Environment Agency

Evidence

eDNA-based metabarcoding as a monitoring tool for fish in large lakes

Report – SC140018/R

We are the Environment Agency. We protect and improve the environment and make it a better place for people and wildlife.

We operate at the place where environmental change has its greatest impact on people's lives. We reduce the risks to people and properties from flooding; make sure there is enough water for people and wildlife; protect and improve air, land and water quality and apply the environmental standards within which industry can operate.

Acting to reduce climate change and helping people and wildlife adapt to its consequences are at the heart of all that we do.

We cannot do this alone. We work closely with a wide range of partners including government, business, local authorities, other agencies, civil society groups and the communities we serve.

This report is the result of research commissioned and funded by the Environment Agency.

Published by:
Environment Agency, Horizon House, Deanery Road, Bristol, BS1 5AH

www.environment-agency.gov.uk

ISBN: 978-1-84911-386-1

Further copies of this report are available from our publications catalogue:
www.gov.uk/government/publications

or our National Customer Contact Centre:
T: 03708 506506

Email: enquiries@environment-agency.gov.uk

Author(s):
Bernd Hänfling, Lori Lawson Handley, Daniel Read, Ian Winfield

Dissemination Status:
Publicly available

Research Contractors:
Dr. Bernd Hänfling
Senior Lecturer in Ecology and Evolution
Director of Postgraduate Studies
Evolutionary Biology Group
School of Biological, Biomedical and Environmental Sciences
University of Hull
Hull, HU6 7RX
UK

Ian J. winfield
Lake Ecosystems Group
Centre for Ecology & Hydrology
Lancaster Environment Centre
Library Avenue
Bailrigg, Lancaster
Lancashire LA1 4AP
U.K.

Environment Agency's Project Manager:
Graeme Peirson and Kerry Walsh

Project Number:
SC140018

Evidence at the Environment Agency

Evidence underpins the work of the Environment Agency. It provides an up-to-date understanding of the world about us, helps us to develop tools and techniques to monitor and manage our environment as efficiently and effectively as possible. It also helps us to understand how the environment is changing and to identify what the future pressures may be.

The work of the Environment Agency's evidence teams are a key ingredient in the partnership between research, guidance and operations that enables the Environment Agency to protect and restore our environment by:

- **Setting the agenda**, by providing the evidence for decisions;
- **Maintaining scientific credibility**, by ensuring that our programmes and projects are fit for purpose and executed according to international standards;
- **Carrying out research**, either by contracting it out to research organisations and consultancies or by doing it ourselves;
- **Delivering information, advice, tools and techniques**, by making appropriate products available.

Doug Wilson

Director of Research Analysis and Evaluation

Executive summary

Organisms continuously release DNA into their environments in the form of shed cells, waste matter, blood, gametes and decaying material. Analysis of this 'environmental DNA' (eDNA) is revolutionising the way biodiversity is monitored. This exceptional new technology has been rapidly adopted for targeted monitoring of single species and studies have consistently shown that it outperforms traditional survey methods in terms of ease of sampling, sensitivity and cost.

Recent developments in next generation sequencing enable DNA from whole communities of organisms to be sequenced simultaneously ('metabarcoding'). Metabarcoding is routinely used for direct analysis of microbial and meiofaunal diversity, but this approach has so far rarely been used to analyse eDNA. eDNA metabarcoding is potentially the most promising technological advance for biodiversity measurement. However, the few studies carried out to date have been performed in aquaria or on a very small scale in natural settings. One of the most promising aspects of eDNA is its potential for detecting rare or elusive species that are challenging to monitor using established survey methods. There is therefore considerable interest in using eDNA to supplement existing survey methods for routine monitoring. However, important questions remain about how sensitive and accurate the method is for detecting species and for estimating their abundance or relative biomass.

The European Union Water Framework Directive requires assessment of the ecological status of lakes, rivers and groundwaters based on biological elements including phytoplankton, macrophytes and phytobenthos, benthic invertebrates and fish. The current fish classification tool for England, Wales and Scotland (Fisheries Classification Scheme 2, FCS2) was devised for rivers and uses electrofishing survey methods to provide information on the composition and abundance, as well as the age structure of fish populations. Routine monitoring of lake fish populations for the Water Framework Directive is desirable but not yet feasible with existing tools and resources. The aim of this project was to investigate the potential of eDNA metabarcoding as a fish classification tool for large UK lakes. The study investigated whether eDNA metabarcoding could recover information on fish presence/absence and abundance by comparison with long-term data and recent gill net surveys in the English Lake District.

