDNA concentration of the introduced species peaks on D2 at the position closest to the keepnets (P1). Thereafter, eDNA concentration of these introduced species declines gradually over time and stabilises by D6 in both ponds. Consequently, the highest community dissimilarity of different sampling positions is observed on D2 and decreases over time in both ponds. The increase in eDNA concentration of the introduced fish after 43 hrs may have been caused by increased eDNA shedding rates as a result of fish being stressed by handling, as observed in other studies (Takahara et al. 2012; Maruyama et al. 2014; Klymus et al. 2015; Sassoubre et al. 2016). Considering the degradation rate of eDNA is less than 48 hrs (see more detail in Section 4.4.2), eDNA concentration may have declined after D2 due to fish acclimation to the keepnets and reduced activity, resulting in less eDNA release. By D6, the rate of eDNA release from the two introduced species seems to reach equilibrium with the rate of eDNA degradation. These patterns are consistent with previous qPCR studies that targeted single species and investigated eDNA production and degradation, including eDNA shedding rate of common carp *Cyprinus carpio* in aquaria (Takahara et al. 2012), different developmental stages of bluegill sunfish *Lepomis macrochirus* in aquaria (Maruyama et al. 2014) and three marine fish (Northern anchovy *Engraulis mordax*, Pacific sardine *Sardinops sagax* and Pacific chub mackerel *Scomber japonicas*) in seawater mesocosms (Sassoubre et al. 2016). However, eDNA concentration of two amphibian species (common spadefoot toad *Pelobates fuscus* and great crested newt) exhibit monotonic increases after introduction into aquaria, which may be the result of a longer sampling period over larger time intervals, i.e., weeks over two months or lower degradation rates in controlled environments (Thomsen et al. 2012b).

4.4.2 eDNA degradation

The detection rates of the introduced species declines with no detectable eDNA signal at any sampling position in both ponds approximately 48 hrs after removal. As a result, there is no significant difference in community dissimilarity of different sampling positions among the sampling days after removal of the introduced species. This observation is in agreement with other studies that documented no eDNA detection shortly after target species were removed from the water in which they occurred. For example, detection of European flounder *Platichthys flesus* or bluegill sunfish fails around 24 hours after removal from aquaria (Thomsen et al. 2012a; Maruyama et al. 2014), and 48 hrs after removal of Atlantic salmon *Salmo salar* from a river ecosystem (Balasingham et al. 2017). By contrast, other studies have reported slower eDNA degradation rates in controlled aquaria or mesocosms. For example, eDNA degrades beyond detection within a week for fish (Thomsen et al. 2012a; Barnes et al. 2014; Sassoubre et al. 2016), several weeks for amphibians (Dejean et al. 2011; Thomsen et al.

2012b), and a month for New Zealand mudsnail *Potamopyrgus antipodarum* (Goldberg et al. 2013), The wide variation observed in the aforementioned studies emphasises the role of the ecosystem and starting eDNA concentration (influenced by shedding rate) on eDNA persistence. The reason for wide variation in eDNA production rates among species is unconfirmed, but animal physiology is suggested to play a role, e.g., stress (Pilliod et al. 2014), breeding readiness (Spear et al. 2015), diet (Klymus et al. 2015), and metabolic rate (Maruyama et al. 2014). Moreover, eDNA is also found to decay faster in the field than in controlled conditions, which can be attributed to the complex effects of environmental conditions on eDNA persistence (Barnes et al. 2014; Pilliod et al. 2014; Strickler et al. 2015; Lance et al. 2017; Stoeckle et al. 2017; Seymour et al. 2018).

4.4.3 eDNA transport

Regarding horizontal transport of eDNA, the eDNA signal and detection probability of the introduced species is highest close to the keepnets (P1) and broadly decreases with increasing distance up to around 104 m from this point. This finding agrees with previous qPCR studies that reported a patchy distribution of eDNA in the lentic ecosystems, and drastic decline in detection probability and eDNA concentration less than a few hundred metres from the target organisms (Takahara et al. 2012; Eichmiller et al. 2014; Dunker et al. 2016). Moreover, all estimates of β-diversity (i.e., community dissimilarity) of different sampling positions between different sampling days and geographic distances are not linearly correlated, except D0 in pond E4, which indicates that geographic distance does not have a significant effect. This result would imply that the eDNA of stocked fish is well homogenised in the ponds, and the eDNA signal released by the introduced species is too low to influence the spatial distribution pattern of the entire fish community present in the ponds. This result is in agreement with Evans et al. (2017) who do not find a significant relationship between sample dissimilarity and geographic distance in a 22,000 m^2 surface-area reservoir in which fish distribution is relatively homogeneous. By contrast, Sato et al. (2017) indicated that geographic distances among sampling locations within lakes ranging in size from 84,000 m^2 to 2,219,000 m^2 have a significantly positive correlation with the abundance-based community dissimilarity index resulting from spatial heterogeneity of eDNA distribution.

In lotic ecosystems, stream discharge plays an important role in horizontal eDNA transport, and can result in eDNA of target species being transported meters to kilometres (Deiner & Altermatt 2014; Pilliod et al. 2014; Jane et al. 2015; Jerde et al. 2016). Furthermore, the spatial community variance observed in other eDNA studies indicated that β-diversity does not increase as a function of distance (up to 12 km) in a stream (Deiner et al. 2016), but does increase with distance in a highly dynamic marine habitat (O'Donnell et al. 2017). Li et al. (2018b) also observed that the β-diversity of fish communities based on Jaccard distance (i.e., incidence data) between sampling sites is correlated with the sampling distance along the stream.

In the small fish ponds sampled in this study, the community variance in eDNA distribution is highly localised in space. The cline of community variance over distance is consistent, where eDNA signal of the introduced species is strongest at the position closest to the keepnets (P1), followed by a reduction in strength from P1 to P3 and growth from P3 to P4. Furthermore, two introduced species are detected at P5 in pond E4, but not at P5 in pond E1. This may explain why there is no significant change in community dissimilarity between P4 and P5 in pond E4, but the community dissimilarity of P5 is significantly reduced from P4 in pond E1. Notably, there are feeding devices and automatic aerators near P2 and P3. Thus, I speculated that food released by feeding devices could attract fish and cause them to aggregate near positions P2 and P3, which would increase the detection of stocked fish and thus reduce the detection probabilities of the introduced species. On the other hand, the automatic aerators could have enhanced water mixing, bringing eDNA from the introduced species into the other corner of the pond (P4). Therefore, the growth trend in eDNA concentration of the introduced species from P3 to P4 in both ponds may be a consequence of anthropogenic interference.

4.5 Conclusions

This study has demonstrated that eDNA metabarcoding is a powerful tool for monitoring change in community structure across time and space. After eDNA is shed and transported away from its source, the increased movement of eDNA particles homogenises community similarity and erodes the distance-decay relationship of eDNA. Notably, after two introduced species have been removed, they are not detectable at any sampling position after 48 hrs. These findings on the spatial and temporal resolution of eDNA support that genetic material present in static environments originates from organisms that are nearby or have been nearby very recently. This work serves as an important case study of eDNA-based community diversity at fine temporal and spatial scales in ponds as a coherent view of eDNA ecology and dynamics begins to come into focus. While our observations are instructive, further quantitative modelling of eDNA transport, retention, and subsequent resuspension are needed to predict species location and estimate abundance (e.g., Jane et al. 2015; Jerde et al. 2016; Shogren et al. 2016; Shogren et al. 2017). This will be critical to take eDNA analysis to the next level as a powerful, diagnostic tool in ecology, conservation, and management. Regardless of modelling approaches, rigorous and spatially standardised sampling designs are key to ensuring the reliability of eDNA surveillance.

4.6 Supporting Information

Appendix S4.1 Supplementary tables and figures

Table S4.1 Number of sampling position of species detection over 14 days (D0–D14) in ponds E1 and E4 at the National Coarse Fish Rearing Unit.

Pond	Species	D0	D2	D4	D6	D8	D10	D12	D14
E1	RUD†	0	2 (P1/P4)	2 (P1/P4)	2 (P1/P2)	3 (P1/P2/P4)	0	0	0
E1	CHU†	0	1 (P1)	2 (P1/P2)	0	2 (P1/P4)	0	0	0
E1	BAR	5	5	5	5	5	5	5	5
E1	BRE	5	5	5	5	5	5	5	5
E1	CAR	5	5	5	5	5	5	5	5
E1	ROA	5	5	5	5	5	5	5	5
E1	TEN	5	5	5	5	5	5	5	5
E1	DAC	5	5	5	5	5	5	5	5
E4	RUD†	0	1 (P1)	1 (P1)	3 (P2/P3/P5)	1 (P1)	0	0	0
E4	CHU	5	5	5	5	5	5	5	5
E4	BAR	5	5	5	5	5	5	5	5
E4	BRE	5	5	5	5	5	5	5	5
E4	CAR	5	5	5	5	5	5	5	5
E4	ROA	5	5	5	5	5	5	5	5
E4	TEN	4 (P1/P2/P3/P5)	4 (P2/P3/P4/P5)	3 (P2/P3/P4)	5	5	2 (P1/P3)	2 (P1/P4)	4 (P1/P2/P3/P5)
E4	DAC†	0	1 (P1)	3 (P1/P3/P5)	4 (P1/P2/P4/P5)	4 (P1/P2/P3/P4)	0	0	0

Notes: Species three letter codes correspond to species given in Table 4.1. "†" indicates the rare species added to each pond for the purposes of this study. The exactly detected positions are showed in bracket if this species were not detected from all sampling positions. The different sampling stages and linear distance between sampling positions are described in Figure 4.3.

Table S4.2 Detection probability of the introduced species during the introductory stage (D2–D8) in ponds E1 and E4 at the National Coarse Fish Rearing Unit.

Pond	Species	P1	P2	P3	P4	P5
E1	RUD	1.00	0.50	0	0.75	0
E1	CHU	0.75	0.25	0	0.25	0
E4	RUD	0.75	0.25	0.25	0	0.25
E4	DAC	1.00	0.50	0.50	0.50	0.50
Mean ± SD		0.88 ± 0.13^{a}	0.38 ± 0.13^{b}	0.19 ± 0.21^{b}	0.38 ± 0.28^{b}	0.19 ± 0.21^{b}

Notes: Species three letter codes correspond to species given in Table 4.1. Linear distance between sampling positions is described in Figure 4.3. Sampling positions that differ significantly ($p < 0.05$) from one another are indicated by different letters.

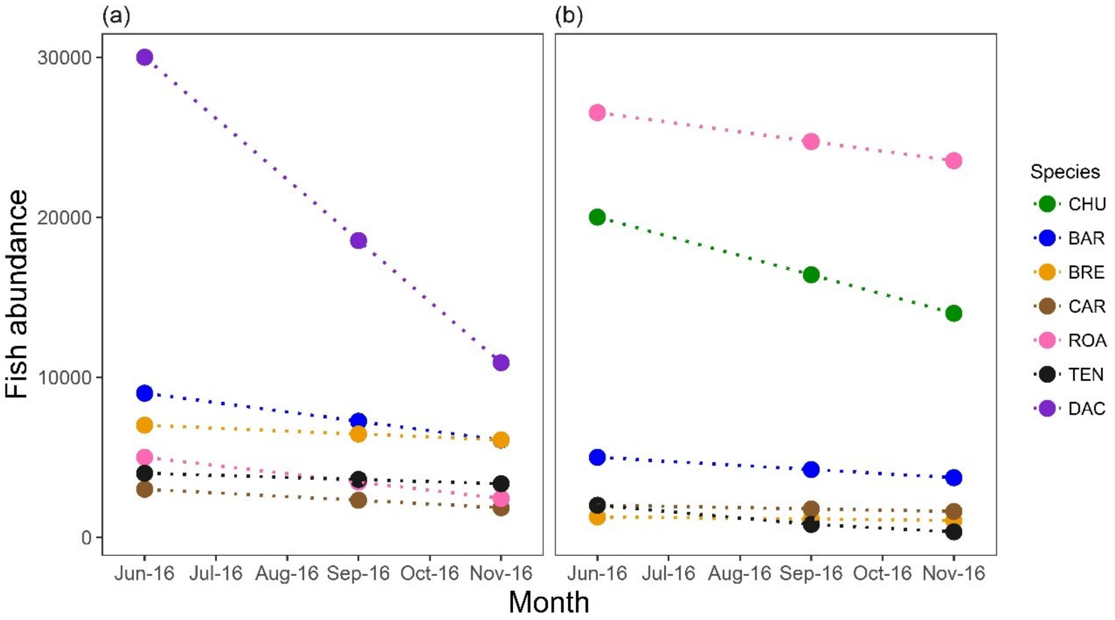


Figure S4.1 Death curves in ponds (a) E1 and (b) E4 at the National Coarse Fish Rearing Unit from June 2016 and to November 2016. Species three letter codes correspond to species given in Table 4.1.

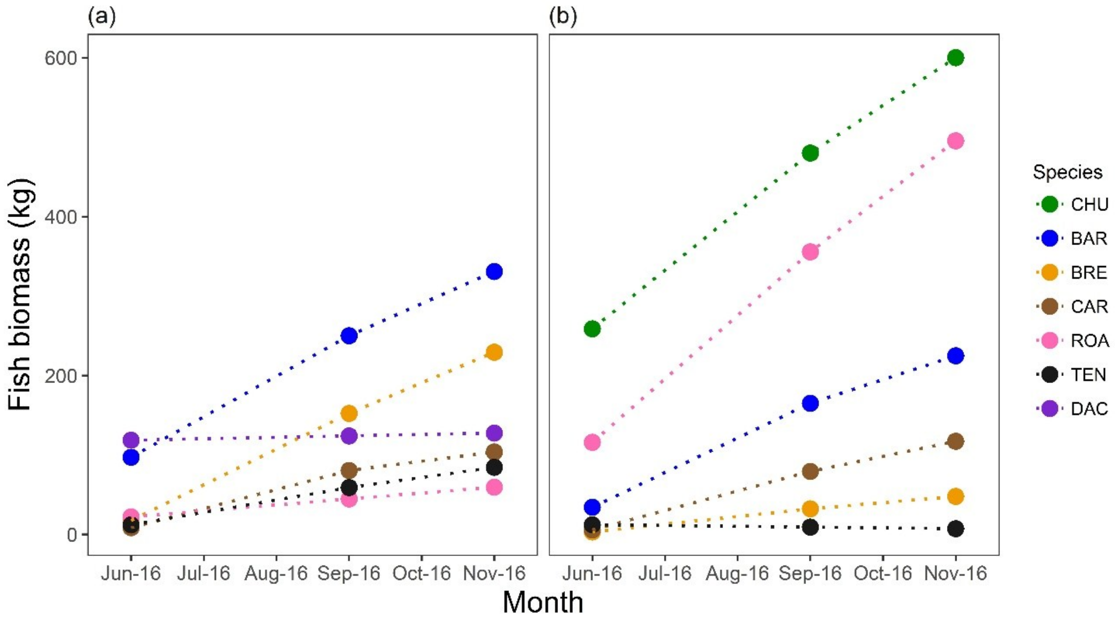


Figure S4.2 Growth curves ponds (a) E1 and (b) E4 at the National Coarse Fish Rearing Unit from June 2016 and to November 2016. Species three letter codes correspond to species given in Table 4.1.

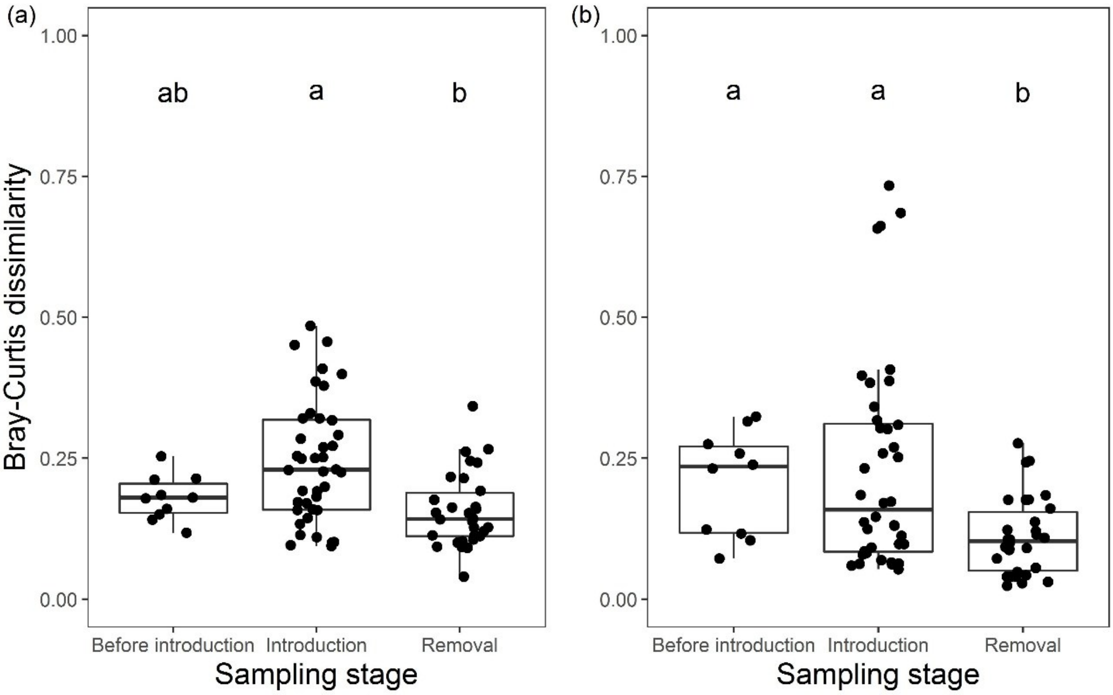


Figure S4.3 Bray-Curtis dissimilarity of different sampling stages across five sampling positions in ponds (a) E1 and (b) E4. The different sampling stages: before introduction (D0), introduction (D2–D8), and removal (D10–D14). Each point represents the Bray-Curtis dissimilarity of two different sampling positions at the same sampling stage. Sampling stages that differ significantly (p < 0.05) from one another are indicated by different letters in each boxplot.

Appendix S4.2 Read counts of OTUs data was used for the R script (.csv; supplied in a separate file)

The file can be viewed or downloaded use the link as below:

https://github.com/HullUni-bioinformatics/Li_et_al_2019_eDNA_dynamic/blob/master/Appendix_S1.csv

Chapter 5 Ground-truthing of a fish-based environmental DNA metabarcoding method for assessing the quality of lakes[4]

Abstract

Accurate, cost-effective monitoring of fish is required to assess the quality of lakes under the European Water Framework Directive (WFD). Recent studies have shown that environmental DNA (eDNA) metabarcoding is an effective and non-invasive method, which can provide semi-quantitative information on fish communities in large lakes. This study further investigates the potential of eDNA metabarcoding as a tool for WFD status assessment by collecting and analysing water samples from eight Welsh lakes and six meres in Cheshire, England, with well-described fish faunas. Water samples (N = 252) are assayed using two mitochondrial DNA regions (Cytb and 12S rRNA). eDNA sampling indicates very similar species to be present in the lakes compared to those expected on the basis of existing and historical information. In total, 24 species with 111 species occurrences by lake are detected using eDNA. There is a significant positive correlation between expected faunas and eDNA data in terms of confidence of species occurrence. eDNA data can estimate relative abundance with the standard five-level classification scale (“DAFOR”). Four ecological fish communities are characterised using eDNA data which are agreed with the pre-defined lake types according to environmental characteristics. Shoreline sampling with 10 samples is

[4] This chapter has been published as ***Li, J.***, Hatton-Ellis, T.W., Lawson Handley, L.J., Kimbell, H.S., Benucci, M., Peirson, G. & Hänfling, B. Ground-truthing of a fish-based environmental DNA metabarcoding method for assessing the quality of lakes. *Journal of Applied Ecology.* https://doi.org/10.1111/1365-2664.13352

adequate for capturing the majority of species and perform better than open water sampling alone. To better estimate abundance of fish living in different habitats in large lakes, however, a combination of both shoreline and offshore sampling is recommended. In summary, this study provides further evidence that eDNA metabarcoding could be a powerful and non-invasive monitoring tool for WFD purpose in a wide range of lake types, considerably outperforming other methods for community-level analysis.

5.1 Introduction

The impact of anthropogenic pressures on aquatic ecosystems is ubiquitous and usually leads to alterations of the environment (Scheffer et al. 2001). To mitigate negative human effects, the first step in the improvement of the ecological quality of degraded water bodies is the development of an assessment system to evaluate the current situation. The main European initiative for water quality assessment and improvements, Water Framework Directive 2000/60/EC (WFD), requires all member states to reach “good” ecological status of lakes, rivers and ground waters based on biological elements including phytoplankton, macrophytes and phytobenthos, benthic invertebrates, and fish (CEC 2000).

Fish are widely considered as relevant for detecting and quantifying impacts of anthropogenic pressures on lakes and reservoirs (Argillier et al. 2013). Although most UK WFD ecological tools have been developed and deployed, an effective fish-based assessment tool for lakes remains problematic (Winfield 2002; Kelly et al. 2012). Development of lake fish classification tools is less well developed than those for rivers; nevertheless, a number of fish indices and metrics based on lake fish assemblages have been proposed. These include the Fish In Lakes classification tool (FIL2) developed for use in Ireland (Kelly et al. 2012) and the fish-based index to assess lake eutrophication status explored at a European scale (Argillier et al. 2013). The current fish-based indices or tools rely on semi-quantitative capture-based methods such as electro-fishing or gill-netting to provide information on fish composition and abundance, as well as age-structure of fish populations (Argillier et al. 2013). Furthermore, the choice of survey methods is heavily dependent on the depth of the lake and to a lesser extent its size. However, both of these sampling methods are relatively laborious and thus expensive, sometimes destructive, taxonomically biased, cannot be deployed in all situations (e.g., in or near dense vegetation, or entire water column in a very deep lake) and have poor sampling accuracy and precision. These drawbacks restrict their ability to meet WFD requirements (Kubecka et al. 2009; Winfield et al. 2009). A future strategy for WFD purposes thus requires a highly cost-effective approach, with a minimum amount of destructive sampling.

A significant "game-changer" in biodiversity monitoring in recent years is the analysis of environmental DNA (eDNA) which is a non-invasive genetic method that takes advantage of intracellular or extra-organismal DNA in the environment to detect the presence of organisms (Lawson Handley 2015; Thomsen & Willerslev 2015; Taberlet et al. 2018). A particularly promising approach is to simultaneously screen whole communities of organisms using eDNA metabarcoding. Several studies have shown that eDNA metabarcoding using High-Throughput Sequencing (HTS) offers tremendous potential as a complementary method tool to established monitoring methods for ecology and conservation of aquatic species (e.g., Hänfling et al. 2016; Port et al. 2016; Valentini et al. 2016). After reviewing and discussing the potential of eDNA metabarcoding as a tool for WFD status assessment, Hering et al. (2018) suggested that this approach is well-suited for fish biodiversity assessment, as the suitability of DNA-based identification is particularly high for fish. A prototype eDNA tool for fish biodiversity assessment was tested in three lakes in the English Lake District (Hänfling et al. 2016), and initial results from this are promising. However, in order to understand factors such as responses to ecological pressures, taxon-specific biases, sampling requirements in different lake types and management of low confident occurrence, a much larger dataset from a range of lakes with different ecological characteristics is required. Thus, the objectives of this study are (1) to broaden the lake fish dataset using eDNA-based metabarcoding analysis by collecting and analysing water samples from 14 UK lakes with well-described fish faunas; (2) to explore a possible approach to evaluate the confidence of species presence based on site occupancy and read counts and (3) to use a relative abundance scale to estimate species abundance by comparing the eDNA metabarcoding results to historical data; and (4) to explore effectiveness of different spatial sampling approaches, particularly comparing shore and offshore sampling.

5.2 Materials and methods

5.2.1 Study sites

Among the 14 lakes, eight Welsh lakes were chosen because they are under Sites of Special Scientific Interest (SSSI) and Special Areas of Conservation (SAC) legislation and have fish data from a variety of sources over the past 30 years. Six Cheshire meres were selected based on the availability of high-quality fishery survey data, collected by Ecological Consultancy Ltd (ECON) during the same period as water sample collection in this study in 2016.

The distribution of the 14 sampling lakes including eight Welsh lakes and six Cheshire meres are shown in Figure 5.1, and the general characteristics are outlined in Appendix S5.1 Table S5.1 based on data from the UK Lakes Portal (Hughes et al. 2004).

In brief, the 14 lakes can be divided into three types according to environmental characteristics. These three types were: Type 1: three low alkalinity lakes that are broadly upland in character (Llyn Cwellyn, Llyn Ogwen and Llyn Padarn); Type 2: two high alkalinity but shallow lakes (Llyn Traffwll and Llyn Penrhyn) on the west coast of Anglesey (North Wales) which are close to sea and accessible for migratory fish; and Type 3: nine high alkalinity but shallow lakes that are dominated by coarse fish (Kenfig Pool, Llan Bwch-llyn, Llangorse Lake, Maer Pool, Chapel Mere, Oss Mere, Fenemere, Watch Lane Flash and Betley Mere) (Appendix S5.2).

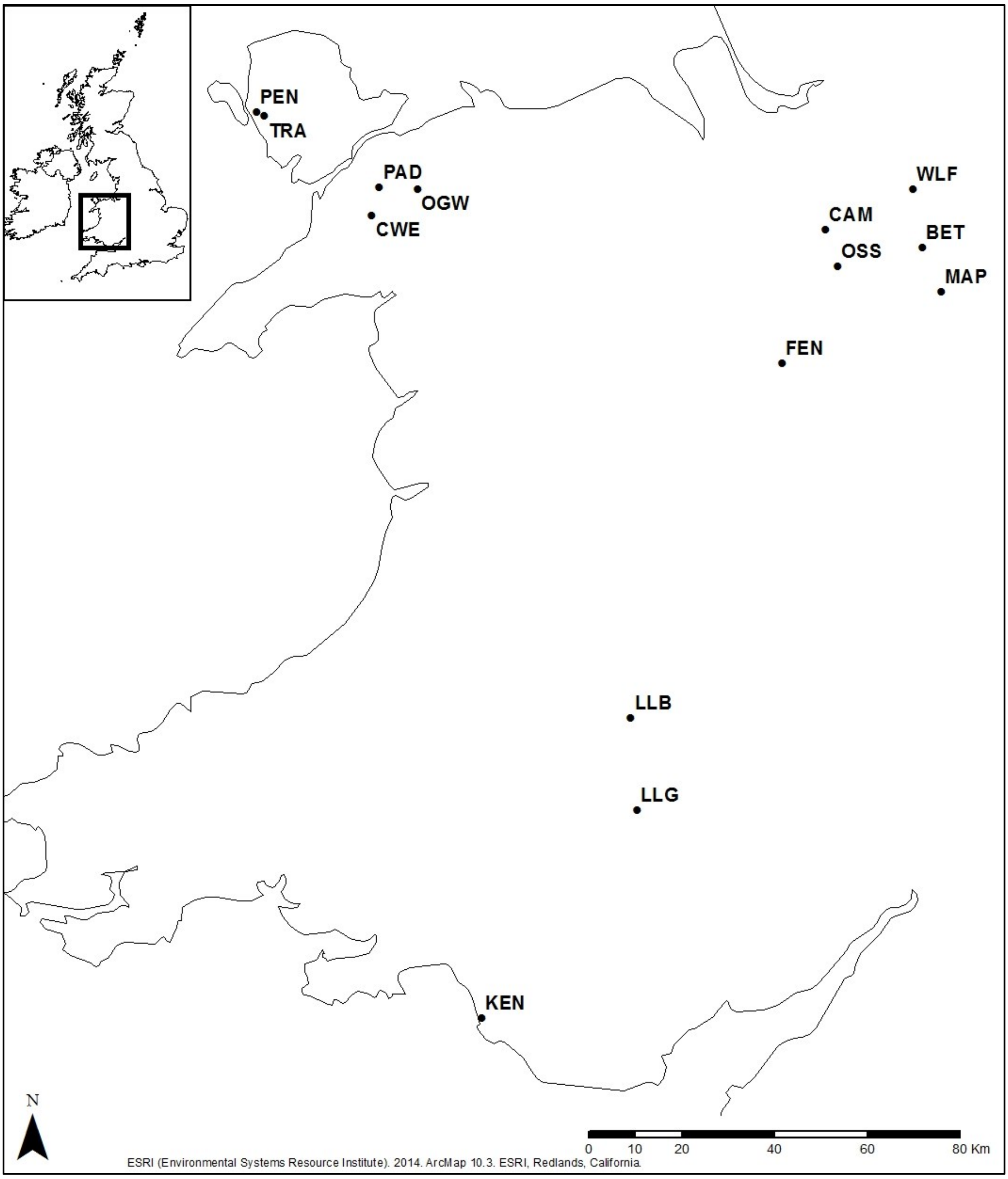

PEN
TRA
PAD
OGW
CWE
WLF
CAM
BET
OSS
MAP
FEN
LLB
LLG
KEN
N
0 10 20 40 60 80 Km
ESRI (Environmental Systems Resource Institute). 2014. ArcMap 10.3. ESRI, Redlands, California.

Figure 5.1 Distribution of sampling lakes. Sampling lake codes: (CWE) Llyn Cwellyn; (PAD) Llyn Padarn; (OGW) Llyn Ogwen; (PEN) Llyn Penrhyn; (TRA) Llyn Traffwll; (KEN) Kenfig Pool; (LLB) Llan Bwch-llyn; (LLG) Llangorse Lake; (MAP) Maer Pool; (CAM) Chapel Mere; (OSS) Oss Mere; (FEN) Fenemere; (WLF) Watch Lane Flash; and (BET) Betley Mere. The GPS coordinates of sampling locations in each lake are listed in Appendix S5.3 & Appendix S5.4.

5.2.2 Sampling strategy

In total, 252 samples were collected from the 14 lakes ranging in size from 8 to 140 ha. The GPS coordinates of sampling locations in each lake are listed in Appendix S5.3 & Appendix S5.4. Sampling involved the collection of water samples from Welsh lakes between 11/01/2016 to 12/07/2016, taking into account the surface area and the mean depth (Appendix S5.1 Table S5.2). Where surface area was < 30 ha and mean depth < 5 m, 10–12 2 L shore samples were collected (Kenfig Pool, Llyn Penrhyn and Llan Bwch-llyn Lake). Twenty shore samples were collected from lakes with the surface area of 30–50 ha and mean depth < 5 m (Llyn Ogwen and Llyn Traffwll). Ten shore and 10 offshore samples were collected from lakes with a surface area > 50 ha, (Llangorse Lake, Llyn Padarn and Llyn Cwellyn). At the two deep lakes (i.e., mean depth > 15 m), Llyn Padarn and Llyn Cwellyn, offshore samples were collected using a 1 L vertical water sampler with a small opening in the lid which was submerged slowly in order to collect throughout the entire water column from the surface to the bottom. Two 1 L samples were taken from at each offshore sampling point and pooled in a 2 L sterile plastic bottle. Between offshore sampling points, the sampler was sterilised by washing in 10% v/v commercial bleach solution (containing ~3% sodium hypochlorite) followed by 5% v/v microsol detergent (Anachem, UK) and rinsed with purified water. The sampler was then rinsed again in lake water at the next sampling point before sampling. At the large and shallow Llangorse Lake (140 ha, mean depth < 5 m), the 10 offshore samples were collected from the surface. The shore samples were collected and pooled from five 400 mL samples by submerging 2 L sterile bottles at arm’s length. All samples were stored in cool boxes until filtration. Twenty (10 shore samples and 10 offshore samples) were collected from each Cheshire mere between 01/12/2015 to 07/01/2016 (Appendix S5.1 Table S5.2). The collection of offshore samples was carried out by a small boat or canoe. Since the mean depth of the Cheshire meres is < 5 m, the offshore samples were taken from the water surface. The methods of shore sample collection and sample storage before filtration were same to Welsh lakes samples.

5.2.2 eDNA capture, extraction, library preparation and sequencing

5.2.2.1 eDNA capture and extraction

The samples from the Cheshire meres and Llyn Traffwll, Kenfig Pool, Llan Bwch-llyn and Llangorse Lake were filtered in the dedicated eDNA laboratory at the University of Hull (UoH). Samples from other Welsh lakes (Llyn Padarn, Llyn Cwellyn, Llyn Ogwen and Llyn Penrhyn) were filtered in a laboratory at the Bangor University that was not used for handling fish or DNA and was decontaminated before filtration by bleaching floors and surfaces.

All samples were filtered within 24 hrs of collection. Two litres of each water sample was filtered through a 47 mm diameter 0.45 µm mixed cellulose acetate and nitrate filter (Whatman) using Nalgene filtration units in combination with a vacuum pump (15–20 in. Hg; Pall Corporation). Our previous study demonstrated that the 0.45 µm filters are suitable for fish metabarcoding, with low variation and high repeatability between the filtration replicates (see more detail in Chapter 3; Li et al. 2018a) compared to other filtration methods. The filtration units were cleaned with 10% v/v commercial bleach solution (containing ~3% sodium hypochlorite) and 5% v/v microsol detergent (Anachem, UK), and then rinsed thoroughly with de-ionised water after each filtration run to prevent cross-contamination. For each sampling lake, one sampling blank and one filtration blank (2 L deionised water) were filtered before sample filtration in order to test for possible contamination at the sampling and filtration stages. After filtration, all filters were placed into 50 mm sterile petri dishes using sterile tweezers, sealed with parafilm and then immediately stored at –20 °C until extraction. DNA extraction was carried out using the PowerWater DNA Isolation Kit (Qiagen) following the manufacturer's protocol.

5.2.2.2 Library preparation and sequencing

Sequencing libraries were generated from PCR amplicons targeting two mitochondrial loci: Cytochrome b (Cytb) and 12S rRNA (12S). We previously tested both fragments *in vitro* on 22 common freshwater fish species and 10 mock communities (see more detail in Chapter 2), and *in situ* on three deep lakes in the English Lake District, and demonstrated their suitability for eDNA metabarcoding of UK lake fish communities (Hänfling et al. 2016).

To enable the detection of possible PCR contamination, I included no-template controls (NTCs) of molecular grade water (Fisher Scientific) and single-template controls (STCs) of cichlid fish DNA (the Eastern happy, *Astatotilapia calliptera,* a cichlid from Lake Malawi, which is not present in natural waters in UK) within each library (*N* = 308). Three PCR technical replicates were performed for each sample then pooled to minimise bias in individual PCRs. All PCRs were set up using eight-strip PCR tubes in a PCR workstation with UV hood and HEPA filter in the eDNA laboratory of UoH to minimise the risk of contamination.

The Cytb locus, targeting a 414-bp vertebrate-specific fragment (Kocher et al. 1989), was amplified using a one-step library preparation protocol (Kozich et al. 2013). The Cytb one-step library preparation protocol for this study was described in Chapter 2 Section 2.2.3.1. The final 10 pM denatured Cytb library mixed with 30% PhiX genomic control was sequenced with the MiSeq reagent kit v3 (2 × 300 cycles) at UoH. The 12S locus targets a 106-bp vertebrate-specific fragment (Riaz et al. 2011). Following the consistently lower sequencing yield of the one-step protocol for the 12S compared to Cytb in previous studies, we decided to switch to a two-step library preparation protocol using nested tagging (Kitson et al. 2019). The detail of the full two-step 12S library preparation was described in Chapter 4 Section 4.2.3. To improve clustering during initial sequencing, the denatured 12S library (13 pM) was mixed with 10% PhiX genomic control. The library was sequenced with the MiSeq reagent kit v2 (2 × 250 cycles) at UoH.

5.2.3 Data analysis

5.2.3.1 Collation of fish data from historical data

Existing fish data for the Welsh lakes and Cheshire meres was collated using a range of data sources (see more detail in Appendix S5.2). These included site-specific surveys carried out by the Natural Resources Wales (NRW) or the Environment Agency (EA) and their predecessor bodies, third-party fishery surveys, data published in the literature, and *ad hoc* records stored on databases such as the National Biodiversity Network Atlas (https://nbnatlas.org). NRW and EA fisheries staff were also consulted for their knowledge of each site, and where available, current and historical stocking records were used. Atlas data (Davies et al. 2004) was also used to provide a general strategic overview of the range of individual taxa.

Despite the existence of survey data, our knowledge of the fish faunas of many sampling sites is imperfect, due to the strongly selective nature of many survey methods and the tendency to overlook species that are not of economic importance. Consequently, data to inform the interpretation of unexpected eDNA records was also collected for each sampling site. This included the presence of inflows and outflows that might explain the presence of rheophilic species, distance to the sea and accessibility to and from it that might explain the presence of diadromous species, habitat suitability (including spawning habitat), and presence of nearby settlements or other human developments (e.g., sewage treatment works, fish farms) that may act as sources of DNA. Only a qualitative analysis of these factors has been attempted here to provide contextual information for records that have not been ground-truthed. Based on the above information and expert opinion (Hatton-Ellis T.W. and Graeme P.), fish species for each sampling site were placed into the four general categories with confidence of species presence/absence for each lake (Table 5.1), reflecting both the available data and the uncertainty around it.

Assessing abundance *a priori* was more challenging. Hänfling et al. (2016) compared eDNA abundance to the long-term rank abundance of the fish community in Windermere which is a very well-studied lake. However, existing datasets here consisted of a heterogeneous mix of data collected using multiple methods and at different dates. Whilst these data were suitable for generating an "expected" fish fauna for each lake, it was not possible to produce rank-abundance estimates. Consequently, for the Welsh lakes, an abundance category approach has been used, with each species being placed on the DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare) using available data where possible and/or Hatton-Ellis T.W.'s expert opinion. The DAFOR scale is a classic multilevel descriptor scale, which is used for semi-quantitative sampling to provide a quick estimate of the relative abundance of species (generally plants) in a given area (Tansley 1993) All records where no abundance was recorded were assumed to be rare. For the Cheshire meres, summaries of fish community composition and fish density per unit area (ind. ha^{-1} and kg ha^{-1}) were produced from ECON fishery survey using point abundance sampling by electro-fishing (PASE) and seine netting in the same period as water samples collection with this study in 2016 (Appendix S5.1 Table S5.3 & Table S5.4). ECON fishery survey in 2016 was undertaken by ECON and commissioned by Natural England as part of an investigation of the impacts of fisheries on SSSI lake condition.

Table 5.1 Criteria for assessing the confidence of species presence/absence using existing data.

Category	**Confidence in Presence**	**Description**
Probably absent (PrA)	Very low	No records and at least one of outside biogeographic range/habitat unsuitable.
Possibly present (PoP)	Low	Either not recorded since 1977 or not recorded from the lake, but present in connected water body or diadromous species with direct access to the sea.
Probably present (PrP)	Moderate	Either multiple records but not recorded since 1997 or only one record since 1997. For Welsh lakes, this category has also been used where the current status of a population is uncertain, even if records data meet the criteria for Established.
Established (E)	High	For Welsh lakes, multiple records including at least one since 2000. For Cheshire meres, records in the Ecological Consultancy Limited (ECON) survey.

5.2.3.2 Bioinformatics analysis

Raw read data from Illumina MiSeq sequencing have been submitted to NCBI (BioProject: PRJNA454866). Bioinformatics analysis was implemented following a custom reproducible metabarcoding pipeline (metaBEAT v0.97.9) (see more detail in

Appendix S2.1 Section S2.1.1; Hänfling et al. 2016) with custom-made reference databases (Cytb and 12S) as described in Chapter 2. Sequences for which the best BLAST hit had a bit score below 80 or had less than 95%/100% (Cytb/12S) similarity to any sequence in the curated databases were considered non-target sequences. To assure full reproducibility of our bioinformatics analysis, the custom reference databases and the Jupyter notebooks for data processing have been deposited in an additional dedicated GitHub repository (https://github.com/HullUni-bioinformatics/Li_et_al_2019_eDNA_fish_monitoring). The Jupyter notebook also performs demultiplexing of the indexed barcodes added in the first PCR reactions.

5.2.3.3 Low-frequency noise threshold

Filtered data were summarised into the number of sequence reads per species/sample for downstream analyses (Appendix S5.3 & Appendix S5.4). After bioinformatics analysis, the low-frequency noise threshold (proportion of STC species read counts in the real sample) was set to 0.07% and 0.3% for Cytb and 12S respectively to filter high-quality annotated reads passing the previous filtering steps that have high-confidence BLAST matches but may result from contamination during the library construction process or sequencing (De Barba et al. 2014; Hänfling et al. 2016; Port et al. 2016). Filtered data were summarised in two ways for downstream analyses: (1) the number of sequence reads per species at each lake (hereafter referred to as read counts) and (2) the proportion of sampling locations in which a given species was detected (hereafter referred to as site occupancy).

5.2.3.4 Estimating abundance with eDNA

Based on other studies (Pilliod et al. 2013; Schmidt et al. 2013; Ficetola et al. 2015; Hänfling et al. 2016; Valentini et al. 2016; Lawson Handley et al. 2019), site occupancy is a better proxy for estimates of abundance than read counts. However, sequencing read count should not be ignored entirely because they still contain important abundance information. The maximum site occupancy across both loci was used as a score for species abundance (Table 5.2). In addition to site occupancy score, the maximum relative read count (i.e., proportion of read counts per species at each sampling lake) across both loci, and the number of loci with which the species was detected were used as confidence indicators to assign fish species into the same general categories which were used for non-eDNA survey data (Table 5.3). Those species that were assigned the lowest confidence scores (probably absent) were excluded from downstream analysis (i.e., roach *Rutilus rutilus* in Llyn Cwellyn, bream *Abramis brama* and bullhead *Cottus gobio* in Llyn Ogwen, pike *Esox lucius* in Llyn Penrhyn, bullhead in Llyn Traffwll, and gudgeon *Gobio gobio* in Betley Mere). Each retained species was assigned a corrected abundance score by multiplying site occupancy score × confidence score. The corrected abundance score was used to assign each species to a relative abundance DAFOR scale (Table 5.4).

Table 5.2 Site occupancy score based on maximum site occupancy across Cytb and 12S.

Site occupancy score	Maximum site occupancy (across Cytb and 12S)
1	> 0 and ≤ 0.1
2	> 0.1 and ≤ 0.3
3	> 0.3 and ≤ 0.6
4	> 0.6 and ≤ 0.8
5	> 0.8

Table 5.3 Criteria for assessing confidence of species occurrence using eDNA data.

Category	Confidence score	Description		
		Locus	Site occupancy score	Maximum relative read count (across Cytb and 12S)
Probably absent (PrA)	Very low (-1)	Any	= 1	> 0 and ≤ 0.005
Possibly present (PoP)	Low (1)	Any	> 1	> 0 and ≤ 0.005
		One	≥ 1	> 0.005 and ≤ 0.05
Probably present (PrP)	Moderate (3)	Both	≥ 1	> 0.005 and ≤ 0.01
		Both	= 1	> 0.01 and ≤ 0.1
		One	≥ 1	> 0.05 and ≤ 0.1
		One	= 1	> 0.1 and ≤ 0.5
Established (E)	High (5)	Both	> 1	> 0.01
		Both	= 1	> 0.1
		One	> 1	> 0.1
		One	= 1	> 0.5

Table 5.4 The relative abundance DAFOR scale based on eDNA data.

Abundance score	**DAFOR scale**	**Corrected abundance score (site occupancy score × confidence score)**
0	None	= 0
1	Rare	≥ 1 and < 4
2	Occasional	≥ 4 and < 9
3	Frequent	≥ 9 and < 15
4	Abundant	≥ 15 and < 25
5	Dominant	= 25

5.2.3.5 Ecological and statistical analyses

All downstream analyses were performed in R v3.3.2 (R_Core_Team 2016), and graphs were plotted using GGPLOT2 v2.2.1 (Wickham & Chang 2016). Before investigating species detection and abundance estimate with eDNA, we first evaluated whether Cytb and 12S datasets produced consistent results by calculating the Pearson's product-moment correlation coefficient for both read counts and site occupancy. Spearman's rank correlations were performed between historical data and eDNA data – firstly in terms of confidence of species presence and secondly in terms of relative abundance. To investigate differences in fish communities between sampling lakes, hierarchical clustering dendrograms was used to assess the existence of distinct community types using the function *fviz_dend* in FACTOEXTRA v1.0.4 (Kassambara & Mundt 2017) to extract and visualise the results calculated from the function *hclust* using Canberra distances method based on site occupancy. Furthermore, non-metric multidimensional scaling (NMDS) in individual sampling location of lakes, allied with analysis of similarities (ANOSIM), were performed using the abundance-based Bray-Curtis dissimilarity index with the function *metaMDS* and *anosim* respectively in

VEGAN v2.4-4 (Oksanen et al. 2017). Finally, sample-based rarefaction (Gotelli & Colwell 2011) was applied to determine the number of shore samples needed to detect the majority of the species present. Rarefaction was performed with 499 randomizations using the function *rich* and *rarc* in VEGAN v2.4-4 (Oksanen et al. 2017). The full R script is available on the GitHub repository (https://github.com/HullUni-bioinformatics/Li_et_al_2019_eDNA_fish_monitoring/tree/master/R_script)

5.3 Results

Libraries generated 17.15 million reads with 15.91 million reads passing filter including 49.01% PhiX for Cytb and 15.97 million reads with 13.40 million reads passing filter including 11.07% PhiX for 12S. After quality filtering and removal of chimeric sequences, the average read count per sample (excluding controls) was 12,503 for Cytb and 21,703 for 12S. After BLAST searches for taxonomic assignment, 59.47% ± 35.56% and 26.46% ± 19.60% reads in each sample were assigned to fish for Cytb and 12S, respectively.

Significant correlations were found between site occupancy and average read counts, for both loci and all sampling lakes apart from Llyn Ogwen (Pearson's r = 0.81, df = 4, p = 0.05) and Llan Bwch-llyn (Pearson's r = 0.94, df = 2, p = 0.06) for Cytb (Appendix S5.1 Figure S5.1a, b). Data from the two loci were significantly correlated (Pearson's r consistently $p < 0.05$) for site occupancy for every lake (Appendix S5.1 Figure S5.1c), and for average read counts apart from in three lakes: Llyn Padarn (Pearson's r = 0.55, df = 7, p = 0.12), Llyn Traffwll (Pearson's r = 0.59, df = 5, p = 0.16), and Llan Bwch-llyn (Pearson's r = 0.89, df = 2, p = 0.11) (Appendix S5.1 Figure S5.1d).

5.3.1 Comparison of eDNA data and expected results

In total, 24 species with 111 species occurrences by lake were detected across Cytb and 12S. Perch *Perca fluviatilis* was the most common species which were detected in all lakes. The second most frequently occurring species was roach with detections in 10 lakes. In addition to these two species, pike were found in all nine coarse fish lakes (Table 5.5; Figure 5.2; Appendix S5.1 Figure S5.2). There was a significant positive correlation between expected faunas and eDNA data in terms of confidence of species occurrence by lake (Spearman's $r = 0.74$, $df = 109$, $p < 0.001$). A total of 73 species occurrences have been recorded across all lakes according to historical data. Of these, 72 (98.6%) were detected with eDNA (Table 5.5). The only false negative in the eDNA datasets was tench *Tinca tinca* in Llyn Penrhyn. Tench were detected by 12S in Llyn Penrhyn but at a read count less than the low-frequency noise threshold (Appendix S5.4; site occupancy = 0.3, average read counts = 3.9). Tench are likely to be very rare in Llyn Penrhyn: only two tench were recorded by the 2016 Royal Society for the Protection of Birds survey, and small numbers have also been detected in previous years (Appendix S5.2 Section S5.2.4).

There were consistent positive correlations between DAFOR scale based on eDNA data and expected DAFOR scale based on both historical data for Welsh lakes and fish density per unit area from ECON survey of Cheshire meres (Figure 5.3). The correlations were significant in three out of eight Welsh lakes (Figure 5.3). Significant correlations were observed in three and four Cheshire meres based on individual density (ind. ha^{-1}) and biomass density (kg ha^{-1}), respectively (Figure 5.3; Appendix S5.1 Figure S5.3).

Table 5.5 Correspondence of confidence of species occurrence between predicted and eDNA data in 14 lakes

Community type		**1**			**2**		**3**							**4**		**Number of lakes**
Scientific name	**Code**	**CWE**	**PAD**	**OGW**	**PEN**	**TRA**	**KEN**	**LLB**	**LLG**	**MAP**	**CAM**	**OSS**	**FEN**	**WLF**	**BET**	
Abramis brama	BRE							PrP/E	E/E	PoP/PoP	PoP/PoP	E/E	E/E	E/E	E/E	8
Alburnus alburnus	BLE			PrA/PoP			PrA/PoP									2
Anguilla anguilla	EEL	E/PrP	E/E	PoP/PrP	E/E	E/E	E/E		E/E				E/E	PoP/PoP		9
Carassius auratus	GOF		PrA/PrP													1
Carassius carassius	CRU														PoP/PrP	1
Cottus gobio	BUL						PrA/PoP		PoP/PrP						PoP/PoP	3
Cyprinus carpio	CAR									E/E	PoP/PoP	E/E	E/E	E/E	E/E	6
Esox lucius	PIK						E/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E	9
Gasterosteus aculeatus	3SS		E/E		E/E	E/E					PoP/E		E/PrP	PoP/E	PoP/PrP	7
Gobio gobio	GUD		PrA/PoP									PoP/PoP		E/PoP		3
Leucaspius delineatus	SUN													E/E		1
Oncorhynchus mykiss	RTR			E/E												1
Perca fluviatilis	PER	PrA/E	PrP/PoP	PrA/PrP	E/E	PrA/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E	14
Phoxinus phoxinus	MIN	E/E	E/E	E/E								PrA/PoP				4
Pseudorasbora parva	TMG														PrA/PoP	1
Pungitius pungitius	9SS				PoP/PrP	PoP/PoP	PoP/PoP									3
Rhodeus amarus	BIT													PoP/PoP		1
Rutilus rutilus	ROA				E/E	E/E	PoP/PoP	E/E	E/E	PrA/PoP		E/E	E/E	E/E	E/E	10
Salmo salar	SAL	E/PoP	E/E													2
Salmo trutta	BTR	E/E	E/E	E/E	PoP/PoP	PoP/E									PoP/PoP	6
Salvelinus alpinus	CHA	E/E	E/PrP													2
Scardinius erythrophthalmus	RUD				E/E	PoP/PrP	E/E		E/E		PrA/PoP	E/E	E/E			7

Squalius cephalus	CHU													PrA/PoP		1
Tinca tinca	TEN				E/PrA		E/E		E/E	E/E	E/E	E/E	PoP/PoP	PoP/PrP	E/E	9
Number of species		6	9	6	8	7	9	4	8	6	7	9	9	12	11	111 in total

Notes: Categories of presence confidence ("PrA" probably absent, "PoP" possibly present, "PrP" probably present and "E" established) are described in Table 5.1 showing as based on predicted/eDNA data. The "Established" species occurrences based on historical data are shown in blue background (N = 73). Solid vertical lines indicate community types (Community 1–3) with similar faunas. Dashed line indicates division between whether bream Abramis brama is dominant species in the lake (Community 4). Sampling lake codes are given in Figure 5.1.