Sixty-six 2-litre water samples were collected along 3 offshore transects and 6 shoreline sites in Lake Windermere in January 2015. Twelve additional samples were collected at shoreline and gill net sites in Bassenthwaite Lake and Derwent Water. Samples were filtered, the eDNA extracted and 2 target gene regions (mitochondrial cytochrome b and 12S) amplified using polymerase chain reaction (PCR). PCR products were sequenced on an Illumina MiSeq Next Generation Sequencer and data analysed using customised bioinformatics pipelines. eDNA data were compared with fish abundance and biomass data from recent gill net surveys (September 2014) and rank abundance based on expert opinion from hydroacoustic and gill net surveys.

The results demonstrate that eDNA metabarcoding is extremely sensitive for detecting species. Fourteen of the 16 species previously recorded in Windermere (based on long-term datasets) were detected using eDNA, compared with only 4 species in the September 2014 gill net survey. eDNA also outperformed gill net surveys in terms of species detection in Bassenthwaite Lake and Derwent Water, despite limited sampling in these water bodies.

The analyses indicate that 10–20 samples may be adequate to accurately represent the species present. Moreover, shoreline sampling may be sufficient for assessing presence/absence since 12 species were detected in just 6 samples collected along a short stretch of shoreline in Windermere.

It is clear that more comprehensive sampling is required for estimating fish abundance. The number of sequence reads per species and the proportion of sampling sites in which a species was detected were used as proxies for abundance from eDNA data and compared with data from established surveys. eDNA abundance data consistently correlated with abundance/biomass estimates from established surveys, suggesting that the potential for extracting quantitative estimates from eDNA metabarcoding data may be greater than previously thought. However, correlations were only consistently statistically significant for Lake Windermere, where sampling was intense.

This study provides the first demonstration that eDNA metabarcoding can effectively classify community diversity in large lakes and has great potential to supplement existing biomonitoring programmes. Repeated sampling and statistical modelling, which were out of the scope of the present study, are needed to investigate the relationship between eDNA and actual abundance further.

Acknowledgements

We are extremely grateful to Ben James and Janice Fletcher for assistance with boat survey work and to Jianlong Li and Rosetta Blackman for help with sample collection and filtration. Sincere thanks are also given to Paul Nichols for exceptional help with the lab work, Anna Oliver for assistance with the next generation sequencing, and Christoph Hahn for the bioinformatics analysis.

Contents

1 Introduction

1.1 Background

Organisms release deoxyribonucleic acid (DNA) into their environments through waste, moulting, secreting mucous or releasing gametes. Recent advances in molecular techniques have made it feasible to detect and analyse traces of this free-floating DNA in the environment – referred to as 'environmental DNA' or 'eDNA'. Such techniques have great potential for biodiversity monitoring since they are non-invasive, can detect rare or elusive species that are difficult to detect using established methods, and can distinguish cryptic species or juvenile stages from different species that are difficult to identify taxonomically (Lawson Handley 2015).

To date, the great majority of eDNA studies have focused on targeted detection of one or a small number of species using assays designed to be species-specific. Targeted eDNA assays have proved highly successful in detecting individual species from a wide range of taxonomic groups including amphibians, fish, reptiles, mammals, crustaceans, molluscs, insects and fungal pathogens in aquatic environments.[1] For example, a recent eDNA study targeting great crested newts, *Triturus cristatus,* demonstrated high repeatability and substantially higher detection rates for eDNA compared with established survey methods (Biggs et al. 2015).

A targeted approach can be very useful if the aim is to survey one or a small number of species. However, species-specific assay design is not trivial and the technique is therefore not suitable for studying many species simultaneously. 'Next-generation' DNA sequencing (NGS) methods provide an alternative approach for analysing DNA from whole communities of organisms and do not require design of specific assays. Instead, a target region or 'barcode' is simultaneously sequenced from DNA of multiple organisms present in the sample on an NGS platform. This technique, called 'metabarcoding', generates hundreds of thousands to millions of DNA sequences in a single experiment. Resulting DNA sequences are then compared with existing DNA sequences in reference databases using bioinformatics pipelines to identify species present in the sample.

Metabarcoding is considered the leading technological advance for biodiversity measurement (Ji et al. 2013) and could lead to a shift in the focus of biodiversity monitoring away from reliance on indicator species. It is now routinely applied for analysing whole animal, fungal, bacterial and plant communities, and is revolutionising understanding of the diversity in understudied environments (Creer et al. 2010, Fonseca et al. 2010, Bik et al. 2012). However, the great majority of metabarcoding studies have so far been performed by direct sampling of microbial or microscopic animal communities. Direct metabarcoding requires destructive sampling and is not appropriate for general biodiversity monitoring of larger organisms such as vertebrates, large invertebrates and plants.

Metabarcoding of environmental DNA is a promising, alternative approach for monitoring of whole communities. So far very few eDNA metabarcoding studies have been carried out. These studies have either been performed in tanks or aquaria (Kelly et al. 2014, Mahon et al. 2014) or on a very small scale in natural settings (Thomsen et al. 2012a, Thomsen et al. 2012b).

Encouragingly, an eDNA metabarcoding survey carried out in coastal waters off Denmark detected DNA from 15 marine fish species, including both common and a rare, vagrant species, in just 3 half litre samples of seawater (Thomsen et al. 2012a).

[1] For a summary see Table 1 in Lawson Handley (2015).

Despite the small number of samples and volumes collected, eDNA outperformed 8 out of 9 conventional survey methods in terms of the number of species detected (Thomsen et al. 2012a). This study demonstrates the potential of eDNA for detecting elusive species that are often missed by conventional methods and is particularly encouraging given the obvious difficulties in sampling in the marine compared with freshwater environment.

Although eDNA metabarcoding appears to be extremely sensitive for detecting rare species and for describing presence/absence, important questions remain about its efficacy in obtaining accurate estimates of species abundance and biomass. This information is critical for end users involved in the monitoring and management of biodiversity or commercially exploited stocks.

Obtaining quantitative estimates from eDNA is challenging because of the large number of factors that influence DNA dynamics in the environment (reviewed by Barnes et al. 2014, Lawson Handley 2015) and because of the many opportunities for bias during laboratory steps (sampling, DNA extraction, polymerase chain reaction) and bioinformatics stages (Yu et al. 2012).

In principle, the number of sequences per taxon (or 'operational taxonomic unit') could be taken in metabarcoding studies as an estimator of species biomass. A recent aquarium-based study demonstrated a perfect correlation between rank abundance of eDNA sequences per taxon and rank biomass, but the actual number of sequence reads was not correlated to biomass (Kelly et al. 2014). The relationship between eDNA metabarcode data and abundance/biomass does not therefore appear to be a simple one. An alternative approach for estimating abundance is to carry out comprehensive sampling of a given environment and estimate the proportion of sites occupied by a species (MacKenzie et al. 2002, MacKenzie and Nichols, 2004). However, since species are not guaranteed to be detected even when present at a site, this simple measure typically underestimates true abundance (MacKenzie et al. 2002).

Repeated surveying of sites and statistical modelling (site occupancy modelling) is required to estimate the probability of detection and obtain more realistic estimates of abundance (MacKenzie et al. 2002). This approach has been advocated for eDNA studies (MacKenzie et al. 2002, Pilliod et al. 2013, Schmidt et al. 2013, Ficetola et al. 2015), but has not yet been trialled with eDNA metabarcoding.

1.2 Aims and objectives

The European Union Water Framework Directive requires assessment of the ecological status of lakes, rivers and groundwaters based on biological elements including phytoplankton, macrophytes and phytobenthos, benthic invertebrates and fish. The current fish classification tool for England, Wales and Scotland (Fisheries Classification Scheme 2, FCS2) was devised for rivers and uses electrofishing survey methods to provide information on the composition and abundance, as well as the age structure of fish populations. Routine monitoring of lake fish populations for the Water Framework Directive is desirable but not yet feasible with existing tools and resources.

The aim of the present study was to evaluate whether eDNA metabarcoding could provide a complementary tool for Water Framework Directive monitoring of freshwater fish in large lakes. In particular, the study investigated whether eDNA metabarcoding could recover accurate qualitative (presence/absence) and quantitative (abundance/biomass) information for lake fish monitoring compared with expert long-term datasets and recent gill net surveys.

The specific objectives were to:

- develop a reference database of fish species in UK lakes for the mitochondrial 12S ribosomal RNA (12S), cytochrome *b* (CytB) and cytochrome oxidase I (COI) genes, and to test several primer sets *in silico* and *in vitro*
- optimise methods for water sampling, filtration and eDNA extraction
- develop customised bioinformatics pipelines for analysis of eDNA metabarcoding data
- carry out comprehensive water sampling along depth–profile transects in Lake Windermere, and at gill net and shoreline locations within Windermere, Bassenthwaite Lake and Derwent Water to:
  - estimate the proportion of sites occupied by each species
  - provide a preliminary estimation of species abundance using eDNA metabarcoding
- compare the qualitative and quantitative results from eDNA metabarcoding with long-term expert opinion and recent gill net survey data