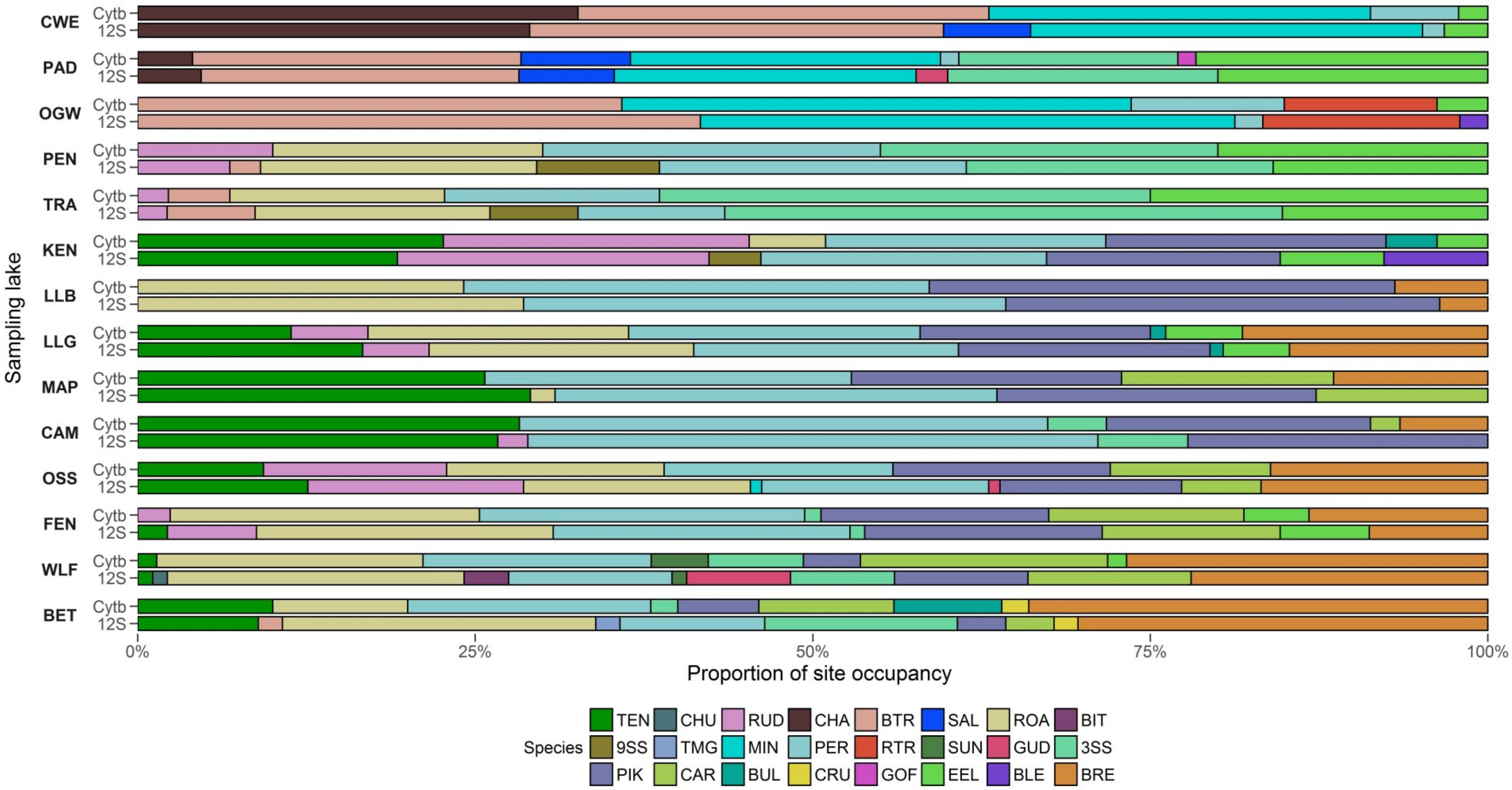


Figure 5.2 Species composition of site occupancy for Cytb and 12S. Species three letter codes and sampling lake codes are given in Table 5.5 and Figure 5.1, respectively.

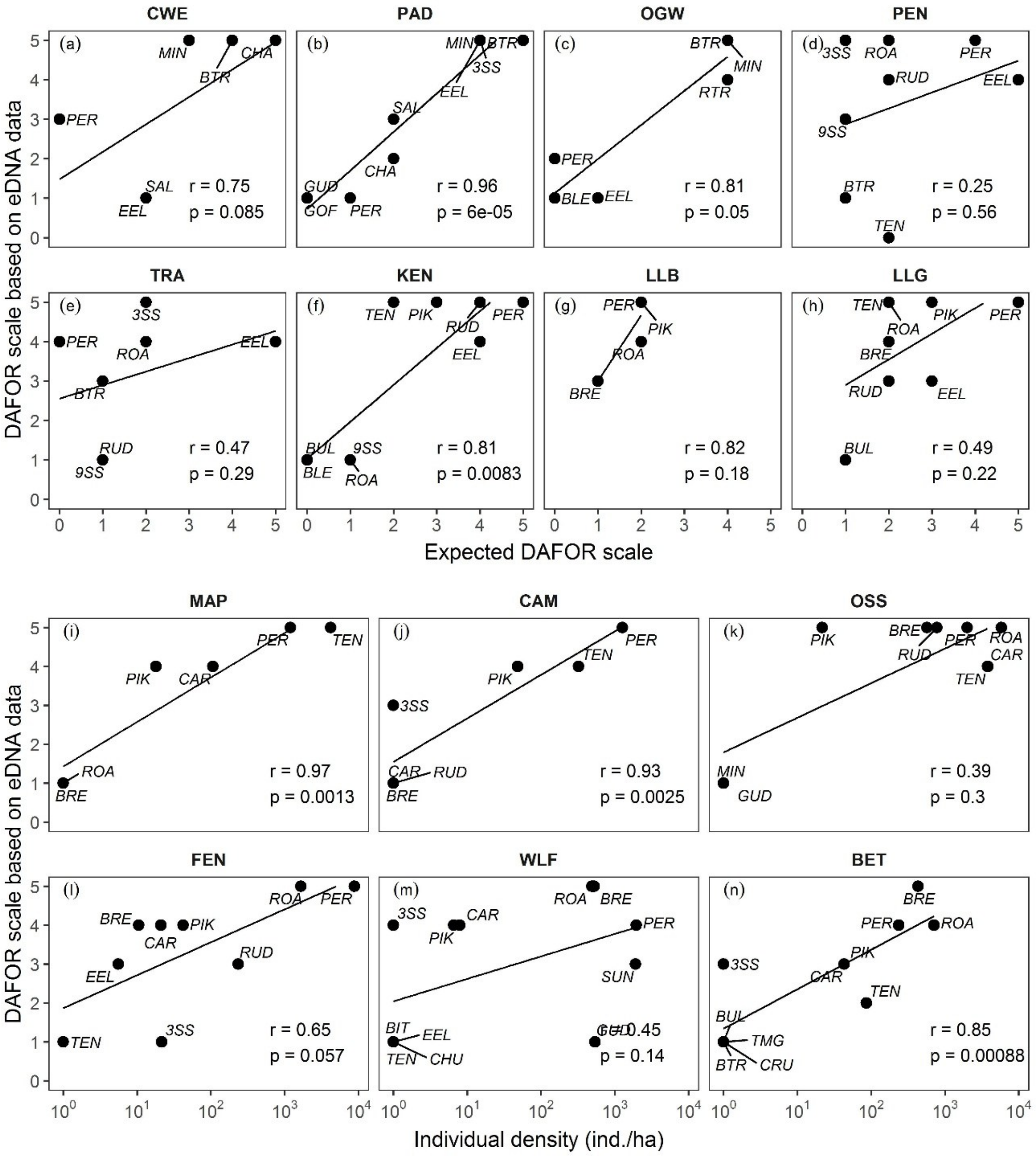


Figure 5.3 Correlations between DAFOR scale based on eDNA data and expected DAFOR scale based on existing data of Welsh lakes or individual density from the Ecological Consultancy Ltd survey of Cheshire meres. Sampling lake codes, corrected abundance scores to DAFOR scale and species three letter codes are given in Figure 5.1, Table 5.4 and Table 5.5, respectively.

5.3.2 Characterisation of fish assemblages using eDNA data

The hierarchical clustering dendrograms based on site occupancy for two loci indicated that there were four distinct community types compared to three pre-defined lake types according to environmental characteristics. Specifically, the clustering dendrograms accorded with the pre-defined lake Type 1 (Llyn Cwellyn, Llyn Ogwen and Llyn Padarn) and Type 2 (Llyn Traffwll and Llyn Penrhyn), but indicated that Watch Lane Flash and Betley Mere can be divided from the pre-defined lake Type 3 (Figure 5.4a1, b1). NMDS ordination allied with ANOSIM based on read counts for two loci confirmed the clustering dendrograms results that there were four distinct community types according to the predominant groups of fish (Figure 5.4a2, b2). These four identified community types were: Community 1: salmonids and minnow *Phoxinus phoxinus*; Community 2: mixed diadromous fish; Community 3: coarse fish; and Community 4: bream-dominated coarse fish (Figure 5.4). The *R* statistics in the ANOSIM global test with these four different communities were high (Appendix S5.1 Table S5.5, 0.75 ± 0.02) supporting statistical differences between communities (Appendix S5.1 Table S5.5, $p = 0.001$). The overall distance pattern of these four fish communities was that both coarse fish communities (Community 3 and Community 4) were close to each other, the mixed diadromous fish community (Community 2) was between the coarse fish communities and the salmonids and minnow community (Community 1) (Figure 5.4; Appendix S5.1 Table S5.5).

The salmonids and minnow community (Community 1) is characteristic of three low alkalinity upland lakes (Llyn Cwellyn, Llyn Ogwen and Llyn Padarn). These sites were generally dominated by brown trout *Salmo trutta*, rainbow trout *Oncorhynchus mykiss*, or Arctic charr *Salvelinus alpinus*, but also included minnow and Atlantic salmon *Salmo salar* (Figure 5.2; Appendix S5.1 Figure S5.2). Llyn Cwellyn and Llyn Padarn are designated as SSSIs under the Wildlife and Countryside Act 1981 (as amended) to conserve Arctic charr, and the oligotrophic lake habitat of Llyn Cwellyn is also protected as a SAC under the EU Habitats Directive. The eDNA data showed that Llyn Padarn contained a small number of Arctic charr, whereas this species was dominant in Llyn Cwellyn (Figure 5.2; Appendix S5.1 Figure S5.2). Rainbow trout were only detected in Llyn Ogwen, where they are regularly stocked by the local angling club (Appendix S5.2 Section 5.2.3). Some of these sites may also contain a low density of other species; usually diadromous species were detected using eDNA as well such as European eel *Anguilla anguilla* in all three upland lakes and three-spined stickleback *Gasterosteus aculeatus* in Llyn Padarn. Perch have recently been recorded from Llyn Padarn for the first time using captured-based survey methods (Appendix S5.2 Section 5.2.2). Surprisingly, this species was detected using eDNA (though shown to be rare), in all of these three lakes (Figure 5.2; Table 5.5). The diadromous fish community (Community 2) is characteristic of Llyn Traffwll and Llyn Penrhyn which were dominated by European eel and three-spined stickleback. Other species included brown trout, nine-spined stickleback *Pungitius pungitius*, roach, perch, and rudd *Scardinius erythrophthalmus* (Figure 5.2). The coarse fish communities consisted of bream, common carp *Cyprinus carpio*, pike, perch, roach, rudd and tench, plus some additional species (e.g., European eel, bullhead, three-spined stickleback and gudgeon). Pike, perch, roach, and tench were dominant in Kenfig Pool, Maer Pool, Chapel Mere, Llan Bwch-llyn, Llangorse Lake, Oss Mere and Fenemere (Community 3); however, the dominant species in Watch Lane Flash and Betley Mere (Community 4) was bream (Figure 5.2; Appendix S5.1 Figure S5.2).

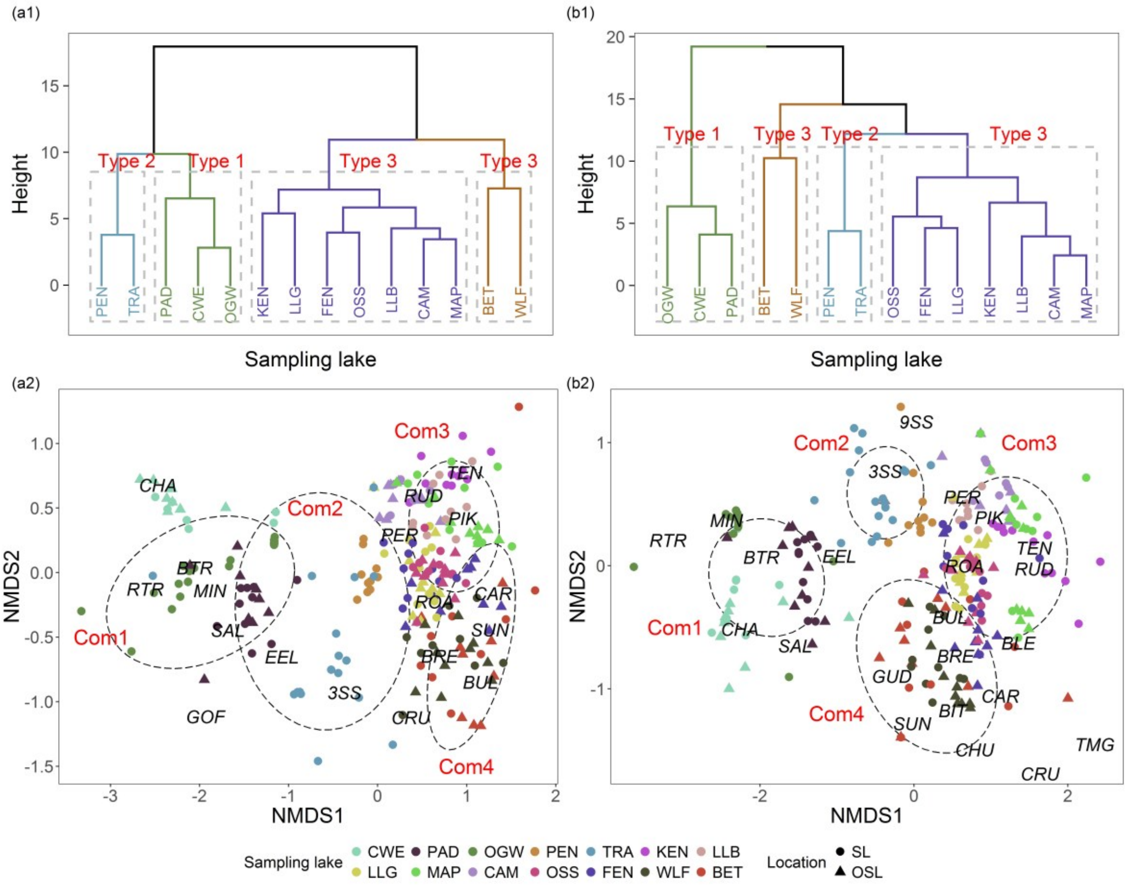


Figure 5.4 Hierarchical clustering dendrograms and non-metric multidimensional scaling (NMDS) ordination of the sampling lakes. Dendrograms using Canberra distances method based on site occupancy for (a1) Cytb and (b1) 12S. The dashed frames are drawn around each cluster in dendrograms. The three pre-defined lake types according to environmental characteristics. NMDS in individual sampling location of lakes using Bray-Curtis dissimilarity index based on read counts for (a2) Cytb and (b2) 12S. Each symbol corresponds to a sampling lake, with circles corresponding to shore samples ("SL") and triangles corresponding to offshore samples ("OSL") in NMDS ordination. The ellipse indicates 70% similarity level within each community type in ordinations. Scores for species taxa are plotted on the same axes to better visualise the ordination in space between species and samples. Species three letter codes and community type (Com 1–4) with individual lake codes are given in Table 5.5 and Figure 5.1, respectively.

5.3.3 Evaluation of shoreline and boat-based samples

Offshore samples were available for nine out of 14 lakes (Appendix S5.1 Table S5.2). The average species richness of shore and offshore samples across two loci was 7.33 ± 3.25 and 7.11 ± 3.11, respectively, and there was no significant difference between them (Paired t-test, t = 1.00, df = 8, p = 0.35). The NMDS ordination showed a high degree of overlap between shore and offshore samples for each lake across two loci (Figure 5.4). Moreover, significant correlations were observed between site occupancy from shore and offshore samples excluding Maer Pool in the Cytb dataset (Spearman's r = 0.68, df = 4, p = 0.14; Appendix S5.1 Figure S5.4). From the perspective of each species across lakes, there were consistently positive correlations between shore and offshore samples (Appendix S5.1 Figure S5.5). In the Cytb dataset, the correlations between shore and offshore samples were significant excluding bream (Pearson's r = 0.72, df = 5, p = 0.07), common carp (Pearson's r = 0.55, df = 4, p = 0.25) and rudd (Pearson's r = 0.92, df = 2, p =0.08). In the 12S dataset, the correlations were significant excluding common carp (Pearson's r = 0.81, df = 4, p = 0.05), rudd (Pearson's r = 0.86, df = 2, p = 0.14), and gudgeon (Pearson's r = 0.60, df = 1, p = 0.59). Common carp tended to be more abundant in offshore samples for both loci (Appendix S5.1 Figure S5.5c1, c2).

Sample-based rarefaction analyses indicated that approximately seven shore samples captured the majority (~85%) of the taxa present in 10 out of 14 lakes across both Cytb and 12S datasets. For the remaining lakes, 10 shore samples were needed to capture the majority of the taxa present in Fenemere, Watch Lane Flash, and Betley Mere; while 11 shore samples were necessary to adequately characterise the fish fauna in Llyn Traffwll (Figure 5.5).

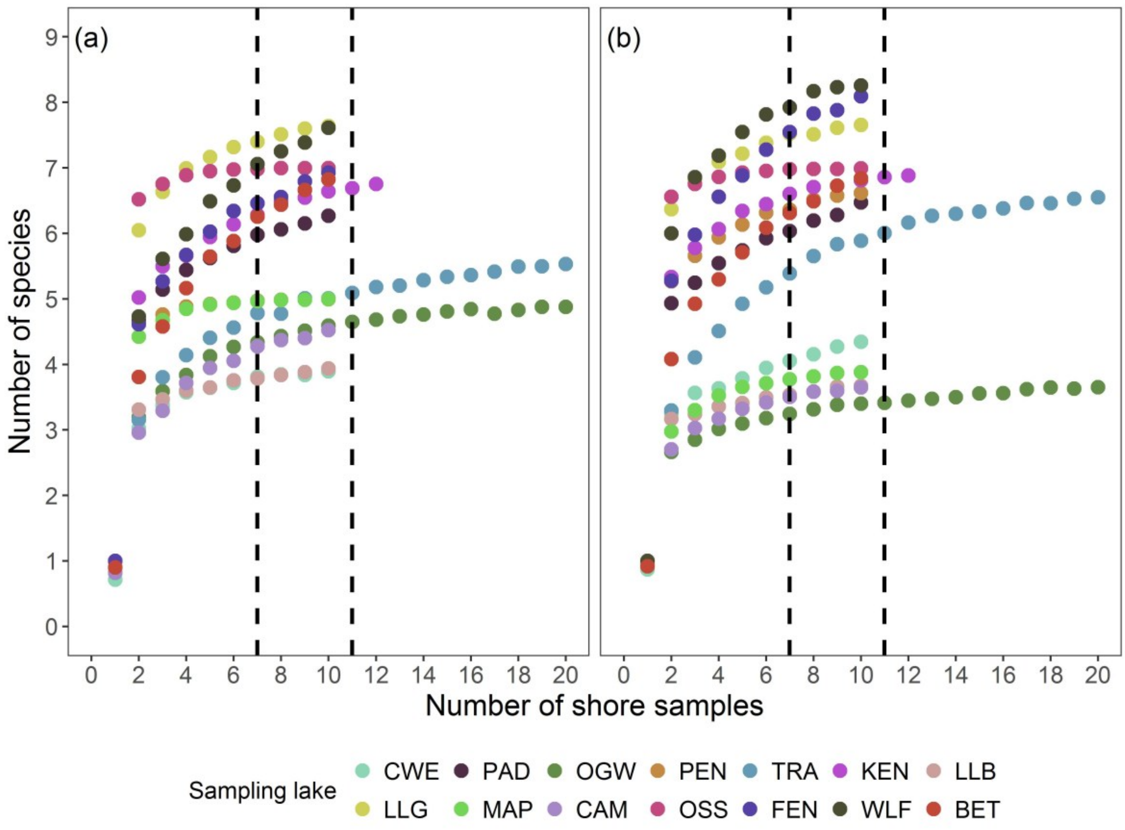


Figure 5.5 Sample-based rarefaction analyses of shore samples based on (a) Cytb and (b) 12S datasets. Sampling site codes are given in Figure 5.1.

5.4 Discussion

In this study, I have extended the geographical, ecological and taxonomic extent of the lake fish eDNA-based metabarcoding dataset, including 14 UK lakes with well-described fish faunas. This study confirms key results from our previous study (Hänfling et al. 2016) in that there is (1) a consistent, strong correlation between Cytb and 12S in terms of read counts and site occupancy and (2) a consistent, strong correlation between site occupancy and average read counts; and (3) site occupancy is consistently better than average read counts for estimating relative abundance, for both Cytb and 12S datasets. Moreover, eDNA metabarcoding outperforms established survey techniques in terms of species detection, relative abundance estimate using the standard five-level classification scale and characterisation ecological fish communities, suggesting eDNA metabarcoding has great potential as a fish-based assessment tool for the WFD lake status assessment.

5.4.1 Comparison of eDNA and existing data for species detection

A total of 73 species occurrences are recorded across the 14 lakes according to existing and historical fish data. Of these, 72 (98.6%) are detected with eDNA, which demonstrats that eDNA metabarcoding give comparable species richness estimates to collations of data using a range of sampling methods over date ranges spanning several decades. This result is consistent with previous studies that have demonstrated that eDNA metabarcoding produces a more comprehensive species list than alternative survey techniques with a similar effort in both marine and freshwater ecosystems (Hänfling et al. 2016; Port et al. 2016; Valentini et al. 2016).

Moreover, there is a significant positive correlation between historical data and eDNA data in terms of confidence of species presence. Species occurrences that are assessed as “probably or possibly present” (25/111) and “probably absent” (13/111) based on distribution data or anecdotal evidence also fitted the eDNA criteria for lowered confidence levels. In most cases, the “probably absent” eDNA occurrences are at very low site occupancy (≤ 0.1) and read counts (~0.5%), just above the threshold for accepting a positive record. These records could be genuine detections of species that have previously been missed. For example, perch were recorded with both Cytb and 12S in Llyn Cwellyn, Llyn Ogwen and Llyn Traffwll. The confidence of species present is either “established present” or “probably present” according to eDNA data. The recent appearance of this species in nearby lakes Llyn Padarn and Llyn Penrhyn (Appendix S5.2) suggest that the species is spreading in North Wales either by natural means or as a result of illegal introductions. In addition, the other "probably absent" eDNA occurrences could be either false positives from sequencing error, laboratory or environmental contamination. For instance, bleak *Alburnus alburnus* in Llyn Ogwen, goldfish *Carassius auratus* and gudgeon in Llyn Padarn, rudd in Chapel Mere, roach in Maer Pool, topmouth gudgeon *Pseudorasbora parva* in Betley Mere, and chub *Squalius cephalus* in Watch Lane Flash are most likely explained by low-level cross-contamination or sequencing barcode misassignment since these species are only detected by either Cytb or 12S in single sample of the lake.

Barcode misassignment and tag jumps have been proved in metabarcoding studies (e.g., Deakin et al. 2014; Schnell et al. 2015). To minimise contamination or false assignments, further work needs to be done to optimise each step, as well as the incorporation of robust controls to identify contamination when it occurs. Together with a growing number of studies (Stat et al. 2017; Li et al. 2018b), the results demonstrate the importance of using two (or more) markers for metabarcoding to avoid problems due to primer bias (e.g., reduced detection of nine-spined stickleback and gudgeon with Cytb; see more detail in Chapter 2 Section 2.3.3), gene copy number, PCR or sequencing artefacts (Schloss et al. 2011) and/or contamination. Other strategies such as using the consistency of presence across technical replicates as used by Port et al. (2016) might be a more suitable approach to control for false positives if rare species are of particular interest. Where verification of rare species detections is considered a high priority, additional confirmation from targeted species approaches (e.g., qPCR) and/or field surveys may be required.

5.4.2 Use of eDNA for assessing relative abundance of fish

WFD specifies that abundance should be considered when determining ecological status; hence, current WFD approaches include estimates of abundance (often as abundance classes). For instance, a five-level scale, adapted from DAFOR scale, is accepted for Austrian standard and national monitoring techniques applied under WFD for surveying aquatic macrophytes (Pall & Moser 2009). DAFOR scale is used to estimate fish relative abundance in the present study based on DNA-based identification and facilitate integration into current WFD approaches.

There are consistently positive correlations in terms of relative abundance between the eDNA data and historical data. However, these correlations are not always statistically significant. The correlations are not significant in Llyn Traffwll and Llyn Penrhyn probably because three-spined stickleback are more abundant based on eDNA data than expected. This species is often under-represented or overlooked in surveys using established fish capture methods (Hänfling et al. 2016). Moreover, tench are not detected by eDNA in Llyn Penrhyn. Llan Bwch-llyn is a species-poor lake with only four species detected, which could reduce the statistical power. The non-significant correlation in Watch Lane Flash could be attributed to under-representation of three-spined stickleback in previous fish surveys and the reduced detection probability of gudgeon with Cytb. Although these results are generally encouraging, further work is critical to obtain enough statistical power to directly test the relationship between abundance estimates from eDNA and surveys using other methods. It is also important to investigate taxon-specific detection probabilities and abundance estimates so that a pressure-sensitive tool can be developed.

5.4.3 Using eDNA to describe ecological communities

According to environmental characteristics, there are three pre-defined lake types. Encouragingly, four distinct community types could be identified based on clustering dendrograms and NMDS ordinations using eDNA data. Basically, the Community 1 and Community 2 are agreed with the pre-defined lake Type 1 and Type 2, respectively. The pre-defined coarse fish lakes (Type 3) can be further divided into Community 3 and Community 4 based on whether bream is a dominant species in the lake. These findings indicated that eDNA metabarcoding has great potential as a fish-based assessment tool for WFD lake status assessment.

Specifically, the results of this study showed that the low alkalinity lakes within a predominantly upland catchment are mainly dominated by salmonids reflecting their relatively deep and oligotrophic nature. Except salmonids, all these sites containe minnow, most likely introduced by anglers using them as live bait (Hatton-Ellis 2005). Compared to the salmonids and minnow lakes (Community 1), perch, pike and cyprinids (such as bream, common carp, roach and tench) are prevalent in the coarse fish lakes (Community 3 & Community 4) reflecting their relatively shallow and eutrophic nature. These findings support that eDNA profiles are suitable to reflect the eutrophic conditions of Lake Windermere in which species that prefer less eutrophic conditions (Atlantic salmon, brown trout, Arctic charr, minnow and bullhead) are more abundant in the mesotrophic North Basin, and species associated with eutrophic conditions (roach, tench, rudd, bream and European eel) are more common in the eutrophic South Basin (Hänfling et al. 2016). Furthermore, both of the diadromous fish lakes (Community 2) containe a varied fauna including various combinations of mostly diadromous species such as European eel, three-spined stickleback, nine-spined stickleback, and brown trout. Alongside these, several coarse fish species such as roach, perch, and rudd also occurr in these lakes. However, of greater conservation interest is the distribution of nine-spined stickleback. This species occurs only locally in Wales, mainly at lowland sites close to the sea (Davies et al. 2004; Hatton-Ellis 2005).

5.4.4 Shoreline and boat-based sampling

Offshore sampling is expensive as it requires dedicated additional site access, the use of a boat, trained staff and capital costs. Sampling from the shore would obviously alleviate some of the sampling costs and reduce potential contamination from transferring boats. Our previous study showed that 12 out of the 16 previously recorded species are detected in only six shoreline samples via eDNA metabarcoding (Hänfling et al. 2016). This suggests eDNA could accumulate on the littoral zone and that shoreline sampling could be adequate for the detection of most species (Hänfling et al. 2016). In the present study, the rigorous sampling of shore and offshore locations is carried out to compare the suitability of the different spatial sampling location. The number of species detected in the shore samples is equal to, or slightly higher than in the offshore samples. Moreover, consistent and significant correlations are observed between shore and offshore samples in terms of site occupancy. From the perspective of each species, there are also consistently positive correlations between shore and offshore samples. Remarkably, Arctic charr and Atlantic salmon, deep water species, are present in both shore and offshore samples with similar site occupancy in Llyn Cwellyn and Llyn Padarn. This may result from eDNA dispersing more widely in winter (when these lakes were sampled) due to increased water movement, a lack of stratification and/or slower temperature-related degradation of eDNA. This result supports our previous results in which Arctic charr and Atlantic salmon are detected in both shoreline and offshore samples during winter sampling campaign in Windermere; however, these two species are only found in offshore samples during the summer (Lawson Handley et al. 2019). Overall, the results from rarefaction analyses with the shoreline samples are broadly consistent with Hänfling et al. (2016), indicating that more species are detected with the same number of samples for 12S than Cytb and approximately 10 shore samples detects the majority (≥85%) of species.

Although shoreline sampling is adequate for capturing the majority of species in winter, there are important differences between sampling locations for some species such as common carp showing higher abundance in offshore samples. More importantly, restricting sampling to the shoreline would potentially skew abundance estimates in favour of littoral and benthic species (Lawson Handley et al. 2019). If abundance estimates are required for all fish species, in large lakes, a combination of both shoreline and offshore sampling is recommended. It should also be noted that 10 shore samples may be adequate for species detection, while abundance estimates based on site occupancy require more comprehensive spatial sampling.

5.5 Conclusions

The present study further demonstrate that eDNA metabarcoding has great power in detecting fish species when compared against data from a heterogeneous range of sources obtained during the past 30 years for Welsh lakes, and for Cheshire meres from PASE and seine netting survey results. Four eDNA communities are characterised in this study, which is consistent with earlier assessments and ecological interpretations. I propose a five-level DAFOR scale to estimate fish relative abundance. There are consistently positive correlations in terms of relative abundance between the eDNA data and historical data. This is an exceptional level of performance and provides strong grounds for further development. In theory, 10 shore samples are adequate for capturing the majority of species. To better estimate abundance of fish living in different habitats, in large lakes, a combination of both shore and offshore sampling is recommended. In conclusion, eDNA metabarcoding shows great promise as a monitoring tool for fish in a wide range of lake types, outperforming conventional survey methods in terms of species detection and consistently providing credible results. Further optimisation and development such as understanding taxon-specific biases, abundance estimates and management of low confident occurrence are strongly recommended with a larger dataset, to improve the accuracy, effectiveness and applicability of WFD monitoring tool.

5.6 Supporting Information

Appendix S5.1 Supplementary tables and figures

Table S5.1 Key environmental descriptors of the 14 sampled lakes.

Lake Name	Grid Reference	Lake Area (ha)	Mean Depth (m)	Depth Type	Altitude Type	Alkalinity Type	Humic Type
Llyn Cwellyn	SH55975492	90	22.6	Deep	Mid	Low	Clear
Llyn Padarn	SH56986145	98	15.9	Deep	Mid	Low	Clear
Llyn Ogwen	SH65906044	39	2.0	Very Shallow	Mid	Low	Clear
Llyn Penrhyn	SH31327689	22	2.2	Very Shallow	Low	High	Clear
Llyn Traffwll	SH32577696	37	2.5	Very Shallow	Low	High	Clear
Kenfig Pool	SS79688155	29	2.0	Very Shallow	Low	High	Clear
Llan Bwch-llyn	SO11924634	10	3.2	Shallow	High	High	Clear
Llangorse Lake	SO13222649	140	2.0	Very Shallow	Low	High	Clear
Maer Pool	SJ78923845	5	2.8	Very shallow	Low	High	-
Chapel Mere	SJ53985183	9	1.0	Very shallow	Low	High	Humic
Oss Mere	SJ56604389	11	1.2	Very shallow	Low	High	Humic
Fenemere	SJ44552290	9	0.9	Very shallow	Low	High	Clear
Watch Lane Flash	SJ72816060	8	3.3	Shallow	Low	High	-
Betley Mere	SJ74874794	9	0.7	Very shallow	Low	High	Humic

Table S5.2 Sampling and filtration dates and sampling strategy.

Lake Name	**Sampling Date**	**Filtration Date**	**Sampling Depth**	**Sample number**
Llyn Cwellyn	13/01/2016	14/01/2016	0-30m	10 shore and 10 offshore samples
Llyn Padarn	13/01/2016	14/01/2016	0-25m	10 shore and 10 offshore samples
Llyn Ogwen	11/01/2016	12/01/2016	Surface	20 shore samples
Llyn Penrhyn	14/01/2016	15/01/2016	Surface	10 shore samples
Llyn Traffwll	15/01/2016	16/01/2016	Surface	20 shore samples
Kenfig Pool	11/07/2016	12/07/2016	Surface	12 shore samples
Llan Bwch-llyn	12/07/2016	13/07/2016	Surface	10 shore samples
Llangorse Lake	12/07/2016	13/07/2016	Surface	10 shore and 10 offshore samples
Maer Pool	05/01/2016	05/01/2016	Surface	10 shore and 10 offshore samples
Chapel Mere	07/01/2016	08/01/2016	Surface	10 shore and 10 offshore samples
Oss Mere	03/12/2015	04/12/2015	Surface	10 shore and 10 offshore samples
Fenemere	08/12/2015	09/12/2015	Surface	10 shore and 10 offshore samples
Watch Lane Flash	01/12/2015	01/12/2015	Surface	10 shore and 10 offshore samples
Betley Mere	05/01/2016	06/01/2016	Surface	10 shore and 10 offshore samples

Table S5.3 Individuals density (ind. ha^{-1}) of Cheshire meres from fisheries survey by the Ecological Consultancy Ltd in 2016.

Species	**MAP**	**CAM**	**OSS**	**FEN**	**WLF**	**BET**
Abramis brama	-	-	569.04	10.57	525.31	431.9
Anguilla anguilla	-	-	-	5.57	-	-
Cyprinus carpio	106.21	-	3811.34	20.99	7.9	43.19
Esox lucius	18.01	48.35	21.77	41.97	6.58	43.19
Gasterosteus aculeatus	-	-	-	21.51	-	-
Gobio gobio	-	-	-	-	539.48	-
Leucaspius delineatus	-	-	-	-	1899.41	-
Perca fluviatilis	1202.48	1273.32	2007.33	8795.86	1955.05	237.54
Rutilus rutilus	-	-	5833.44	1655.62	487.38	712.63
Scardinius erythrophthalmus	-	-	777.11	234.47	-	-
Tinca tinca	4188.83	322.36	3769.32	-	-	86.38

Notes: Sampling lake codes are given in Figure 5.1.

Table S5.4 Biomass density (kg ha^{-1}) of Cheshire meres from fisheries survey by the Ecological Consultancy Ltd in 2016.

Species	**MAP**	**CAM**	**OSS**	**FEN**	**WLF**	**BET**
Abramis brama	-	-	9.51	11.63	81.76	39.04
Anguilla anguilla	-	-	-	3.90	-	-
Cyprinus carpio	10.56	-	65.96	83.94	44.80	70.47
Esox lucius	22.43	67.91	21.77	21.60	16.45	72.45
Gasterosteus aculeatus	-	-	-	1.02	-	-
Gobio gobio	-	-	-	-	6.20	-
Leucaspius delineatus	-	-	-	-	3.93	-
Perca fluviatilis	24.70	7.53	9.83	25.20	9.28	3.64
Rutilus rutilus	-	-	18.99	19.79	16.91	1.72
Scardinius erythrophthalmus	-	-	2.37	3.29	-	-
Tinca tinca	44.96	6.16	9.97	-	-	137.41

Notes: Sampling lake codes are given in Figure 5.1.

Table S5.5 Analysis of similarities (ANOSIM) comparisons of different community types based on read counts for Cytb and 12S datasets.

Comparison	**Cytb**		**12S**		**Mean ± *SD***	
	R	***p***	***R***	***p***	***R***	***p***
Global test	0.765	0.001	0.734	0.001	0.750 ± 0.016	0.001
Community 1 within	0.518	0.001	0.338	0.001	0.428 ± 0.090	0.001
Community 2 within	0.269	0.004	0.224	0.004	0.247 ± 0.023	0.004
Community 3 within	0.365	0.001	0.373	0.001	0.369 ± 0.004	0.001
Community 4 within	0.257	0.001	0.251	0.001	0.254 ± 0.003	0.001
Community 1: Community 2	0.692	0.001	0.768	0.001	0.730 ± 0.038	0.001
Community 1: Community 3	0.967	0.001	0.969	0.001	0.968 ± 0.001	0.001
Community 1: Community 4	0.946	0.001	0.936	0.001	0.941 ± 0.005	0.001
Community 2: Community 3	0.583	0.001	0.545	0.001	0.564 ± 0.019	0.001
Community 2: Community 4	0.753	0.001	0.619	0.001	0.686 ± 0.067	0.001
Community 3: Community 4	0.570	0.001	0.431	0.001	0.501 ± 0.070	0.001

Notes: Community types of sampling lake are the same as in Table 5.5. R values were derived from Bray-Curtis dissimilarity matrices.

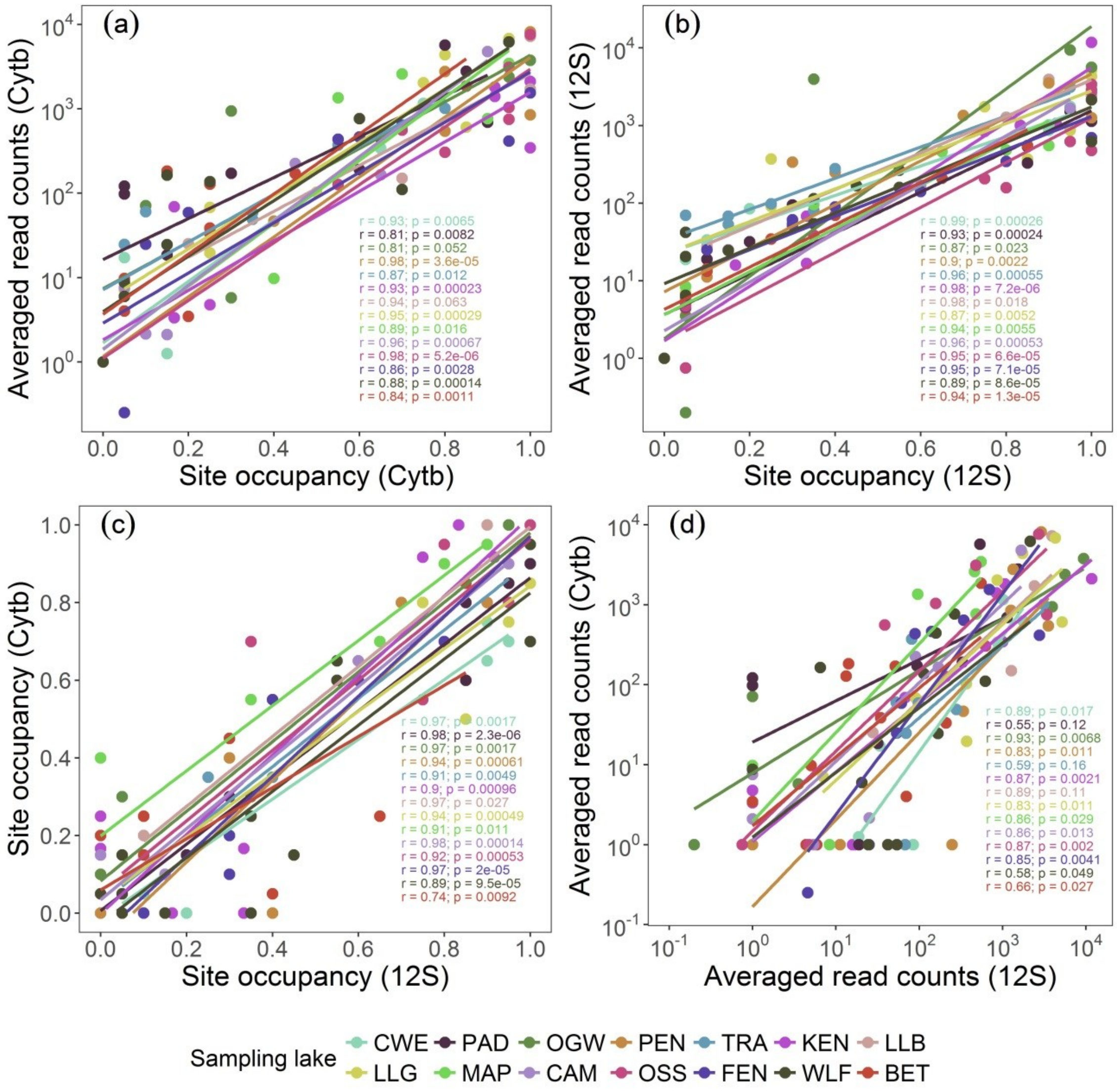


Figure S5.1 Correlations between averaged read counts and site occupancy across Cytb and 12S datasets. Sampling lake codes are given in Figure 5.1.

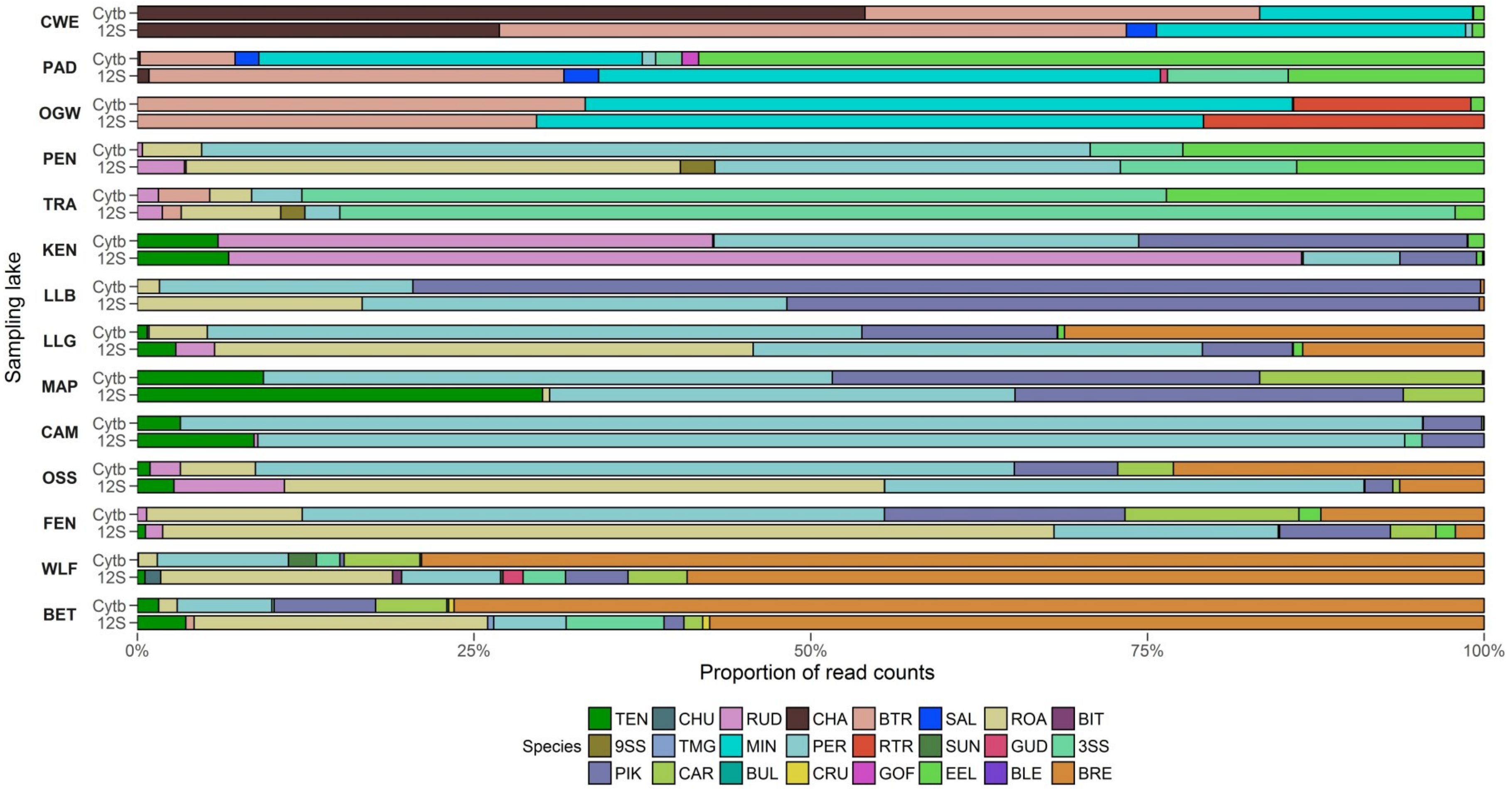


Figure S5.2 Species composition of read counts for Cytb and 12S datasets. Species three letter codes and sampling lake codes are given in Table 5.5 and Figure 5.1, respectively.

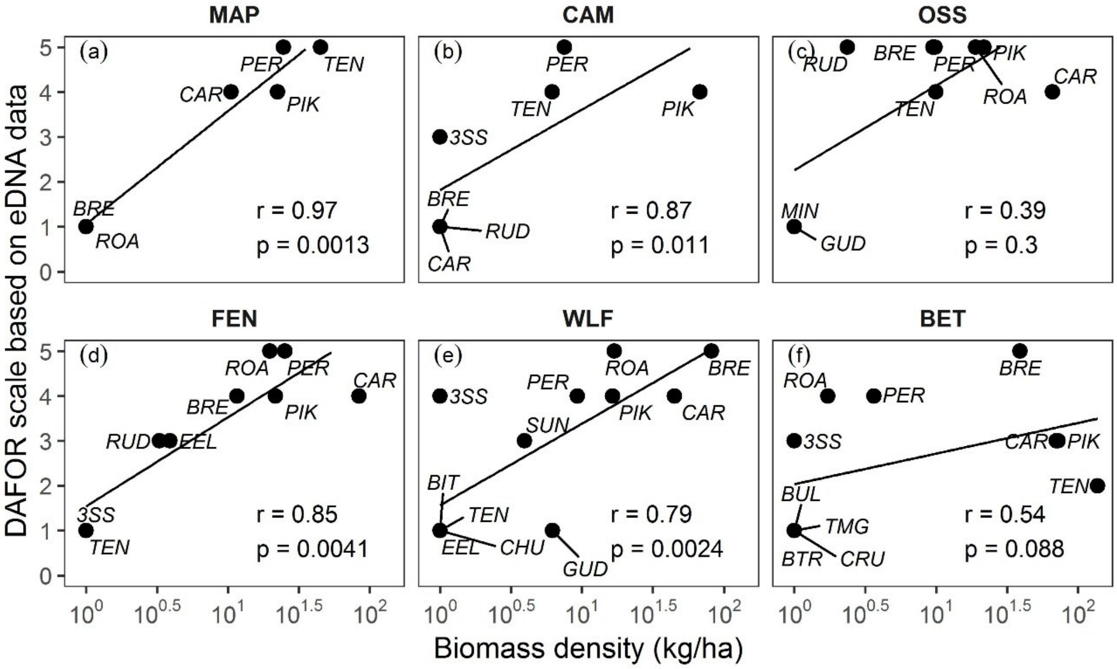


Figure S5.3 Correlations between DAFOR scale based on eDNA data and biomass density from the Ecological Consultancy Ltd survey of Cheshire meres. Sampling lake codes are given in Figure 5.1; Abundance scores to DAFOR scale and species three letter codes are given in Table 5.4 and Table 5.5, respectively.

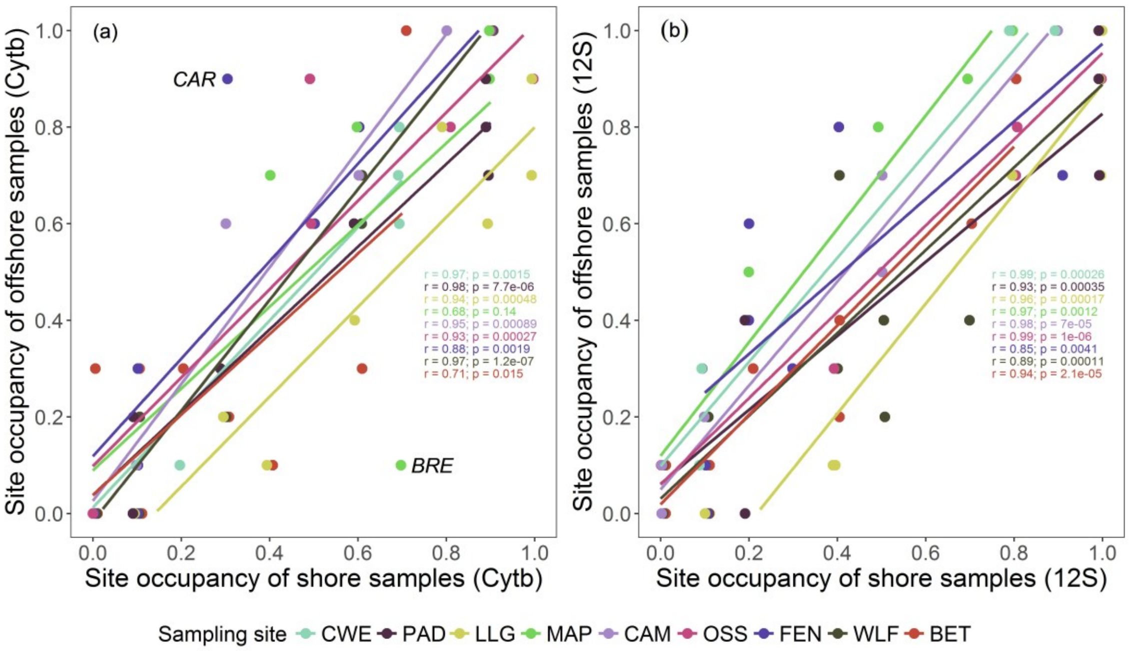


Figure S5.4 Correlation of site occupancy between two sampling locations (offshore and shore) in terms of sampling sites based on (a) Cytb and (b) 12S datasets. Species three letter codes and sampling site codes are given in Table 5.5 and Figure 5.1, respectively.

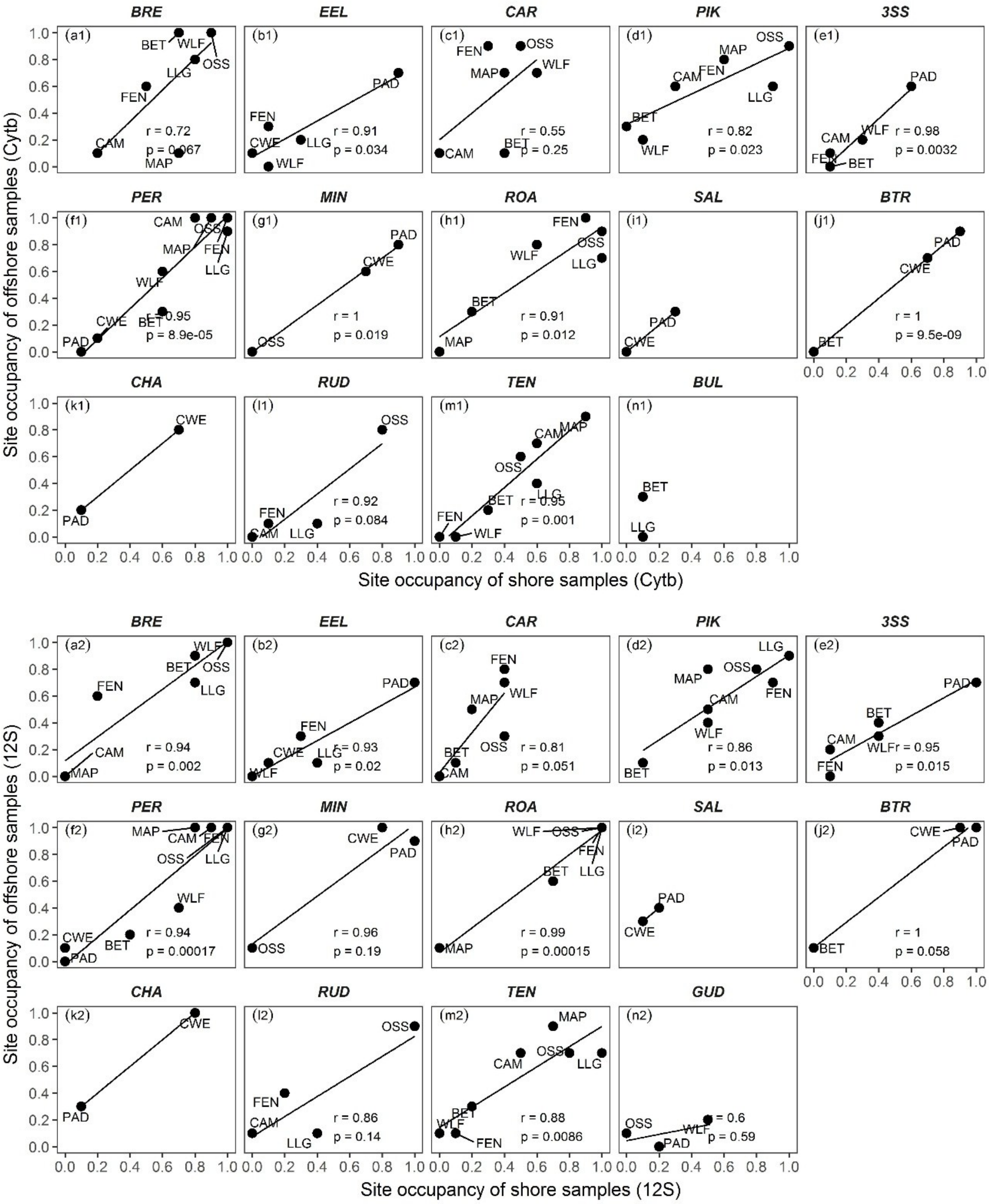


Figure S5.5 Correlation of site occupancy between two sampling locations (offshore and shore) in terms of species based on Cytb and 12S datasets. Species with only one occurrence is not included. Species three letter codes and sampling site codes are given in Table 5.5 and Figure 5.1, respectively.

Appendix S5.2 Collation of fish data from historical data

S5.2.1 Llyn Cwellyn historical fish data

The expected fish fauna of Llyn Cwellyn is shown in Table S5.6. Five species are known from the lake (European eel, minnow, Atlantic salmon, brown trout, and Arctic charr), but several other diadromous species and three-spined stickleback may also be present. The lake is generally dominated by salmonids reflecting its relatively deep and oligotrophic nature, and is notable for containing one of a small number of natural Arctic charr populations in North Wales (Child 1977; Milner 1983; McCarthy 2007).

The Afon Gwyrfai connects the river to the Menai Strait approximately 11 km away, and there are no serious barriers to migration although a structure controlling water levels at the lake outflow may limit access for species such as flounder. Due to its location and physico-chemical characteristics, the Gwyrfai catchment was not colonised by coarse fish species from continental Europe, and therefore the natural fish fauna consists only of diadromous and glacial relict species. The oligotrophic and exposed nature of the lake also makes it generally unsuitable for many coarse fish.

Table S5.6 Expected fish fauna and relative species abundance of Llyn Cwellyn.

Species	**Presence**	**Abundance**	**Notes**
Anguilla anguilla	E	O	Recorded in seine netting
Phoxinus phoxinus	E	F	Detected by electrofishing
Salmo salar	E	O	Diadromous in catchment, mainly rheophilic
Salmo trutta	E	A	Diadromous in catchment and resident in lake
Salvelinus alpinus	E	D	Well documented population, long-standing records
Gasterosteus aculeatus	PoP	R	No records; connected to sea
Lampetra fluviatilis	PoP	R	No records; connected to sea
Lampetra planeri	PoP	R	No records; connected to sea
Petromyzon marinus	PoP	R	No records; connected to sea
Platichthys flesus	PoP	R	No records; connected to sea though access is relatively poor

Notes: Presence categories are described in Table 5.1. The relative abundance DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare).

S5.2.2 Llyn Padarn historical fish data

The expected fish fauna and relative species abundance of Llyn Padarn is shown in Table S5.7. Seven species are known from the lake (European eel, three-spined stickleback, perch, minnow, Atlantic salmon, brown trout and Arctic charr). The lake is generally dominated by salmonids reflecting its relatively deep and oligotrophic nature, and is notable for containing one of a small number of natural Arctic charr populations in North Wales (McCarthy 2007). A large number of Arctic charr surveys have been carried out due to concerns about the decline of this population (Child 1977; Hanks 1998; Hanks 2003; Clabburn et al. 2011; Clabburn & Griffiths 2013; Clabburn et al.

2016), but structured surveys of other species are limited. Perch have recently been recorded from the lake for the first time (Jones H.P. pers. comm.)

The Afon Seiont connects the river to the Menai Strait approximately 10 km away, and there are no significant barriers to migration. Due to its location and environmental characteristics facing the Irish Sea, the Seiont catchment was not colonised by coarse fish species from continental Europe, and therefore the natural fish fauna consists only of diadromous and glacial relict species.

Table S5.7 Expected fish fauna and relative species abundance of Llyn Padarn.

Species	**Presence**	**Abundance**	**Notes**
Anguilla anguilla	E	A	Multiple records, 1966-2007
Gasterosteus aculeatus	E	A	Recorded as abundant in netting survey
Phoxinus phoxinus	E	A	Records between 1966 and 2013
Salmo salar	E	O	Diadromous in catchment, mainly rheophilic
Salmo trutta	E	D	Diadromous in catchment and resident in lake
Salvelinus alpinus	E	O	Well documented population, recent data indicate a decline
Perca fluviatilis	PrP	R	Recently recorded in the lake for the first time
Lampetra fluviatilis	PoP	R	No records; connected to sea
Lampetra planeri	PoP	R	Unidentified lamprey presumed to be this species recorded in 2005
Petromyzon marinus	PoP	R	No records; connected to sea
Platichthys flesus	PoP	R	No records; connected to sea

Notes: Presence categories are described in Table 5.1. The relative abundance DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare).

S5.2.3 Llyn Ogwen historical fish data

Llyn Ogwen is the highest altitude lake in this study, a base-poor shallow, stony lake at the head of the Nant Ffrancon valley in the upper Ogwen catchment. The lake is isolated from the sea by the Rhaeadr Ogwen falls, preventing access by all migratory fish except possibly European eel. Consequently, the lake is very species-poor, with the only species being brown trout, minnow, rainbow trout, and possibly a few European eel (Table S5.8).

Table S5.8 Expected fish fauna and relative species abundance of Llyn Ogwen.

Species	**Presence**	**Abundance**	**Notes**
Oncorhynchus mykiss	E	A	Regularly stocked by the local angling club
Phoxinus phoxinus	E	A	Widespread in the catchment
Salmo trutta	E	A	Angler record
Anguilla anguilla	PoP	R	Doubtfully present, Ogwen falls is a significant obstacle but a few eels may successfully reach the lake

Notes: Presence categories are described in Table 5.1. The relative abundance DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare).

S5.2.4 Llyn Penrhyn historical fish data

Llyn Penrhyn is one of three alkaline shallow lakes on the west coast of Anglesey in this study. The lake is close to the sea, but access for migratory fish may be hindered by the sluice used to maintain water levels in summer by the Royal Society for the Protection of Birds (RSPB). Llyn Penrhyn has been well studied due to a lengthy history of eutrophication from sewage effluent (Duigan et al. 1996), and this has included various fish surveys (Bray 1995; White 2000). The lake is an RSPB reserve, and fish surveys have also been carried out to determine the quality of the lake and its fringing reedbeds for bittern (Self & Muirhead 2003; Self & Lyons 2007; Self 2016).

Six species (European eel, three-spined stickleback, perch, roach, rudd, and tench) are considered established in the lake; although the community is dominated by European eel and perch (Table S5.9).

Table S5.9 Expected fish fauna and relative species abundance of Llyn Penrhyn.

Species	**Presence**	**Abundance**	**Notes**
Anguilla anguilla	E	D	High densities recorded using fyke nets
Gasterosteus aculeatus	E	R	Recorded in 2005 and 2007
Perca fluviatilis	E	A	Multiple records and abundant in samples
Rutilus rutilus	E	O	Present probably at low densities overall although large shoals have been noted in some locations
Scardinius erythrophthalmus	E	O	Multiple records, present at low densities
Tinca tinca	E	O	Stocked to the lake at low densities; two tench were recorded by the 2016 RSPB survey and small numbers have also been detected in previous years
Platichthys flesus	PoP	R	Close to the sea with good accessibility
Pungitius pungitius	PoP	R	Present in nearby and directly connected Llyn Dinam; habitat

			suitable
Salmo trutta	PoP	R	Old records but doubtful if still present. Little spawning habitat

Notes: Presence categories are described in Table 5.1. The relative abundance DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare).

S5.2.5 Llyn Traffwll historical fish data

Another western Anglesey coastal lake, close to Llyn Penrhyn, Llyn Traffwll would originally have had a species-poor fauna dominated by diadromous species (European eel, three-spined stickleback, and brown trout). The lake has no inflows, and a short, sluggish tributary of the Afon Crigyll connects the lake to the sea, a distance of about 4 km.

However, several coarse fish species were introduced during the mid-20th century, and at least roach has persisted. Eutrophication has also rendered the habitat less suitable for brown trout, and it is unclear whether this species still occurs in the lake. The lake has been surveyed by RSPB in support of their bittern conservation programme (Self & Muirhead 2003; Self & Lyons 2007).

Only three species have recently been recorded: European eel (which occurs at high densities) and small quantities of roach and three-spined stickleback. However, it is possible that five other species (European flounder *Platichthys flesus*, nine-spined stickleback, brown trout, rudd, and tench) inhabit the lake (Table S5.10). It should be noted that fish surveys at this site mainly concentrate on open areas in reedbeds and may not be representative of the entire lake.

Table S5.10 Expected fish fauna and relative species abundance of Llyn Traffwll.

Species	Presence	Abundance	Notes
Anguilla anguilla	E	D	Regularly recorded by fyke netting
Gasterosteus aculeatus	E	O	Present but at low abundance
Rutilus rutilus	E	O	Detected via electrofishing
Platichthys flesus	PoP	R	Close to the sea and expected to be accessible
Pungitius pungitius	PoP	R	No records, but accessible by sea, habitat is suitable and present in similar nearby lakes
Salmo trutta	PoP	R	Old records and habitat suitable for adults but doubtful if there is sufficient spawning habitat
Scardinius erythrophthalmus	PoP	R	Possibly present; recorded from nearby lakes and suitable habitat
Tinca tinca	PoP	R	Recorded from nearby lakes and habitat is suitable

Notes: Presence categories are described in Table 5.1. The relative abundance DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare).

S5.2.6 Kenfig Pool historical fish data

Kenfig Pool is a shallow lake at the head of the Kenfig Burrows dune system in Glamorgan. The lake is groundwater fed and lacks either substantial inflows or outflows, so is naturally very species-poor. However, it has been used as a fishery by the local angling club for decades, and during this time various species have been stocked or introduced. A number of surveys have been carried out (Favager 1997; Giles 2003; Hatton-Ellis 2005), though methods used have not been consistent.

The current fauna is known to consist of European eel, pike, perch, rudd, and tench (Giles 2003) but a few common carp may still persist in the lake, the legacy of an earlier introduction. Roach has also previously been recorded, and either or both of the stickleback species may also be present (Table S5.11).

Table S5.11 Expected fish fauna and relative species abundance of Kenfig Pool.

Species	**Presence**	**Abundance**	**Notes**
Anguilla anguilla	E	A	Regularly recorded; an eel fisherman has rights to fish the lake using fyke nets
Esox lucius	E	F	Includes some large individuals
Perca fluviatilis	E	D	Stocked to the lake; multiple records
Scardinius erythrophthalmus	E	A	Stocked to the lake; multiple records
Tinca tinca	E	O	Stocked to the lake at low densities
Cyprinus carpio	PrP	R	Doubtfully present. Stocked to the lake during the 1980s and was recorded in the early 2000s, but they do not seem to reproduce and are in poor condition
Pungitius pungitius	PoP	R	Recorded in 1977 but often overlooked
Rutilus rutilus	PoP	R	Recorded in 1977 but may no longer be present

Notes: Presence categories are described in Table 5.1. The relative abundance DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare).

S5.2.7 Llan Bwch-llyn historical fish data

Llan Bwch-llyn is a shallow alkaline lake at the head of the Bachawy, a tributary of the River Wye in mid-Wales. This is a naturally species-poor lake because access for most migratory and rheophilic species in the Wye catchment is blocked by a natural falls low down on the Bachawy. Llan Bwch-llyn is managed as a fishery and has been stocked with coarse fish in the past, but no stocking has taken place for around 10 years.

There is no survey data from the lake, but the fish fauna is reasonably well-known from angler records and consists of just three species: pike, perch, and roach (Table S5.12). Bream was formerly stocked to the lake but none have been caught recently, and the population may have died out. It is possible that some other species are present but not recorded.

Table S5.12 Expected fish fauna and relative species abundance of Llan Bwch-llyn.

Species	**Presence**	**Abundance**	**Notes**
Esox lucius	E	O	Angler records
Perca fluviatilis	E	O	Angler records
Rutilus rutilus	E	O	Angler records
Abramis brama	PrP	R	Previously stocked but no recent angler catches

Notes: Presence categories are described in Table 5.1. The relative abundance DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare).

S5.2.8 Llangorse Lake historical fish data

Llangorse Lake has potentially the most diverse fish fauna of any of the lakes in this study, with seven species considered established and a further sixteen species possibly occurring (Table S5.13). This reflects the location of the lake in south-east Wales where the fish fauna is naturally more diverse, the fact that the lake lies in the Wye catchment and hence has good connectivity for rheophilic and migratory species, and that the lake habitat is suitable for a relatively wide range of species. The lake has been surveyed by NRW and its predecessor EA for the EU Water Framework Directive purposes. Bream, European eel, pike, perch, roach, rudd and tench are all present or highly likely to occur. A wide variety of other species may occur including brown trout, Atlantic salmon, bleak, silver bream *Blicca bjoerkna*, bullhead, common carp, dace *Leuciscus leuciscus*, three-spined stickleback, all three lamprey species (river lamprey *Lampetra fluviatilis*, European brook lamprey *Lampetra planeri*, sea lamprey *Petromyzon marinus*), minnow, chub, and grayling *Thymallus thymallus*.

Table S5.13 Expected fish fauna and relative species abundance of Llangorse Lake.

Species	**Presence**	**Abundance**	**Notes**
Abramis brama	E	O	Seems to be rather scarce
Anguilla anguilla	E	F	Multiple records and subject to a stocking programme
Esox lucius	E	F	Well-known from the lake, a large specimen attacked a water-skier in 1999
Perca fluviatilis	E	D	Probably the most abundant fish species in the lake
Rutilus rutilus	E	O	Several records including from recent fyke netting
Scardinius erythrophthalmus	E	O	Recorded in 2016 fyke netting
Tinca tinca	E	O	Angler record
Alburnus alburnus	PoP	R	Recorded in 1978
Blicca bjoerkna	PoP	R	Roach x this species hybrid recorded in 1997
Cottus gobio	PoP	R	Widespread in the connected

			Wye catchment; 1978 record from the lake. Population could occur in exposed shore areas
Cyprinus carpio	PoP	R	1978 record. However not recently stocked and doubtful if still extant
Gasterosteus aculeatus	PoP	R	No records; connected to sea
Lampetra fluviatilis	PoP	R	No records; common in catchment, connected to sea
Lampetra planeri	PoP	R	No records; very common in catchment
Leuciscus leuciscus	PoP	R	In outflow
Petromyzon marinus	PoP	R	No records; connected to sea
Phoxinus phoxinus	PoP	R	1978 record and widespread in the catchment
Platichthys flesus	PoP	R	No records; connected to sea though access is poor
Salmo salar	PoP	R	Diadromous in catchment, mainly rheophilic
Salmo trutta	PoP	R	Although coarse fish dominated, there may be trout present
Squalius cephalus	PoP	R	Rheophilic; widespread in catchment and 1978 record from inflow
Thymallus thymallus	PoP	R	In outflow

Notes: Presence categories are described in Table 5.1. The relative abundance DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare).

S5.2.9 Maer Pool historical fish data

Maer Pool is a very small and shallow mere with high nutrient levels due to diffuse water pollution. Relatively little is known about the fish community of Maer Pool with just one survey undertaken by APEM in 2009 prior to the present Ecological

Consultancy Ltd (ECON) survey, with similar species found, though it is notable that neither roach nor bream have previously been caught or stocked into Maer Pool. There are no waters obviously connected to Maer Pool that would be a likely source of these albeit relatively common species. It is possible that they have been introduced without being recorded.

S5.2.10 Chapel Mere historical fish data

There is very limited information on this site other than the current ECON survey. Bream and possibly common carp were said to have been present historically, but there is no record of rudd either being present or stocked.

S5.2.11 Oss Mere historical fish data

Gudgeon are quite possibly present in small numbers though there is no reference to them in any of the supplementary information in the site report. Earlier surveys by Brian Moss in 1994 suggested crucian carp *Carassius carassius* were present in addition to the species found in the ECON survey, but these were not found in surveys by APEM in 2009 and 2011.

S5.2.12 Fenemere historical fish data

Fenemere is a shallow, enriched lake with significant macrophyte growth. There have been a number of surveys of Fenemere as far back as 1994. In addition to the species found in the current ECON survey, crucian carp have also been reported, and tench were historically present. Hence the eDNA record of tench may reflect a small relict population of this species which is likely to have been both native and stocked to this site in the past.

S5.2.13 Watch Lane Flash historical fish data

Watch Lane Flash has poor water quality with few macrophytes and is brackish. It appears to be connected via small watercourses to the river Wheelock and the river Dane and is therefore theoretically accessible to European eel, though as this is well away from the sea, they are probably present in only very small numbers. European bitterling *Rhodeus amarus* could be present as they are present in the adjacent Trent and Mersey canal and if present in only small numbers could easily have been missed by the conventional survey methods especially as electro-fishing was not used at this site due to high water conductivity. There is some previous survey information, EA undertook two previous surveys in recent years in Isle Pool and Watch Lane Flash and found, in addition to the species found in the current ECON survey, goldfish, crucian carp and Wels catfish *Siluris glanis* though these latter were removed after capture. No chub have ever been recorded, but it is possible that some have been introduced illegally. It is highly likely that three-spined stickleback could be present due to suitable habitats; however, this species could be missed by all methods if it is present only at very low abundance.

S5.2.14 Betley Mere historical fish data

There have been two previous surveys undertaken on Betley Mere in 1979 and 2009. In addition to the species found in the current ECON survey, European eel, white bream, ruffe *Gymnocephalus cernua*, rudd, and brown trout have also been recorded. Brown trout are less certain since although they have been stocked, they would almost certainly not breed in the mere and they are not usually long-lived. It is possible, in the absence of further information, that there might be a small brown trout population in the in- or out-flowing streams. Bullhead have not been recorded in any other survey but may not be particularly vulnerable to capture methods and are rarely caught by angling, and could, in any case, exist in in- and out-flowing streams. It is highly likely that three-spined stickleback could be present due to suitable habitats; however, this species could be missed by all methods if it is present only at very low abundances. Topmouth Gudgeon have never been recorded here, and there are no obvious links with anywhere where they have been present, though there have been waters in the region where they did exist but have been eradicated.

Appendix S5.3 Read counts of OTUs data for the Cytb dataset was used for the R script (.csv; supplied in a separate file)

The file can be viewed or downloaded use the link as below:

https://github.com/HullUni-bioinformatics/Li_et_al_2019_eDNA_fish_monitoring/blob/master/Appendix_S3_Cytb.csv

Appendix S5.4 Read counts of OTUs data for the 12S dataset was used for the R script (.csv; supplied in a separate file)

The file can be viewed or downloaded use the link as below:

https://github.com/HullUni-bioinformatics/Li_et_al_2019_eDNA_fish_monitoring/blob/master/Appendix_S4_12S.csv

Appendix S5.5 The generalised linear model (GLM) results of Welsh lake dataset

Model	**Residual deviance**	**Akaike information criterion (AIC)**	**Note**
glm(DAFOR ~ SO + Species, data)	77.75	319.89	The best model
glm(DAFOR ~ RRC + Species, data)	128.34	377.02	
glm(DAFOR ~ SO + RRC + Species, data)	77.46	321.46	*p* = 0.55 for 'RRC'
glm(DAFOR ~ SO + Lake + Species, data)	69.46	321.03	No significant difference under 'Lake'
glm(DAFOR ~ SO + Locus + Species, data)	77.43	321.43	No significant difference under 'Locus'
lmer(DAFOR ~ SO + (1|Species))	329.40	339.50	
lmer(DAFOR ~ RRC + (1|Species))	382.8	390.10	
lmer(DAFOR ~ SO + RRC + (1|Species))	328.5	339.00	
lmer(DAFOR ~ SO + (1|Lake) + (1|Species))	329.40	341.50	Intercept under 'Lake' was 0.00
lmer(DAFOR ~ SO + (1|Locus) + (1|Species))	329.40	341.50	Intercept under 'Locus' was 0.00

Notes: "DAFOR": Conventional DAFOR scale (D = Dominant 5; A = Abundant 4; F = Frequent 3; O = Occasional 2; R = Rare 1); "SO": site occupancy; "RRC": relative read count; "Locus": Cytb or 12S.

Chapter 6 General discussion

Environmental DNA (eDNA) analysis is starting to change the way we design and implement biodiversity monitoring programmes and has opened up new possibilities for the future. However, there are a number of challenges distributed throughout the workflow including choice of DNA capture, preservation and extraction methods, the specificity and taxonomic resolution of PCR primers, choice of parameters and pipelines during bioinformatics analysis, and availability of reference databases for taxonomic assignment. These challenges must be overcome to achieve accurate, standardised tools that can be routinely and reproducibly implemented. In this thesis, I evaluated and optimised various aspects (marker selection and primer design, compilation marker-specific reference database and eDNA capture) of the workflow for eDNA-based metabarcoding of freshwater fish communities. I also investigated the spatial and temporal distribution of eDNA to inform the eDNA sampling strategies and ensure the accuracy and reliability of eDNA biodiversity assessments, and the potential of eDNA as a tool for biodiversity monitoring with a larger dataset from a range of lakes with different ecological characteristics. In this chapter, I discuss the main findings of the studies with current knowledge of eDNA from other studies to fully exploit the potential of metabarcoding data and improve the accuracy and precision of their analysis for the future integration of eDNA metabarcoding to routine biological monitoring programmes (Figure 6.1).

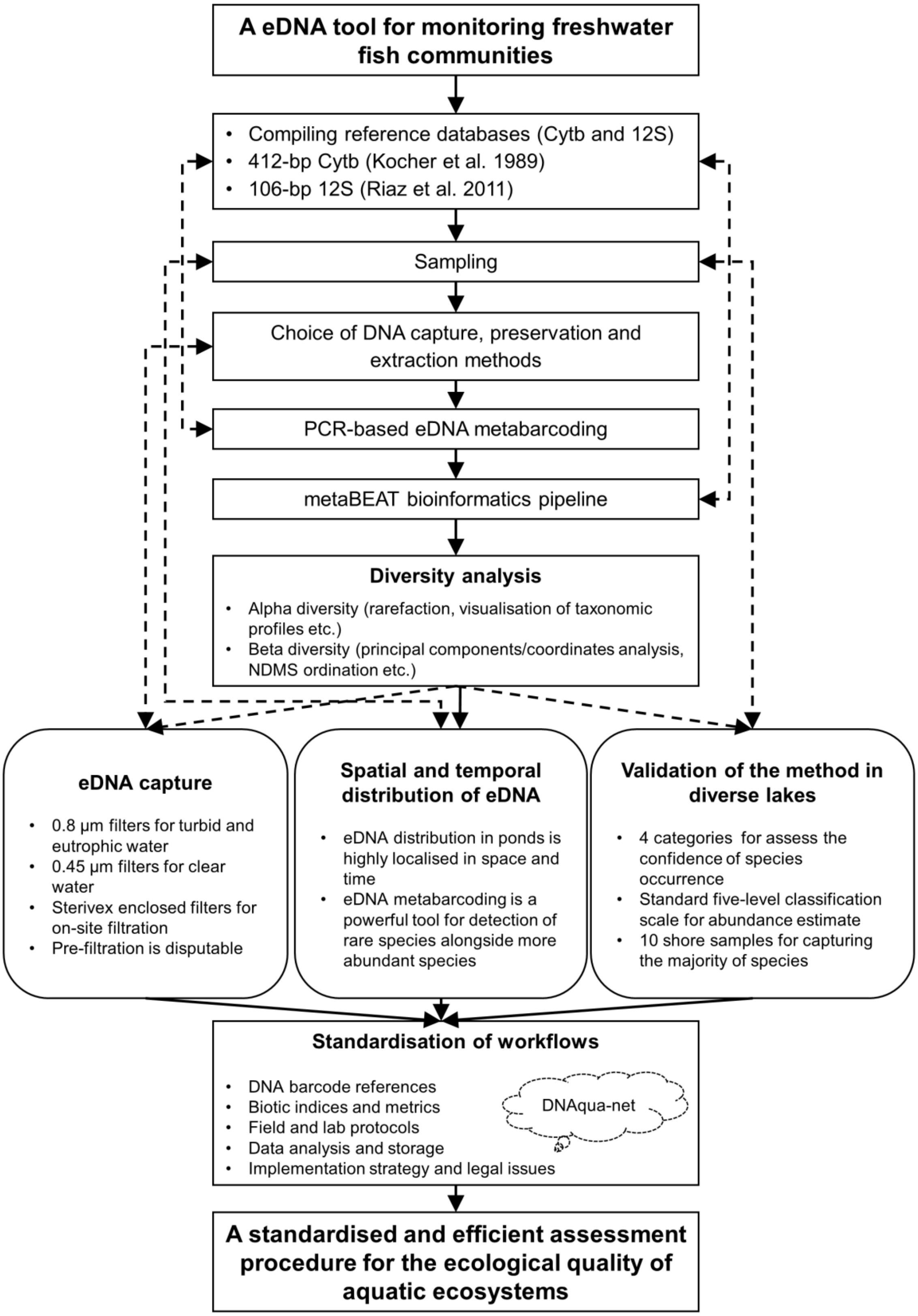

A eDNA tool for monitoring freshwater fish communities
• Compiling reference databases (Cytb and 12S)
• 412-bp Cytb (Kocher et al. 1989)
• 106-bp 12S (Riaz et al. 2011)
Sampling
Choice of DNA capture, preservation and extraction methods
PCR-based eDNA metabarcoding
metaBEAT bioinformatics pipeline
Diversity analysis
• Alpha diversity (rarefaction, visualisation of taxonomic profiles etc.)
• Beta diversity (principal components/coordinates analysis, NDMS ordination etc.)
eDNA capture
• 0.8 µm filters for turbid and eutrophic water
• 0.45 µm filters for clear water
• Sterivex enclosed filters for on-site filtration
• Pre-filtration is disputable
Spatial and temporal distribution of eDNA
• eDNA distribution in ponds is highly localised in space and time
• eDNA metabarcoding is a powerful tool for detection of rare species alongside more abundant species
Validation of the method in diverse lakes
• 4 categories for assess the confidence of species occurrence
• Standard five-level classification scale for abundance estimate
• 10 shore samples for capturing the majority of species
Standardisation of workflows
• DNA barcode references
• Biotic indices and metrics
• Field and lab protocols
• Data analysis and storage
• Implementation strategy and legal issues
DNAqua-net
A standardised and efficient assessment procedure for the ecological quality of aquatic ecosystems

Figure 6.1 Overview of the eDNA metabarcoding workflow for monitoring freshwater fish communities developed in this thesis. Dotted arrows indicate that the steps directly interact with each other.

6.1 Choice of markers and primers bias in metabarcoding

Although mitochondria are found in vast copy numbers in metazoa, mitochondrial DNA only accounts for a small fraction of the total DNA compared to nuclear sequences. For example, less than 0.5% useful mitochondrial sequences reads were retrieved from bulk arthropod samples using shotgun sequencing (Zhou et al. 2013). In water samples, eDNA from macro-organisms generally occurs at very low concentration and can be heterogeneously distributed throughout a water body (Deiner et al. 2017a; Taberlet et al. 2018). The concentration of mitochondrial DNA is even lower. For example, Turner et al. (2014) estimated that DNA from an established population of common carp *Cyprinus carpio* makes up ⩽ 0.0004% of total DNA in water samples; equivalent to < 0.01 ng mitochondrial DNA/litre. Therefore, enrichment of target mitochondria DNA is required before it can be sequenced. This is usually achieved through PCR-based amplification for obtaining community-level data (i.e., PCR-based metabarcoding).

PCR primers for amplifying mixed samples need to be highly conserved across target taxa in order avoid preferential amplification of certain taxa (i.e., primer biases). Despite best efforts, primer biases have been observed in a number of metabarcoding studies which results in skewed relative sequence abundances and failure to detect some taxa known to be present, particularly those that are rare (Clarke et al. 2014; Elbrecht & Leese 2015; Piñol et al. 2018).

Primer bias is generally unavoidable when a broad taxonomic range is targeted as it becomes more difficult to identify a region which is completely conserved across all taxa, but careful primers design can minimise it. Recently, a number of software programmes have been developed to aid the design and *in silico* evaluation of metabarcoding primers such as ecoPCR (Ficetola et al. 2010), PrimerMiner (Elbrecht &

Leese 2017), PrimerTree (Cannon et al. 2016), and Metacoder (Foster et al. 2017). I evaluated the metabarcoding primers for UK freshwater fish communities *in silico* with ecoPCR (Ficetola et al. 2010), since, to my knowledge, it was the only available software programme when I conducted metabarcoding studies. The *in silico* resolution results are influenced by the reliable and robust marker-specific reference databases. It is unfortunate; however, that many of the sequences submitted to public databases do not include the conserved priming sites used for their amplification, which limits their value for primer design and evaluation. For example, a large proportion of sequences in the curated Cytb database of this study do not cover the primer binding site of forward primer, which reduces the reliability of the evaluation. Taberlet et al. (2018) suggested that, if feasible, the amplification for the local reference database should be performed using external primers and not using the primers that will be used for the metabarcoding study. Thus, a set of novel primers was designed in order to generate reference sequences of the entire 12S region of freshwater fish in this study. The main take home message of this study is that whether *in silico* evaluation against the curated 12S reference database using ecoPCR (Ficetola et al. 2010), *in vitro* validation on 10 mock communities or *in situ* application in aquaculture fish ponds and diverse lakes, the 12S_V5_F and 12S_V5_R primer pairs (Riaz et al. 2011) targeting a 106-bp fragment in fish is highly suitable for eDNA metabarcoding of UK freshwater fish communities. Primer bias was therefore not found to be a problem for the combination of primers and fish communities analysed in the present thesis.

As demonstrated and discussed in Chapter 5, together with a growing number of studies (Hänfling et al. 2016; Evans et al. 2017; Stat et al. 2017; Li et al. 2018b), use of more than one locus for a target group in metabarcoding can allow for tests of consistency between loci and increase stringency of species detection and avoid problems due to primer bias, PCR or sequencing artefacts and/or contamination. To further address primer bias, if it is suspected in future studies, a hybridisation capture enrichment approach has been applied effectively for targeting mitochondrial DNA from bulk samples (Dowle et al. 2016; Liu et al. 2016), benthic samples (Shokralla et al. 2016), and water samples (Wilcox et al. 2018). With this approach, sheared DNA fragments are hybridised to probes (baits), and then recovered by streptavidin-coated magnetic beads (Dowle et al. 2016; Jones & Good 2016). Concerning the quantitative aspect, capture probe bias still exists although strong correlations in relative abundances of before and after capture were observed (Dowle et al. 2016; Liu et al. 2016; Shokralla et al. 2016; Wilcox et al. 2018). While these initial studies have evaluated the performance of hybridisation probes against metabarcoding, the design of unbiased probes and further tests especially with environmental samples are needed to get this method beyond the proof of concept stage.

6.2 eDNA capture from water

eDNA analysis often deals with small quantities of short and degraded DNA fragments; therefore methods that maximise eDNA recovery are required to increase detectability. Furthermore, the efficiency of different DNA recovery methods needs to be understood to ensure the comparability of different studies. This is especially relevant for the development of eDNA tools for biodiversity assessment where results from different laboratories are required to produce the same outcome, and therefore, there should be no methodological bias. There is an interactive influence among the eDNA recovery steps of capture, preservation, and extraction. Several studies have demonstrated that different combinations of recovery methods vary the quantity of eDNA and efficiency of species detection (e.g., Deiner et al. 2015; Renshaw et al. 2015;

Eichmiller et al. 2016; Piggott 2016; Hinlo et al. 2017). DNA capture is a crucial step in eDNA analysis. Most studies focussing on individual target species to investigate the impact of different types and pore sizes of filter on DNA quantity, indicated that the smaller pore size (0.2 or 0.45 μm) of filters were more likely to clog and increase filtration time (Turner et al. 2014; Eichmiller et al. 2016; Minamoto et al. 2016). Minamoto et al. (2016) advocated using 0.7 μm glass fibre filters for capture eDNA from ponds and lakes when considering DNA quantity. Nevertheless, Eichmiller et al. (2016) found that 0.2–0.6 μm polycarbonate filters are optimal for biomass quantification in the laboratory.

In this thesis, I investigated the impact of different pore sizes of the membrane filter (0.45, 0.8 and 1.2 μm), different types of filter (0.45 μm membrane filters and 0.45 μm Sterivex enclosed filters), and pre-filtration (20 μm pre-filters with 0.45 μm filters) on eDNA recovery and fish community composition via metabarcoding (Chapter 3). The results showed that pore sizes of mixed cellulose ester (MCE) filter ranging from 0.45 to 1.2 μm are broadly consistent in their DNA recovery for metabarcoding analysis when a sufficient number of replicates (*N* = 5) are carried out (Chapter3; see also in Li et al. 2018a). However, in terms of representing the community composition, the performance of 0.45 μm filters is consistently higher than for other filter types, in agreement with other studies (Miya et al. 2016; Majaneva et al. 2018).

Where water is turbid, to process greater water volumes and reduce the filtration time, centrifugation, increased pore size, or pre-filtration will be necessary (Takahara et al. 2012; Minamoto et al. 2016; Robson et al. 2016). Majaneva et al. (2018) demonstrated that pre-filtration could potentially reduce the number of detected metazoan taxa, although it recovered higher diversity index values and more consistent community composition. In my study, pre-filtration prevents the suspended particulate matter from clogging finer filters without reducing the probability of species detection and repeatability of community composition. However, pre-filters increase cost, and the larger pore sizes trade capture of smaller particle sizes for greater proportions of target DNA, reducing total eDNA yield (Turner et al. 2014). Additionally, there is a drawback of pre-filtration in terms of more handling, which could increase the opportunity for contamination. These issues make it difficult to standardise the exact filtration method or volume of water processed for a standardised tool that can be routinely and reproducibly implemented. Encouragingly though, the probability of species detection might not significantly decrease if using a larger pore size (i.e., 0.6 to 6 µm) as the greater volume of water filtered likely compensates for loss of small particle sizes (Eichmiller et al. 2016; Minamoto et al. 2016; Goldberg et al. 2018). Similarly, in this study, the 0.8 and 1.2 µm filters provide dramatic improvements in filtration time, with a comparable recovery of eDNA, and did not impede probability of species detection, repeatability of community composition, and the relationship between read counts and abundance or biomass. By contrast, total DNA yield, the probability of species detection and repeatability reduces using very large pore size filters (20 µm). These findings are consistent with fractionation studies that showed that the modal size for fish eDNA is between 1 and 10 µm ((Turner et al. 2014; Wilcox et al. 2015). The optimal pore size of the filter or filter type will strongly depend on the water type under study. In summary, for development of eDNA tools for biological monitoring of freshwater ecosystems, small pore sizes of 0.45 µm are appropriate for lakes and rivers sampling to obtain consistent results. For some other special circumstances and other conservation purposes, the 0.8 µm filters are recommended for turbid and eutrophic water such as ponds to process a greater volume of water and increase the detection probability and the 0.45 μm Sterivex enclosed filters are suitable in situations where on-site filtration is required.

6.3 Sampling strategies

Design of sampling strategies for eDNA studies should be adjusted to maximise the biological signal obtained while minimising the sampling effort. However, the distribution and dispersion of eDNA in freshwater habitats complicates the design of sampling strategies. Previous studies have shown eDNA is heterogeneously distributed in the lotic ecosystems across both spatial and temporal scales (Deiner & Altermatt 2014; Pilliod et al. 2014; Jane et al. 2015; Spear et al. 2015; Jerde et al. 2016; Wilcox et al. 2016; Tillotson et al. 2018).

In Chapter 4, I investigated the spatial and temporal distribution of eDNA in fish ponds following the introduction of two rare species with keepnets and removal the keepnets after eight days via metabarcoding. The results revealed that eDNA concentration (i.e., proportional read counts abundance) of the introduced species typically peaked after two days, which is supported by the highest community dissimilarity of different sampling positions observed on the second day after introduction (Chapter 4). But this may have been caused by increased eDNA shedding rates as a result of fish being stressed by handling, as observed in other studies (Takahara et al. 2012; Maruyama et al. 2014; Klymus et al. 2015; Sassoubre et al. 2016). There is also a strong evidence base linking eDNA detection and concentration to life stage, body condition (a consequence of reproduction), seasonality, and behaviour of species (e.g., Spear et al. 2015; de Souza et al. 2016; Bylemans et al. 2017; Buxton et al. 2018; Takahashi et al. 2018; Tillotson et al. 2018). Therefore, the behaviour or activity of target species should be taken into consideration when choosing a sampling time frame.

Moreover, the introduced species were no longer detected at any sampling positions 48 hrs after removal from the ponds. As a result, there is no significant difference in community dissimilarity of different sampling positions among the sampling days after removal of the introduced species (Chapter 4). This observation is in agreement with other studies that documented no eDNA detection shortly (less than 48 hrs) after target species were removed from the water in which they occurred (Thomsen et al. 2012a; Maruyama et al. 2014; Balasingham et al. 2017). Comparing to other studies in controlled aquaria or mesocosms (e.g., Dejean et al. 2011; Thomsen et al. 2012a; Thomsen et al. 2012b; Goldberg et al. 2013; Barnes et al. 2014; Sassoubre et al. 2016), eDNA is found to decay faster in the field than in controlled conditions, which can be attributed to the complex effects of environmental conditions on eDNA persistence (Barnes et al. 2014; Pilliod et al. 2014; Strickler et al. 2015; Lance et al. 2017; Stoeckle et al. 2017; Seymour et al. 2018).

In terms of eDNA dispersion in ponds, there is a strong decrease in eDNA detection probability with distance from the keepnet, with the introduced species nearly undetectable after a few metres, which indicated that eDNA distribution in ponds is highly localised in space (Chapter 4). Similarly, eDNA barcoding studies have shown that eDNA accumulated nearby to the target species (Takahara et al. 2012; Eichmiller et al. 2014; Dunker et al. 2016). Together with other studies in large lakes (Hänfling et al. 2016; Sato et al. 2017; Lawson Handley et al. 2019), these results demonstrate that there is considerable spatial heterogeneity of eDNA distribution in lentic ecosystems. In summary, my research work reveals that eDNA distribution in ponds is highly localised in space and time, which adding to the growing weight of evidence that eDNA signal provides an accurate description of aquatic communities.

To explore effectiveness of different spatial sampling approaches, the results from nine lakes ranging in size from 50, 000 to 1,400,000 m^2 showed that the number of species detected in the shore samples is equal to, or slightly higher than in the offshore samples during winter, and suggested that 10 samples is adequate for capturing the majority of species (Chapter 5). Similar results were observed from our previous work in Lake Windermere (England, UK) during the winter sampling campaign (Hänfling et al. 2016). But there is a slightly difference in species detection between the shore and offshore samples that reflects the species ecology, with greater spatial structuring during the summer months in Lake Windermere, due to water stratification (Lawson Handley et al. 2019). For example, Arctic charr *Salvelinus alpinus*, a deep-water species, are only detected in deep, mid-lake samples in the summer, while littoral or benthic species such as minnow *Phoxinus phoxinus* and three-spined stickleback *Gasterosteus aculeatus* are more frequently detected in shore samples (Lawson Handley et al. 2019).

At present, there is no consensus on volume of water, number of samples, and frequency of sampling from an individual freshwater habitat. Further investigation is required to determine the number of samples needed to achieve a set detection probability for a target species or representative community composition. The results generated in this thesis (Chapter 5) and in our additional studies (Hänfling et al. 2016; Lawson Handley et al. 2019) indicated that to maximise the number of species that can be detected, while minimising the costs and effort associated with sampling, 10 shore samples distributed along the full perimeter of lakes at least up to 14,800,000 m^2 in size and 64 m in depth, is adequate for capturing the majority of species during winter months. However, if abundance estimation using site occupancy data as opposed to sequencing read count, it makes more sense to collect as many, spatially and temporally representative, samples as possible (a detailed discussion is given in Section 6.4).

6.4 Integrating eDNA metabarcoding data to freshwater biodiversity monitoring programmes

6.4.1 Abundance estimate with eDNA

After reviewing and discussing the potential of eDNA metabarcoding as a tool for the Water Framework Directive (WFD) status assessment, Hering et al. (2018) suggested that this approach is well-suited for fish biodiversity assessment, as suitability of DNA-based identification is particularly high for fish considering representativeness, sensitivity, precision, comparability, cost-effectiveness, and environmental impact. A prototype eDNA tool for fish biodiversity assessment was tested in three lakes in the English Lake District; the results also indicated that distribution of eDNA in Lake Windermere are correlated with the expected distribution of eutrophic tolerant and less tolerant fish species (Hänfling et al. 2016). However, before eDNA metabarcoding can be implemented for WFD status assessment, there is one important question needing to be addressed: How best to fulfil the legal requirement of recording abundance? (Hering et al. 2018; Pawlowski et al. 2018).

Sequencing read count is a valid proxy for abundance estimate under the assumption that no significant bias is introduced during sampling, subsequent PCR or sequencing. However, this assumption is unrealistic, and previous studies have demonstrated that the relationship between abundance and sequencing read count is complex (e.g., Yu et al. 2012; Kelly et al. 2014; Ficetola et al. 2015). Encouragingly though, the results from the comparison of the efficiency of different filter types and pore sizes on eDNA capture demonstrated that there are consistent, positive correlations between sequencing read counts and fish abundance or biomass across the six treatments and four ponds (Chapter 3). This finding is in agreement with other studies that documented there are positive relationships between the relative sequencing read counts and relative abundance and/or biomass density or rank abundance estimated with conventional survey methods (e.g., Evans et al. 2016; Hänfling et al. 2016; Lawson Handley et al. 2019). In contrast, Lim et al. (2016) suggested that sequencing read count is a poor proxy for both abundance and biomass estimated using a conventional survey method. One possible option to improve estimates of abundance, without relying on correlations, is the addition of internal standard DNAs followed by use of a copy number correction (Ushio et al. 2018). An alternative approach to sequencing read count is to use the site occupancy, which can be considered as a proxy for species abundance. Site occupancy modelling or site occupancy data has been successfully applied to the analysis of eDNA data for imperfect species detection resulting from spatial heterogeneity of eDNA distribution in freshwater ecosystems (Pilliod et al. 2013; Schmidt et al. 2013; Ficetola et al. 2015; Hänfling et al. 2016; Valentini et al. 2016; Lawson Handley et al. 2019). However, the full site occupancy modelling requires the estimation of detection probability from temporal sampling (i.e., multiple visits with a number of sampling sites) (MacKenzie et al. 2002; MacKenzie & Royle 2005). Moreover, sequencing read count should not be ignored entirely because they still contain important abundance information.

To further address the abundance estimate, errors (i.e., false negatives and false positives) from DNA-based species detection, I collected and analysed water samples from 14 UK lakes, with well-described fish faunas (Chapter 5). Results indicated that it is possible to evaluate the confidence of species presence and estimate species relative abundance based on both site occupancy and read counts (Chapter 5). Specifically, four categories (probably absent, possibly present, probably present, and established) were used to assess the confidence of species occurrence using eDNA data, and the five-level classification scale (relative abundance DAFOR scale) was used to estimate species relative abundance. Thus, reasonable sampling design on spatial and temporal scales will improve the site occupancy model for estimate species relative abundance (see more detail in Section 6.3).

So far, more extensive validation and demonstration of the accuracy, effectiveness and applicability of the method is required in order to develop a robust tool for biodiversity monitoring. We have also sampled and analysed eDNA metabarcoding data from 30 small lakes (0.1–45 m^2) along the proposed London Flood Channel and 35 Scottish lochs (420,000–70,730,000 m^2) across a wide range of ecological conditions with good data from conventional survey methods on the fish communities under WFD monitoring. Together with these two datasets, the results will further inform the number of water samples that should be taken from a lake, the accuracy and effectiveness of the confidence and methods for relative abundance evaluation.

6.4.2 Molecular biotic indices obtain through eDNA metabarcoding

The biodiversity assessment of aquatic ecosystems is currently based on various biotic metrics/indices that use the occurrence and/or abundance of selected taxonomic groups to define ecological status. However, there are some limitations when using conventional survey methods to obtain these biotic metrics/indices, such as morphological identification bias, recording small-bodied, rare and/or elusive species, and destructive impacts on the environment (Stribling et al. 2008; Deiner et al. 2017a). eDNA metabarcoding could potentially alleviate some of these limitations by using genetic information from environmental samples instead of morphology to identify organisms and to characterise a given ecosystem, which is a promising tool for rapid, non-invasive biodiversity monitoring (reviewed in Rees et al. 2014; Barnes & Turner 2016; Jackson et al. 2016; Deiner et al. 2017a; Hering et al. 2018). The challenge for DNA-based assessment is to find a fit within current biodiversity assessment frameworks such as WFD that will enhance our ability to detect and identify stressor impacts. Therefore, we need to consider how to integrate DNA metabarcoding data to development of metrics/indices for biological monitoring and assessment.

The pilot eDNA metabarcoding studies applied to biodiversity assessment can be classified into three categories according to their scope: (1) studies that use metabarcoding data to infer existing morphotaxonomy-based biotic indices (e.g., diatoms Apothéloz-Perret-Gentil et al. 2017), (2) studies that explore the potential of new biological indicator taxa such as cyanobacteria (Mateo et al. 2015), and (3) studies that search for alternative analytical methods to develop new molecular indices such as Supervised Machine Learning (SML) (Cordier et al. 2017; Gerhard & Gunsch 2019). The challenges addressed by each of these categories are not the same. The first group of studies is mainly concerned with testing and improving the match between indices derived from morphological and molecular data. The key challenges of the second and third categories are to develop new analytical methods and indices based on metabarcoding data for the taxonomic groups that are not currently used in ecological quality assessment. In addition, there are still some technical challenges distributed throughout the eDNA metabarcoding workflow including choice of DNA capture, preservation and extraction methods, the specificity and taxonomic resolution of PCR primers, choice of parameters and pipelines during bioinformatics analysis, and availability of reference databases for taxonomic assignment (see more detail in Chapter 1, Section 1.4).

In view of these potential limitations, a two-step implementation of metabarcoding in routine biological monitoring programmes is recommended. In the short term, the metabarcoding data could be integrated into the existing biotic indices. For example, WFD requires ecological status assessment of surface waters to be based on Biological Quality Elements (BQEs), which depending on the water body type, include phytoplankton, diatoms, macrophytes and phytobenthos, benthic invertebrate fauna, and fish fauna. The use of metabarcoding data will provide considerable advantages for any biotic indices based on BQEs given that the adequate effort to complete comprehensive group specific databases is provided. This could be easily done for fish which have been the focus of most metabarcoding studies (e.g., Hänfling et al. 2016; Port et al. 2016; Valentini et al. 2016; Li et al. 2019). In the case of fish-based biotic indices, eDNA analyses offer the possibility to survey fish populations without killing or disturbing them, and to use genetic diversity as a new way to measure degradation. This first step integration could be done locally, with each country being able to use its own biotic indices to test and validate the use of molecular data, applied to the reference water bodies, as highlighted in Leese et al. (2018). In parallel, special efforts need to be provided in order to increase accuracy and precision of the biotic indices by ensuring that the databases are covering at least the important taxa for the biotic index calculations.

In the long term, the new molecular indices could be developed based entirely on metabarcoding data. Such biotic indices could provide a more holistic view of biological community response to the anthropogenic stressors by including new potential BQEs, in particular various groups of prokaryotic and eukaryotic microbiota and meiofauna. They could be based on predictive models established using machine-learning and other algorithms capable of assessing ecological status and identifying ecologically meaningful Molecular Operational Taxonomic Units (MOTUs) in the metabarcoding datasets. Last but not least, to comply with the WFD, these new biotic indices should be benchmarked against both currently existing indices and directly against the pressure data in order to redefine the boundary settings, which will require large-scale intercalibration exercises. The final outcome of such exercises could be the development of pan-European or global molecular biotic indices, which will constitute a major advance towards a standardised and efficient assessment of the ecological quality of aquatic ecosystems (Figure 6.1).

6.5 Conclusions

In my thesis, two metabarcoding primer pairs targeting different mitochondrial genes (Cytb and 12S) are rigorously tested, and two marker-specific reference databases have been compiled with a reproducible workflow, which are fundamental for other eDNA metabarcoding studies focusing on freshwater fish communities. The 0.8 μm filters are advocated for turbid and eutrophic water such as ponds to reduce the filtration time, the 0.45 μm filters are appropriate for clear water sampling to obtain consistent results, and the 0.45 μm Sterivex enclosed filters are suitable in situations where on-site filtration is required. Pre-filters are applied only if absolutely essential for reducing the filtration time or increasing the throughput volume of the capture filters. Furthermore, eDNA distribution in ponds is highly localised in space and time, and 10 shore samples distributed along the full perimeter of lakes is adequate for capturing the majority of species. Lastly, my thesis provides further evidence that eDNA metabarcoding is considerably outperforming other established survey techniques in a wide range of lake types for community-level analysis whether in species detection, relative abundance estimate using the standard five-level classification scale or characterisation ecological fish communities (Figure 6.1).

Therefore, there is a great potential for eDNA-based metabarcoding to be used for assessment procedures to fulfil the requirements of biodiversity monitoring programmes such as WFD. However, a broad standardisation of eDNA workflows, ranging from sampling and lab protocols to the calculation of biotic metrics/indices with highly standardised analysis pipelines, will ensure more robust, comparable, and ecologically meaningful data to guide effective management and conservation of freshwater biodiversity. This process has begun in Europe with the establishment of DNAqua-net. DNAqua-net consortium aims at developing novel molecular tools for biodiversity assessment and biological monitoring of aquatic ecosystems and is composed of five working groups that contribute to these overarching goals: DNA barcode references, biotic indices and metrics, field and lab protocols, data analysis and storage, and implementation strategy and legal issues (Leese et al. 2018). These standardisation outcomes will constitute a major advance towards a standardised and efficient assessment procedure for the ecological monitoring of aquatic ecosystems (Figure 6.1).

References

Adams, S.M. (2002) Biological indicators of aquatic ecosystem stress. pp. 644. American Fisheries Society, Bethesda.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology,* **215,** 403-410.

Andújar, C., Arribas, P., Yu, D.W., Vogler, A.P. & Emerson, B.C. (2018) Why the COI barcode should be the community DNA metabarcode for the Metazoa. *Molecular Ecology,* **27,** 3968-3975.

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P. & Pawlowski, J. (2017) Taxonomy ‐ free molecular diatom index for high ‐ throughput eDNA biomonitoring. *Molecular Ecology Resources,* **17,** 1231-1242.

Argillier, C., Causse, S., Gevrey, M., Pedron, S., De Bortoli, J., Brucet, S., ... Holmgren, K. (2013) Development of a fish-based index to assess the eutrophication status of European lakes. *Hydrobiologia,* **704,** 193-211.

Baker, C.S., Steel, D., Nieukirk, S. & Klinck, H. (2018) Environmental DNA (eDNA) from the wake of the whales: droplet digital PCR for detection and species identification. *Frontiers in Marine Science,* **5,** 133.

Balasingham, K.D., Walter, R.P. & Heath, D.D. (2017) Residual eDNA detection sensitivity assessed by quantitative real-time PCR in a river ecosystem. *Molecular Ecology Resources,* **17,** 523-532.

Baldigo, B.P., Sporn, L.A., George, S.D. & Ball, J.A. (2017) Efficacy of environmental DNA to detect and quantify brook trout populations in headwater streams of the Adirondack Mountains, New York. *Transactions of the American Fisheries Society,* **146,** 99-111.

Barnes, M.A. & Turner, C.R. (2016) The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics,* **17,** 1-17.

Barnes, M.A., Turner, C.R., Jerde, C.L., Renshaw, M.A., Chadderton, W.L. & Lodge, D.M. (2014) Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science & Technology,* **48,** 1819-1827.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. (2013) GenBank. *Nucleic Acids Research,* **41,** D36-D42.

Berger, S.A., Krompass, D. & Stamatakis, A. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology,* **60,** 291-302.

Bergman, P.S., Schumer, G., Blankenship, S. & Campbell, E. (2016) Detection of adult green sturgeon using environmental DNA analysis. *PLoS ONE,* **11,** e0153500.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R.A., ... Brotherton, P. (2015) Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus*). *Biological Conservation,* **183,** 19-28.

Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., ... Hering, D. (2012) Three hundred ways to assess Europe's surface waters: an almost complete overview of biological methods to implement the Water Framework Directive. *Ecological Indicators,* **18,** 31-41.

Bista, I., Carvalho, G.R., Walsh, K., Seymour, M., Hajibabaei, M., Lallias, D., ... Creer, S. (2017) Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nature Communications,* **8,** 14087.

Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* **30,** 2114-2120.

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

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) OBITOOLS: a unix ‐ inspired software package for DNA metabarcoding. *Molecular Ecology Resources,* **16,** 176-182.

Brady, A. & Salzberg, S.L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods,* **6,** 673.

Bray, J. (1995) An assessment of the relative abundances and species composition of fish populations in Llyn Penrhyn and Llyn Dinam (Anglesey). National Rivers Authority.

Bronner, I.F., Quail, M.A., Turner, D.J. & Swerdlow, H. (2014) Improved protocols for Illumina sequencing. *Current Protocols in Human Genetics,* **80,** 18.12.11-18.12.42.

Burgener, M. & Hübner, P. (1998) Mitochondrial DNA enrichment for species identification and evolutionary analysis. *Zeitschrift für Lebensmitteluntersuchung und-Forschung A,* **207,** 261-263.

Burk, A., Douzery, E.J. & Springer, M.S. (2002) The secondary structure of mammalian mitochondrial 16S rRNA molecules: refinements based on a comparative phylogenetic approach. *Journal of Mammalian Evolution,* **9,** 225-252.

Buxton, A.S., Groombridge, J.J. & Griffiths, R.A. (2018) Seasonal variation in environmental DNA detection in sediment and water samples. *PloS ONE,* **13,** e0191737.

Bylemans, J., Furlan, E.M., Hardy, C.M., McGuffie, P., Lintermans, M. & Gleeson, D.M. (2017) An environmental DNA‐based method for monitoring spawning activity: a case study, using the endangered Macquarie perch (*Macquaria australasica*). *Methods in Ecology and Evolution,* **8,** 646-655.

Cannon, M., Hester, J., Shalkhauser, A., Chan, E.R., Logue, K., Small, S.T. & Serre, D. (2016) *In silico* assessment of primers for eDNA studies using PrimerTree and application to characterize the biodiversity surrounding the Cuyahoga River. *Scientific Reports,* **6,** 22908.

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics,* **25,** 1972-1973.

CEC (2000) Directive 2000/60/EC of the European Parliament and of the Council: establishing a framework for Community action in the field of water policy. Official Journal of the European Communities, Luxembourg.

Chao, A., Ma, K.H., Hsieh, T.C. & Chiu, C.H. (2016) SpadeR: species-richness prediction and diversity estimation with R): an R package in CRAN. Retrieved from https://CRAN.R-project.org/package=SpadeR.

Child, A. (1977) Biochemical polymorphism in char (*Salvelinus alpinus* L.) from Llynnau Peris, Padarn, Cwellyn and Bodlyn. *Heredity,* **38,** 359.

Civade, R., Dejean, T., Valentini, A., Roset, N., Raymond, J.-C., Bonin, A., ... Pont, D. (2016) Spatial representativeness of environmental DNA metabarcoding signal for fish biodiversity assessment in a natural freshwater system. *PloS ONE,* **11,** e0157366.

Clabburn, P., Davies, R. & Griffiths, J. (2011) Monitoring the Arctic charr spawning run on the Afon y Bala with DIDSON multibeam sonar between November 2010 and January 2011. Environment Agency Wales.

Clabburn, P., Davies, R. & Griffiths, J. (2016) Summary of the results of hydroacoustic surveys of Llyn Padarn and Llyn Cwellyn, 2014 and 2015. Natural Resources Wales.

Clabburn, P. & Griffiths, J. (2013) Summary of the results of hydroacoustic surveys of Llyn Padarn and Llyn Cwellyn in 2012. Environment Agency Wales.

Clarke, L.J., Soubrier, J., Weyrich, L.S. & Cooper, A. (2014) Environmental metabarcodes for insects: *in silico* PCR reveals potential for taxonomic bias. *Molecular Ecology Resources,* **14,** 1160-1170.

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

Collen, B., Whitton, F., Dyer, E.E., Baillie, J.E., Cumberlidge, N., Darwall, W.R., ... Böhm, M. (2014) Global patterns of freshwater species diversity, threat and endemism. *Global Ecology and Biogeography,* **23,** 40-51.

Comtet, T., Sandionigi, A., Viard, F. & Casiraghi, M. (2015) DNA (meta) barcoding of biological invasions: a powerful tool to elucidate invasion processes and help managing aliens. *Biological Invasions,* **17,** 1-18.

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., ... Pawlowski, J. (2017) Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology,* **51,** 9118-9126.

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

Cristescu, M.E. (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution,* **29,** 566-571.

Davies, C.E., Shelley, J., Harding, P.T., McLean, I.F.G., Gardiner, R. & Peirson, G. (2004) Freshwater Fishes in Britain - the species and their distribution. pp. 176. Harley Books, Colchester.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E. & Taberlet, P. (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources,* **14,** 306-323.

de Souza, L.S., Godwin, J.C., Renshaw, M.A. & Larson, E. (2016) Environmental DNA (eDNA) detection probability is influenced by seasonal activity of organisms. *PloS ONE,* **11,** e0165273.

Deagle, B.E., Eveson, J.P. & Jarman, S.N. (2006) Quantification of damage in DNA recovered from highly degraded samples - a case study on DNA in faeces. *Frontiers in Zoology,* **3,** 11.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. & Taberlet, P. (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters,* **10,** 20140562.

Deakin, C.T., Deakin, J.J., Ginn, S.L., Young, P., Humphreys, D., Suter, C.M., ... Hallwirth, C.V. (2014) Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. *Nucleic Acids Research,* **42,** e129.

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

Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ... Vere, N. (2017a) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology,* **26,** 5872-5895.

Deiner, K., Fronhofer, E.A., Machler, E., Walser, J.C. & Altermatt, F. (2016) Environmental DNA reveals that rivers are conveyer belts of biodiversity information. *Nature Communications,* **7,** 1-9.

Deiner, K., Renshaw, M.A., Li, Y., Olds, B.P., Lodge, D.M. & Pfrender, M.E. (2017b) Long - range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods in Ecology and Evolution,* **8,** 1888-1898.

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

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P. & Miaud, C. (2011) Persistence of environmental DNA in freshwater ecosystems. *PLoS ONE,* **6,** e23398.

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E. & Miaud, C. (2012) Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. *Journal of Applied Ecology,* **49,** 953-959.

Diaz, N.N., Krause, L., Goesmann, A., Niehaus, K. & Nattkemper, T.W. (2009) TACOA–Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics,* **10,** 56.

Djurhuus, A., Port, J., Closek, C.J., Yamahara, K.M., Romero-Maraccini, O., Walz, K.R., ... Breitbart, M. (2017) Evaluation of filtration and DNA extraction methods for environmental DNA biodiversity assessments across multiple trophic levels. *Frontiers in Marine Science,* **4,** 314.

Doi, H., Inui, R., Akamatsu, Y., Kanno, K., Yamanaka, H., Takahara, T. & Minamoto, T. (2017a) Environmental DNA analysis for estimating the abundance and biomass of stream fish. *Freshwater Biology,* **62,** 30-39.

Doi, H., Takahara, T., Minamoto, T., Matsuhashi, S., Uchii, K. & Yamanaka, H. (2015) Droplet digital polymerase chain reaction (PCR) outperforms real-time PCR in the detection of environmental DNA from an invasive fish species. *Environmental Science & Technology,* **49,** 5601-5608.

Doi, H., Uchii, K., Matsuhashi, S., Takahara, T., Yamanaka, H. & Minamoto, T. (2017b) Isopropanol precipitation method for collecting fish environmental DNA. *Limnology and Oceanography-Methods,* **15,** 212-218.

Dowle, E.J., Pochon, X., C. Banks, J., Shearer, K. & Wood, S.A. (2016) Targeted gene enrichment and high ‐ throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. *Molecular Ecology Resources,* **16,** 1240-1254.

Dudgeon, D., Arthington, A.H., Gessner, M.O., Kawabata, Z.-I., Knowler, D.J., Lévêque, C., ... Stiassny, M.L. (2006) Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews,* **81,** 163-182.

Duigan, C., Allott, T., Bennion, H., Lancaster, J., Monteith, D., Patrick, S., ... Seda, J. (1996) The Anglesey lakes, Wales, UK—a conservation resource. *Aquatic Conservation: Marine and Freshwater Ecosystems,* **6,** 31-55.

Dunker, K.J., Sepulveda, A.J., Massengill, R.L., Olsen, J.B., Russ, O.L., Wenburg, J.K. & Antonovich, A. (2016) Potential of environmental DNA to evaluate Northern pike

(*Esox lucius*) eradication efforts: an experimental test and case study. *PloS ONE,* **11,** e0162277.

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

Egan, S.P., Grey, E., Olds, B., Feder, J.L., Ruggiero, S.T., Tanner, C.E. & Lodge, D.M. (2015) Rapid molecular detection of invasive species in ballast and harbor water by integrating environmental DNA and light transmission spectroscopy. *Environmental Science & Technology,* **49,** 4113-4121.

Eichmiller, J.J., Bajer, P.G. & Sorensen, P.W. (2014) The relationship between the distribution of common carp and their environmental DNA in a small lake. *PloS ONE,* **9,** e112611.

Eichmiller, J.J., Miller, L.M. & Sorensen, P.W. (2016) Optimizing techniques to capture and extract environmental DNA for detection and quantification of fish. *Molecular Ecology Resources,* **16,** 56-68.

Elbrecht, V. & Leese, F. (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PloS ONE,* **10,** e0130324.

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

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.-N., ... Leese, F. (2016) Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ,* **4,** e1966.

Elbrecht, V., Vamos, E.E., Meissner, K., Aroviita, J. & Leese, F. (2017) Assessing strengths and weaknesses of DNA metabarcoding ‐ based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution,* **8,** 1265-1275.

Ellegren, H. (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology,* **17,** 1629-1631.

Epp, L.S., Boessenkool, S., Bellemain, E.P., Haile, J., Esposito, A., Riaz, T., ... Brochmann, C. (2012) New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology,* **21,** 1821-1833.

Evans, N.T., Li, Y., Renshaw, M.A., Olds, B.P., Deiner, K., Turner, C.R., ... Pfrender, M.E. (2017) Fish community assessment with eDNA metabarcoding: effects of sampling design and bioinformatic filtering. *Canadian Journal of Fisheries and Aquatic Sciences,* **74,** 1362-1374.

Evans, N.T., Olds, B.P., Renshaw, M.A., Turner, C.R., Li, Y., Jerde, C.L., ... Lodge, D.M. (2016) Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources,* **16,** 29-41.

Favager, R. (1997) An investigation into the reported decline of Kenfig Pool fishery. M.Sc Thesis, University of Glamorgan.

Fernandez, S., Sandin, M.M., Beaulieu, P.G., Clusa, L., Martinez, J.L., Ardura, A. & García-Vázquez, E. (2018) Environmental DNA for freshwater fish monitoring: insights for conservation within a protected area. *PeerJ,* **6,** e4486.

Ficetola, G.F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., ... Pompanon, F. (2010) An *in silico* approach for the evaluation of DNA barcodes. *BMC Genomics,* **11,** 434.

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

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

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., ... Bateman, A. (2010) The Pfam protein families database. *Nucleic Acids Research,* **38,** D211-222.

Foote, A.D., Thomsen, P.F., Sveegaard, S., Wahlberg, M., Kielgast, J., Kyhn, L.A., ... Gilbert, M.T.P. (2012) Investigating the potential use of environmental DNA (eDNA) for genetic monitoring of marine mammals. *PLoS ONE,* **7,** e41781.

Foster, Z.S., Sharpton, T.J. & Grünwald, N.J. (2017) Metacoder: an R package for visualization and manipulation of community taxonomic diversity data. *PLoS Computational Biology,* **13,** e1005404.

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

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* **28,** 3150-3152.

Gerhard, W.A. & Gunsch, C.K. (2019) Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environment International,* **124,** 312-319.

Giles, N. (2003) A fishery management plan for Kenfig Pool cSAC. Countryside Council for Wales.

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

Gleick, P.H. (2011) Water resources. *Encyclopedia of Climate and Weather* (ed. Schneider, S.H.), pp. 817-823. Oxford University Press, New York.

Goldberg, C.S., Pilliod, D.S., Arkle, R.S. & Waits, L.P. (2011) Molecular detection of vertebrates in stream water: a demonstration using Rocky Mountain tailed frogs and Idaho giant salamanders. *PLoS ONE,* **6,** e22746.

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

Goldberg, C.S., Strickler, K.M. & Fremier, A.K. (2018) Degradation and dispersion limit environmental DNA detection of rare amphibians in wetlands: increasing efficacy of sampling designs. *Science of the Total Environment,* **633,** 695-703.

Goldberg, C.S., Turner, C.R., Deiner, K., Klymus, K.E., Thomsen, P.F., Murphy, M.A., ... Taberlet, P. (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution,* **7,** 1299-1307.

Gotelli, N.J. & Colwell, R.K. (2011) Estimating species richness. *Biological diversity: frontiers in measurement and assessment* (eds Magurran, A.E. & McGill, B.J.), pp. 39-54. Oxford University Press, New York.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A. & Baird, D.J. (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE,* **6,** e17497.

Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J., Whitfield, J.B. & Hebert, P.D. (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes,* **6,** 959-964.

Hänfling, B., Lawson Handley, L., Read, D.S., Hahn, C., Li, J., Nichols, P., ... Winfield, I.J. (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology,* **25,** 3101-3119.

Hanks, W. (1998) Arctic charr *(Salvelinus alpinus*) fyke netting survey, Afon Bala - Padarn Lake, Llanberis. Environment Agency Wales.

Hanks, W. (2003) Arctic Charr (*Salvelinus alpinus*) fyke netting survey, Padarn Lake, Llanberis. Environment Agency Wales.

Harper, L.R., Buxton, A.S., Rees, H.C., Bruce, K., Brys, R., Halfmaerten, D., ... Hänfling, B. (2019) Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds. *Hydrobiologia,* **826,** 25-41.

Harper, L.R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H.C., Gough, K.C., ... Hanfling, B. (2018) Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (*Triturus cristatus*). *Ecology and Evolution,* **8,** 6330-6341.

Hassan, R., Scholes, R. & Ash, N. (2005) Ecosystems and human well-being: current state and trends., pp. 948. Island Press, Washington, DC.

Hatton-Ellis, T. (2005) Fish communities and fisheries in Wales's National Nature Reserves: a review. *Freshwater Forum,* **24,** 82-104.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & DeWaard, J.R. (2003a) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B-Biological Sciences,* **270,** 313-321.

Hebert, P.D.N., Ratnasingham, S. & deWaard, J.R. (2003b) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B-Biological Sciences,* **270,** S96-S99.

Hering, D., Borja, A., Jones, J.I., Pont, D., Boets, P., Bouchez, A., ... Kelly, M. (2018) Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research,* **138,** 192-205.

Hickson, R.E., Simon, C., Cooper, A., Spicer, G.S., Sullivan, J. & Penny, D. (1996) Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. *Molecular Biology and Evolution,* **13,** 150-169.

Hinlo, R., Gleeson, D., Lintermans, M. & Furlan, E. (2017) Methods to maximise recovery of environmental DNA from water samples. *PLoS ONE,* **12,** e0179251.

Hofreiter, M., Mead, J.I., Martin, P. & Poinar, H.N. (2003) Molecular caving. *Current Biology,* **13,** R693-R695.

Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., ... Little, D.P. (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences,* **106,** 12794-12797.

Hsieh, H.-M., Chiang, H.-L., Tsai, L.-C., Lai, S.-Y., Huang, N.-E., Linacre, A. & Lee, J.C.-I. (2001) Cytochrome b gene for species identification of the conservation animals. *Forensic Science International,* **122,** 7-18.

Hughes, M., Hornby, D.D., Bennion, H., Kernan, M., Hilton, J., Phillips, G. & Thomas, R. (2004) The development of a GIS-based inventory of standing waters in Great Britain together with a risk-based prioritisation protocol. *Water, Air and Soil Pollution: Focus,* **4,** 73-84.

Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Research,* **17,** 377-386.

Irwin, D.M., Kocher, T.D. & Wilson, A.C. (1991) Evolution of the cytochrome*b* gene of mammals. *Journal of Molecular Evolution,* **32,** 128-144.

Jackson, M.C., Weyl, O., Altermatt, F., Durance, I., Friberg, N., Dumbrell, A., ... Krug, C. (2016) Recommendations for the next generation of global freshwater biological monitoring tools. *Advances in Ecological Research,* **55,** 615-636.

Jane, S.F., Wilcox, T.M., McKelvey, K.S., Young, M.K., Schwartz, M.K., Lowe, W.H., ... Whiteley, A.R. (2015) Distance, flow and PCR inhibition: eDNA dynamics in two headwater streams. *Molecular Ecology Resources,* **15,** 216-227.

Jerde, C.L., Chadderton, W.L., Mahon, A.R., Renshaw, M.A., Corush, J., Budny, M.L., ... Lodge, D.M. (2013) Detection of Asian carp DNA as part of a Great Lakes basin-wide surveillance program. *Canadian Journal of Fisheries and Aquatic Sciences,* **70,** 522-526.

Jerde, C.L., Mahon, A.R., Chadderton, W.L. & Lodge, D.M. (2011) "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters,* **4,** 150-157.

Jerde, C.L., Olds, B.P., Shogren, A.J., Andruszkiewicz, E.A., Mahon, A.R., Bolster, D. & Tank, J.L. (2016) Influence of stream bottom substrate on retention and transport of vertebrate environmental DNA. *Environmental Science & Technology,* **50,** 8770-8779.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., ... Yu, D.W. (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters,* **16,** 1245-1257.

Jo, T., Murakami, H., Masuda, R., Sakata, M.K., Yamamoto, S. & Minamoto, T. (2017) Rapid degradation of longer DNA fragments enables the improved estimation of distribution and biomass using environmental DNA. *Molecular Ecology Resources,* **17,** e25-e33.

Jones, M., Ghoorah, A. & Blaxter, M. (2011) jMOTU and taxonerator: turning DNA barcode sequences into annotated operational taxonomic units. *PLoS ONE,* **6,** e19259.

Jones, M.R. & Good, J.M. (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology,* **25,** 185-202.

Kassambara, A. & Mundt, F. (2017) factoextra: extract and visualize the results of multivariate data analyses: an R package in CRAN. Available from https://CRAN.R-project.org/package=factoextra.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution,* **30,** 772-780.

Kelly, F.L., Harrison, A.J., Allen, M., Connor, L. & Rosell, R. (2012) Development and application of an ecological classification tool for fish in lakes in Ireland. *Ecological Indicators,* **18,** 608-619.

Kelly, R.P., Gallego, R. & Jacobs-Palmer, E. (2018) The effect of tides on nearshore environmental DNA. *PeerJ,* **6,** e4521.

Kelly, R.P., Port, J.A., Yamahara, K.M. & Crowder, L.B. (2014) Using environmental DNA to census marine fishes in a large mesocosm. *PLoS ONE,* **9,** e86175.

Keskin, E. (2014) Detection of invasive freshwater fish species using environmental DNA survey. *Biochemical Systematics and Ecology,* **56,** 68-74.

Kitson, J.J.N., Hahn, C., Sands, R.J., Straw, N.A., Evans, D.M. & Lunt, D.H. (2019) Detecting host-parasitoid interactions in an invasive Lepidopteran using nested tagging DNA-metabarcoding. *Molecular Ecology,* **28,** 471-483.

Klymus, K.E., Richter, C.A., Chapman, D.C. & Paukert, C. (2015) Quantification of eDNA shedding rates from invasive bighead carp *Hypophthalmichthys nobilis* and silver carp *Hypophthalmichthys molitrix*. *Biological Conservation,* **183,** 77-84.

Kocher, A., Thoisy, B., Catzeflis, F., Huguin, M., Valière, S., Zinger, L., ... Murienne, J. (2017) Evaluation of short mitochondrial metabarcodes for the identification of Amazonian mammals. *Methods in Ecology and Evolution,* **8,** 1276-1283.

Kocher, T.D., Thomas, W.K., Meyer, A., Edwards, S.V., Pääbo, S., Villablanca, F.X. & Wilson, A.C. (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences,* **86,** 6196-6200.

Kõljalg, U., Larsson, K.H., Abarenkov, K., Nilsson, R.H., Alexander, I.J., Eberhardt, U., ... Larsson, E. (2005) UNITE: a database providing web - based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist,* **166,** 1063-1068.

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

Kozlov, A.M., Zhang, J., Yilmaz, P., Glöckner, F.O. & Stamatakis, A. (2016) Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research,* **44,** 5022-5033.

Kubecka, J., Hohausova, E., Matena, J., Peterka, J., Amarasinghe, U.S., Bonar, S.A., ... Winfield, I.J. (2009) The true picture of a lake or reservoir fish stock: a review of needs and progress. *Fisheries Research,* **96,** 1-5.

Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology,* **12,** 118-123.

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

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

Lance, R.F., Klymus, K.E., Richter, C.A., Guan, X., Farrington, H.L., Carr, M.R., ... Baerwaldt, K.L. (2017) Experimental observations on the decay of environmental DNA from bighead and silver carps. *Management of Biological Invasions,* **8,** 343-359.

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E.M. & Troedsson, C. (2017) DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *PLoS ONE,* **12,** e0179443.

Lawson Handley, L. (2015) How will the 'molecular revolution' contribute to biological recording? *Biological Journal of the Linnean Society,* **115,** 750-766.

Lawson Handley, L.J., Read, D., Winfield, I., Kimbell, H., Johnson, H., Li, J., ... Hänfling, B. (2019) Temporal and spatial variation in distribution of fish environmental DNA in England's largest lake. *Environmental DNA,* https://doi.org/10.1002/edn3.5.

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., ... Costa, F.O. (2018) Why we need sustainable networks bridging countries, disciplines, cultures and generations for aquatic biomonitoring 2.0: a perspective derived from the DNAqua-Net COST action. *Advances in Ecological Research,* **58,** 63-99.

Levy-Booth, D.J., Campbell, R.G., Gulden, R.H., Hart, M.M., Powell, J.R., Klironomos, J.N., ... Dunfield, K.E. (2007) Cycling of extracellular DNA in the soil environment. *Soil Biology and Biochemistry,* **39,** 2977-2991.

Li, J., Hatton-Ellis, T.W., Handley, L.-J.L., Kimbell, H.S., Benucci, M., Peirson, G. & Hänfling, B. (2019) Ground-truthing of a fish-based environmental DNA metabarcoding method for assessing the quality of lakes. *Journal of Applied Ecology***,** https://doi.org/10.1111/1365-2664.13352.

Li, J., Lawson Handley, L.J., Read, D.S. & Hänfling, B. (2018a) The effect of filtration method on the efficiency of environmental DNA capture and quantification via metabarcoding. *Molecular Ecology Resources,* **18,** 1102-1114.

Li, Y., Evans, N.T., Renshaw, M.A., Jerde, C.L., Olds, B.P., Shogren, A.J., ... Pfrender, M.E. (2018b) Estimating fish alpha- and beta-diversity along a small stream with environmental DNA metabarcoding. *Metabarcoding and Metagenomics,* **2,** e24262.

Liang, Z. & Keeley, A. (2013) Filtration recovery of extracellular DNA from environmental water samples. *Environmental Science & Technology,* **47,** 9324-9331.

Lim, N.K., Tay, Y.C., Srivathsan, A., Tan, J.W., Kwik, J.T., Baloğlu, B., ... Yeo, D.C. (2016) Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal Society Open Science,* **3,** 160635.

Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., ... Tang, M. (2016) Mitochondrial capture enriches mito‐DNA 100 fold, enabling PCR‐free mitogenomics biodiversity analysis. *Molecular Ecology Resources,* **16,** 470-479.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J. & Pallen, M.J. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology,* **30,** 434-439.

Lundberg, J.G., Kottelat, M., Smith, G.R., Stiassny, M.L. & Gill, A.C. (2000) So many fishes, so little time: an overview of recent ichthyological discovery in continental waters. *Annals of the Missouri Botanical Garden***,** 26-62.

Mächler, E., Deiner, K., Spahn, F. & Altermatt, F. (2015) Fishing in the water: effect of sampled water volume on environmental DNA-based detection of macroinvertebrates. *Environmental Science & Technology,* **50,** 305-312.

Mächler, E., Deiner, K., Steinmann, P. & Altermatt, F. (2014) Utility of environmental DNA for monitoring rare and indicator macroinvertebrate species. *Freshwater Science,* **33,** 1174-1183.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology,* **83,** 2248-2255.

MacKenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology,* **42,** 1105-1114.

Magoč, T. & Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics,* **27,** 2957-2963.

Mahon, A.R., Jerde, C.L., Galaska, M., Bergner, J.L., Chadderton, W.L., Lodge, D.M., ... Nico, L.G. (2013) Validation of eDNA surveillance sensitivity for detection of Asian carps in controlled and field experiments. *PLoS ONE,* **8,** e58316.

Majaneva, M., Diserud, O.H., Eagle, S.H.C., Bostrom, E., Hajibabaei, M. & Ekrem, T. (2018) Environmental DNA filtration techniques affect recovered biodiversity. *Scientific Reports,* **8,** 4682.

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M. & Minamoto, T. (2014) The release rate of environmental DNA from juvenile and adult fish. *PLoS ONE,* **9,** e114639.

Mateo, P., Leganés, F., Perona, E., Loza, V. & Fernández-Piñas, F. (2015) Cyanobacteria as bioindicators and bioreporters of environmental analysis in aquatic ecosystems. *Biodiversity and Conservation,* **24,** 909-948.

Matsen, F.A., Kodner, R.B. & Armbrust, E.V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics,* **11,** 538.

McCarthy, I. (2007) The Welsh Torgoch (*Salvelinus alpinus*): a short review of its distribution and ecology. *Ecology of Freshwater Fish,* **16,** 34-40.

McKee, A.M., Spear, S.F. & Pierson, T.W. (2015) The effect of dilution and the use of a post-extraction nucleic acid purification column on the accuracy, precision, and inhibition of environmental DNA samples. *Biological Conservation,* **183,** 70-76.

Milner, N. (1983) Diver observations on Charr (*Salvelinus alpinus* L.) spawning grounds in Llyn Cwellyn, North Wales. Welsh Water Authority.

Minamoto, T., Naka, T., Moji, K. & Maruyama, A. (2016) Techniques for the practical collection of environmental DNA: filter selection, preservation, and extraction. *Limnology,* **17,** 23-32.

Mioduchowska, M., Czyż, M.J., Gołdyn, B., Kur, J. & Sell, J. (2018) Instances of erroneous DNA barcoding of metazoan invertebrates: are universal *cox1* gene primers too "universal"? *PloS ONE,* **13,** e0199609.

Miya, M., Minamoto, T., Yamanaka, H., Oka, S.-i., Sato, K., Yamamoto, S., ... Doi, H. (2016) Use of a filter cartridge for filtration of water samples and extraction of environmental DNA. *Journal of Visualized Experiments,* **117,** e54741.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J.Y., Sato, K., ... Iwasaki, W. (2015) MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science,* **2,** 150088.

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

Nilsson, R.H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K.-H. & Kõljalg, U. (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PloS ONE,* **1,** e59.

O'Brien, H.E., Parrent, J.L., Jackson, J.A., Moncalvo, J.M. & Vilgalys, R. (2005) Fungal community analysis by large-scale sequencing of environmental samples. *Applied and Environmental Microbiology,* **71,** 5544-5550.

O'Donnell, J.L., Kelly, R.P., Shelton, A.O., Samhouri, J.F., Lowell, N.C. & Williams, G.D. (2017) Spatial distribution of environmental DNA in a nearshore marine habitat. *PeerJ,* **5,** e3044.

Ogram, A., Sayler, G.S. & Barkay, T. (1987) The extraction and purification of microbial DNA from sediments. *Journal of Microbiological Methods,* **7,** 57-66.

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

Olds, B.P., Jerde, C.L., Renshaw, M.A., Li, Y.Y., Evans, N.T., Turner, C.R., ... Lamberti, G.A. (2016) Estimating species richness using environmental DNA. *Ecology and Evolution,* **6,** 4214-4226.

Pall, K. & Moser, V. (2009) Austrian Index Macrophytes (AIM-Module 1) for lakes: a Water Framework Directive compliant assessment system for lakes using aquatic macrophytes. *Hydrobiologia,* **633,** 83-104.

Parducci, L., Jorgensen, T., Tollefsrud, M.M., Elverland, E., Alm, T., Fontana, S.L., ... Willerslev, E. (2012) Glacial survival of boreal trees in northern Scandinavia. *Science,* **335,** 1083-1086.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., ... Domaizon, I. (2018) The future of biotic indices in the ecogenomic era: Integrating (e) DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment,* **637,** 1295-1310.

Pečnikar, Ž.F. & Buzan, E.V. (2014) 20 years since the introduction of DNA barcoding: from theory to application. *Journal of Applied Genetics,* **55,** 43-52.

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

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R., Scholes, R.J., ... Cardoso, A. (2013) Essential biodiversity variables. *Science,* **339,** 277-278.

Piaggio, A.J., Engeman, R.M., Hopken, M.W., Humphrey, J.S., Keacher, K.L., Bruce, W.E. & Avery, M.L. (2014) Detecting an elusive invasive species: a diagnostic PCR to detect Burmese python in Florida waters and an assessment of persistence of environmental DNA. *Molecular Ecology Resources,* **14,** 374-380.

Pietramellara, G., Ascher, J., Borgogni, F., Ceccherini, M., Guerri, G. & Nannipieri, P. (2009) Extracellular DNA in soil and sediment: fate and ecological relevance. *Biology and Fertility of Soils,* **45,** 219-235.

Piggott, M.P. (2016) Evaluating the effects of laboratory protocols on eDNA detection probability for an endangered freshwater fish. *Ecology and evolution,* **6,** 2739-2750.

Pilliod, D.S., Goldberg, C.S., Arkle, R.S. & Waits, L.P. (2013) Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences,* **70,** 1123-1130.

Pilliod, D.S., Goldberg, C.S., Arkle, R.S. & Waits, L.P. (2014) Factors influencing detection of eDNA from a stream-dwelling amphibian. *Molecular Ecology Resources,* **14,** 109-116.

Piñol, J., Senar, M.A. & Symondson, W.O. (2018) The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*.

Polz, M.F. & Cavanaugh, C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology,* **64,** 3724-3730.

Port, J.A., O'Donnell, J.L., Romero-Maraccini, O.C., Leary, P.R., Litvin, S.Y., Nickols, K.J., ... Kelly, R.P. (2016) Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology,* **25,** 527-541.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution,* **26,** 1641-1650.

Proença, V., Martin, L.J., Pereira, H.M., Fernandez, M., McRae, L., Belnap, J., ... Gregory, R.D. (2017) Global biodiversity monitoring: from data sources to essential biodiversity variables. *Biological Conservation,* **213,** 256-263.

Qiu, X., Wu, L., Huang, H., McDonel, P.E., Palumbo, A.V., Tiedje, J.M. & Zhou, J. (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Applied and Environmental Microbiology,* **67,** 880-887.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F.O. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research,* **41,** D590-D596.

R_Core_Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

R_Core_Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org.

Ratnasingham, S. & Hebert, P.D. (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes,* **7,** 355-364.

Rees, H.C., Maddison, B.C., Middleditch, D.J., Patmore, J.R. & Gough, K.C. (2014) The detection of aquatic animal species using environmental DNA–a review of eDNA as a survey tool in ecology. *Journal of Applied Ecology,* **51,** 1450-1459.

Renshaw, M.A., Olds, B.P., Jerde, C.L., McVeigh, M.M. & Lodge, D.M. (2015) The room temperature preservation of filtered environmental DNA samples and assimilation into a phenol-chloroform-isoamyl alcohol DNA extraction. *Molecular Ecology Resources,* **15,** 168-176.

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P. & Coissac, E. (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research,* **39,** e145.

Robson, H.L., Noble, T.H., Saunders, R.J., Robson, S.K., Burrows, D.W. & Jerry, D.R. (2016) Fine-tuning for the tropics: application of eDNA technology for invasive fish detection in tropical freshwater ecosystems. *Molecular Ecology Resources,* **16,** 922-932.

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

Sassoubre, L.M., Yamahara, K.M., Gardner, L.D., Block, B.A. & Boehm, A.B. (2016) Quantification of environmental DNA (eDNA) shedding and decay rates for three marine fish. *Environmental Science & Technology,* **50,** 10456-10464.

Sato, H., Sogo, Y., Doi, H. & Yamanaka, H. (2017) Usefulness and limitations of sample pooling for environmental DNA metabarcoding of freshwater fish communities. *Scientific Reports,* **7,** 14860.

Sayers, E. (2008) E-utilities quick start. In: entrez programming utilities help. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25500/.

Scheffer, M., Carpenter, S., Foley, J.A., Folke, C. & Walker, B. (2001) Catastrophic shifts in ecosystems. *Nature,* **413,** 591-596.

Schloss, P.D., Gevers, D. & Westcott, S.L. (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS ONE,* **6,** e27310.

Schmidt, B.R., Kery, M., Ursenbacher, S., Hyman, O.J. & Collins, J.P. (2013) Site occupancy models in the analysis of environmental DNA presence/absence surveys: a case study of an emerging amphibian pathogen. *Methods in Ecology and Evolution,* **4,** 646-653.

Schnell, I.B., Bohmann, K. & Gilbert, M.T.P. (2015) Tag jumps illuminated ‐ reducing sequence ‐ to ‐ sample misidentifications in metabarcoding studies. *Molecular Ecology Resources,* **15,** 1289-1303.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., ... Crous, P.W. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences,* **109,** 6241-6246.

Scholes, R.J., Walters, M., Turak, E., Saarenmaa, H., Heip, C.H., Tuama, É.Ó., ... Jongman, R.H. (2012) Building a global observing system for biodiversity. *Current Opinion in Environmental Sustainability,* **4,** 139-146.

Self, M. (2016) Anglesey wetlands fish assessment-March 2016. Royal Society for the Protection of Birds.

Self, M. & Lyons, G. (2007) Anglesey wetlands fish assessment-October 2007. Royal Society for the Protection of Birds.

Self, M. & Muirhead, L. (2003) Fish in reedbeds survey report, Winter 2003. Royal Society for the Protection of Birds.

Seymour, M., Durance, I., Cosby, B.J., Ransom-Jones, E., Deiner, K., Ormerod, S.J., ... de Bruyn, M. (2018) Acidity promotes degradation of multi-species environmental DNA in lotic mesocosms. *Communications Biology,* **1,** 4.

Shaw, J.L., Clarke, L.J., Wedderburn, S.D., Barnes, T.C., Weyrich, L.S. & Cooper, A. (2016a) Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biological Conservation,* **197,** 131-138.

Shaw, J.L., Weyrich, L. & Cooper, A. (2016b) Using environmental (e) DNA sequencing for aquatic biodiversity surveys: a beginner's guide. *Marine and Freshwater Research,* **68,** 20–33.

Shendure, J. & Ji, H. (2008) Next-generation DNA sequencing. *Nature biotechnology,* **26,** 1135-1145.

Shogren, A.J., Tank, J.L., Andruszkiewicz, E., Olds, B., Mahon, A.R., Jerde, C.L. & Bolster, D. (2017) Controls on eDNA movement in streams: Transport, Retention, and Resuspension. *Scientific Reports,* **7,** 5065.

Shogren, A.J., Tank, J.L., Andruszkiewicz, E.A., Olds, B., Jerde, C. & Bolster, D. (2016) Modelling the transport of environmental DNA through a porous substrate using continuous flow-through column experiments. *Journal of the Royal Society Interface,* **13,** 20160290.

Shokralla, S., Gibson, J., King, I., Baird, D., Janzen, D., Hallwachs, W. & Hajibabaei, M. (2016) Environmental DNA barcode sequence capture: targeted, PCR-free sequence capture for biodiversity analysis from bulk environmental samples. *bioRxiv***,** https://doi.org/10.1101/087437.

Shokralla, S., Spall, J.L., Gibson, J.F. & Hajibabaei, M. (2012) Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology,* **21,** 1794-1805.

Spear, S.F., Groves, J.D., Williams, L.A. & Waits, L.P. (2015) Using environmental DNA methods to improve detectability in a hellbender (*Cryptobranchus alleganiensis*) monitoring program. *Biological Conservation,* **183,** 38-45.

Spens, J., Evans, A.R., Halfmaerten, D., Knudsen, S.W., Sengupta, M.E., Mak, S.S., ... Hellström, M. (2017) Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods in Ecology and Evolution,* **8,** 635-645.

Springer, M.S. & Douzery, E. (1996) Secondary structure and patterns of evolution among mammalian mitochondrial 12S rRNA molecules. *Journal of Molecular Evolution,* **43,** 357-373.

Stat, M., Huggett, M.J., Bernasconi, R., DiBattista, J.D., Berry, T.E., Newman, S.J., ... Bunce, M. (2017) Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports,* **7,** 12240.

Stéphane, D., Anne-Béatrice, D. & Jean, T. (2018) ade4: analysis of ecological data: exploratory and euclidean methods in environmental sciences. Retrieved from https://CRAN.R-project.org/package=ade4.

Stoeckle, B.C., Beggel, S., Cerwenka, A.F., Motivans, E., Kuehn, R. & Geist, J. (2017) A systematic approach to evaluate the influence of environmental conditions on eDNA detection success in aquatic ecosystems. *PloS ONE,* **12,** e0189119.

Strayer, D.L. & Dudgeon, D. (2010) Freshwater biodiversity conservation: recent progress and future challenges. *Journal of the North American Benthological Society,* **29,** 344-358.

Stribling, J.B., Pavlik, K.L., Holdsworth, S.M. & Leppo, E.W. (2008) Data quality, performance, and uncertainty in taxonomic identification for biological assessments. *Journal of the North American Benthological Society,* **27,** 906-919.

Strickler, K.M., Fremier, A.K. & Goldberg, C.S. (2015) Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biological Conservation,* **183,** 85-92.

Suyama, M., Torrents, D. & Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research,* **34,** W609-W612.

Suzuki, M.T. & Giovannoni, S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology,* **62,** 625-630.

Szitenberg, A., John, M., Blaxter, M.L. & Lunt, D.H. (2015) ReproPhylo: an environment for reproducible phylogenomics. *PLoS Computational Biology,* **11,** e1004447.

Taberlet, P., Bonin, A., Zinger, L. & Coissac, E. (2018) Environmental DNA: for biodiversity research and monitoring. pp. 253. Oxford University Press, Oxford.

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

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

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

Takahara, T., Minamoto, T., Yamanaka, H., Doi, H. & Kawabata, Z.i. (2012) Estimation of fish biomass using environmental DNA. *PLoS ONE,* **7,** e35868.

Takahashi, M.K., Meyer, M.J., Mcphee, C., Gaston, J.R., Venesky, M.D. & Case, B.F. (2018) Seasonal and diel signature of eastern hellbender environmental DNA. *The Journal of Wildlife Management,* **82,** 217-225.

Tansley, A.G. (1993) An introduction to plant ecology. pp. 228. Discovery Publishing House, New Delhi.

Taylor, H.R. & Harris, W.E. (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources,* **12,** 377-388.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Møller, P.R., Rasmussen, M. & Willerslev, E. (2012a) Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS ONE,* **7,** e41732.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Wiuf, C., Rasmussen, M., Gilbert, M.T.P., ... Willerslev, E. (2012b) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology,* **21,** 2565-2573.

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

Tillotson, M.D., Kelly, R.P., Duda, J.J., Hoy, M., Kralj, J. & Quinn, T.P. (2018) Concentrations of environmental DNA (eDNA) reflect spawning salmon abundance at fine spatial and temporal scales. *Biological Conservation,* **220,** 1-11.

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

Tsai, Y.-L. & Olson, B.H. (1992) Detection of low numbers of bacterial cells in soils and sediments by polymerase chain reaction. *Applied and Environmental Microbiology,* **58,** 754-757.

Turak, E., Harrison, I., Dudgeon, D., Abell, R., Bush, A., Darwall, W., ... Hermoso, V. (2017) Essential Biodiversity Variables for measuring change in global freshwater biodiversity. *Biological Conservation,* **213,** 272-279.

Turner, C.R., Barnes, M.A., Xu, C.C., Jones, S.E., Jerde, C.L. & Lodge, D.M. (2014) Particle size distribution and optimal capture of aqueous macrobial eDNA. *Methods in Ecology and Evolution,* **5,** 676-684.

Turner, C.R., Uy, K.L. & Everhart, R.C. (2015) Fish environmental DNA is more concentrated in aquatic sediments than surface water. *Biological Conservation,* **183,** 93-102.

Ushio, M., Murakami, H., Masuda, R., Sado, T., Miya, M., Sakurai, S., ... Kondoh, M. (2018) Quantitative monitoring of multispecies fish environmental DNA using high-throughput sequencing. *Metabarcoding and Metagenomics,* **2,** e23297.

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

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P.F., ... Dejean, T. (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology,* **25,** 929-942.

Vasselon, V., Rimet, F., Tapolczai, K. & 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.

Vilgalys, R. (2003) Taxonomic misidentification in public DNA databases. *New Phytologist,* **160,** 4-5.

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., ... Liermann, C.R. (2010) Global threats to human water security and river biodiversity. *Nature,* **467,** 555.

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology,* **73,** 5261-5267.

White, H. (2000) The eutrophication of Llyn Penrhyn, Anglesey. M.Sc Thesis, Napier University.

Wickham, H. & Chang, W. (2016) ggplot2: create elegant data visualisations using the grammar of graphics. Retrieved from https://CRAN.R-project.org/package=ggplot2.

Wilcox, T.M., McKelvey, K.S., Young, M.K., Jane, S.F., Lowe, W.H., Whiteley, A.R. & Schwartz, M.K. (2013) Robust detection of rare species using environmental DNA: the importance of primer specificity. *PLoS ONE,* **8,** e59520.

Wilcox, T.M., McKelvey, K.S., Young, M.K., Lowe, W.H. & Schwartz, M.K. (2015) Environmental DNA particle size distribution from Brook Trout (Salvelinus fontinalis). *Conservation Genetics Resources,* **7,** 639-641.

Wilcox, T.M., McKelvey, K.S., Young, M.K., Sepulveda, A.J., Shepard, B.B., Jane, S.F., ... Schwartz, M.K. (2016) Understanding environmental DNA detection probabilities: a case study using a stream-dwelling char *Salvelinus fontinalis*. *Biological Conservation,* **194,** 209-216.

Wilcox, T.M., Zarn, K.E., Piggott, M.P., Young, M.K., McKelvey, K.S. & Schwartz, M.K. (2018) Capture enrichment of aquatic environmental DNA: a first proof of concept. *Molecular Ecology Resources,* **18,** 1392-1401.

Willerslev, E., Cappellini, E., Boomsma, W., Nielsen, R., Hebsgaard, M.B., Brand, T.B., ... Collins, M.J. (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science,* **317,** 111-114.

Willerslev, E. & Cooper, A. (2005) Ancient DNA. *Proceedings of the Royal Society of London B: Biological Sciences,* **272,** 3-16.

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M.E., ... Taberlet, P. (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature,* **506,** 47-51.

Willerslev, E., Hansen, A.J., Binladen, J., Brand, T.B., Gilbert, M.T., Shapiro, B., ... Cooper, A. (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science,* **300,** 791-795.

Winfield, I.J. (2002) Monitoring lake fish communities for the Water Framework Directive: a UK perspective. *TemaNord,* **566,** 69-72.

Winfield, I.J., Fletcher, J.M., James, J.B. & Bean, C.W. (2009) Assessment of fish populations in still waters using hydroacoustics and survey gill netting: experiences with Arctic charr (*Salvelinus alpinus*) in the UK. *Fisheries Research,* **96,** 30-38.

Yang, C., Wang, X., Miller, J.A., de Blécourt, M., Ji, Y., Yang, C., ... Yu, D.W. (2014) Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators,* **46,** 379-389.

Yoccoz, N.G. (2012) The future of environmental DNA in ecology. *Molecular Ecology,* **21,** 2031-2038.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012) Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution,* **3,** 613-623.

Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology,* **7,** 203-214.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., ... Huang, Q. (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience,* **2,** 4.