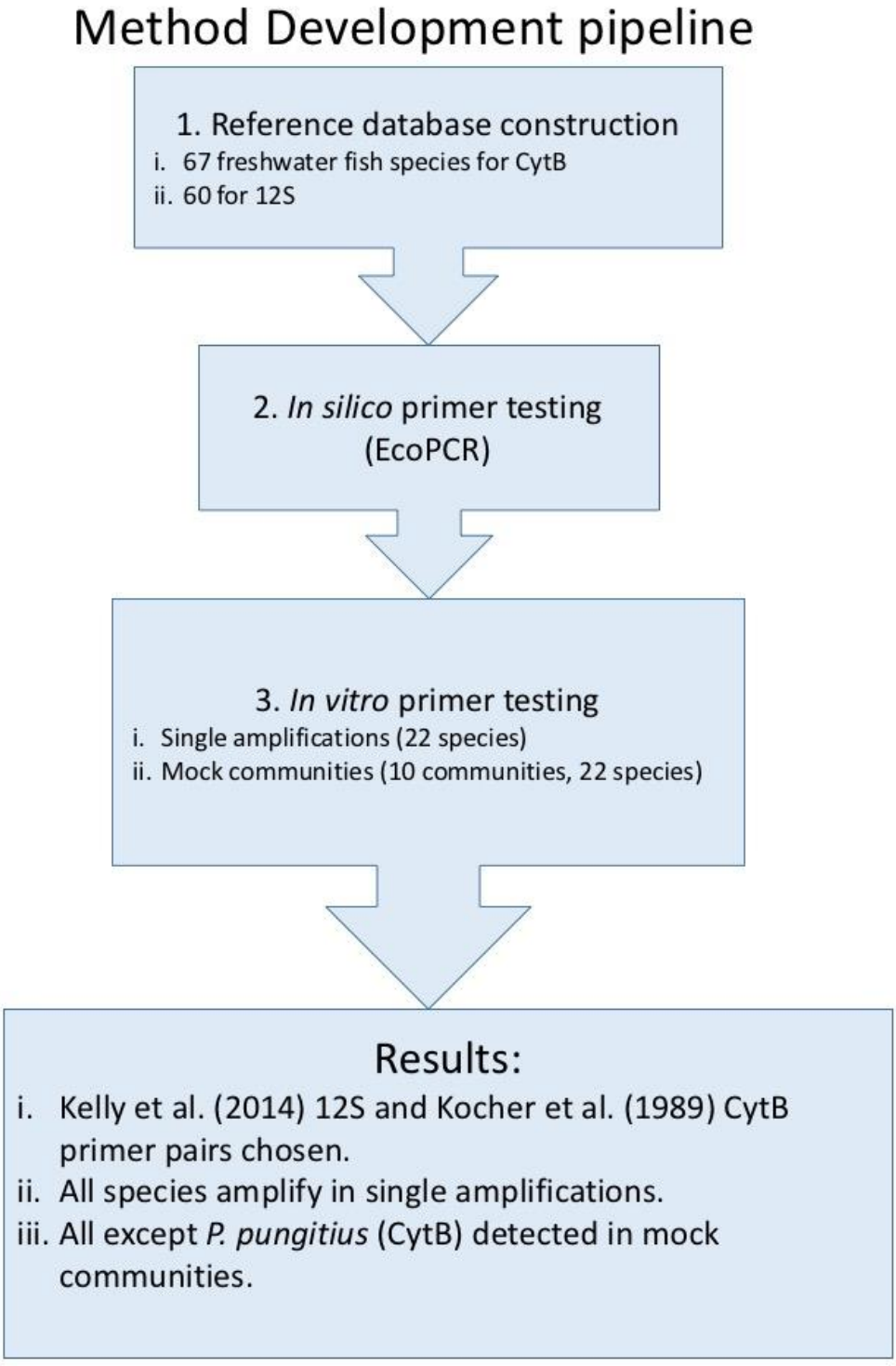
1.3 Project approach and workflows

The study consisted of 2 major workflows:

1) Development of a methodology for eDNA-based metabarcoding of lake fish communities, suitable for the purposes of the project
2) Testing the approach on the fish communities of 3 natural lakes (Windermere, Bassenthwaite Lake and Derwent Water)

The 3 lakes were chosen because they have been routinely monitored by the Centre for Ecology and Hydrology (CEH) for many years using standardised, established techniques (hydroacoustics, survey gill netting) and have the most comprehensive data on species composition and environmental metadata for lakes in the UK. This provided a unique opportunity to compare the results from NGS and established fish survey techniques.

Figure 1.1 presents an overview of the project and the two workflows, which had to take into account an extremely constrained time frame. The methods used in the two workflows are detailed in Section 2.


Method Development pipeline
1. Reference database construction
i. 67 freshwater fish species for CytB
ii. 60 for 12S
2. In silico primer testing (EcoPCR)
3. In vitro primer testing
i. Single amplifications (22 species)
ii. Mock communities (10 communities, 22 species)
Results:
i. Kelly et al. (2014) 12S and Kocher et al. (1989) CytB primer pairs chosen.
ii. All species amplify in single amplifications.
iii. All except P. pungitius (CytB) detected in mock communities.

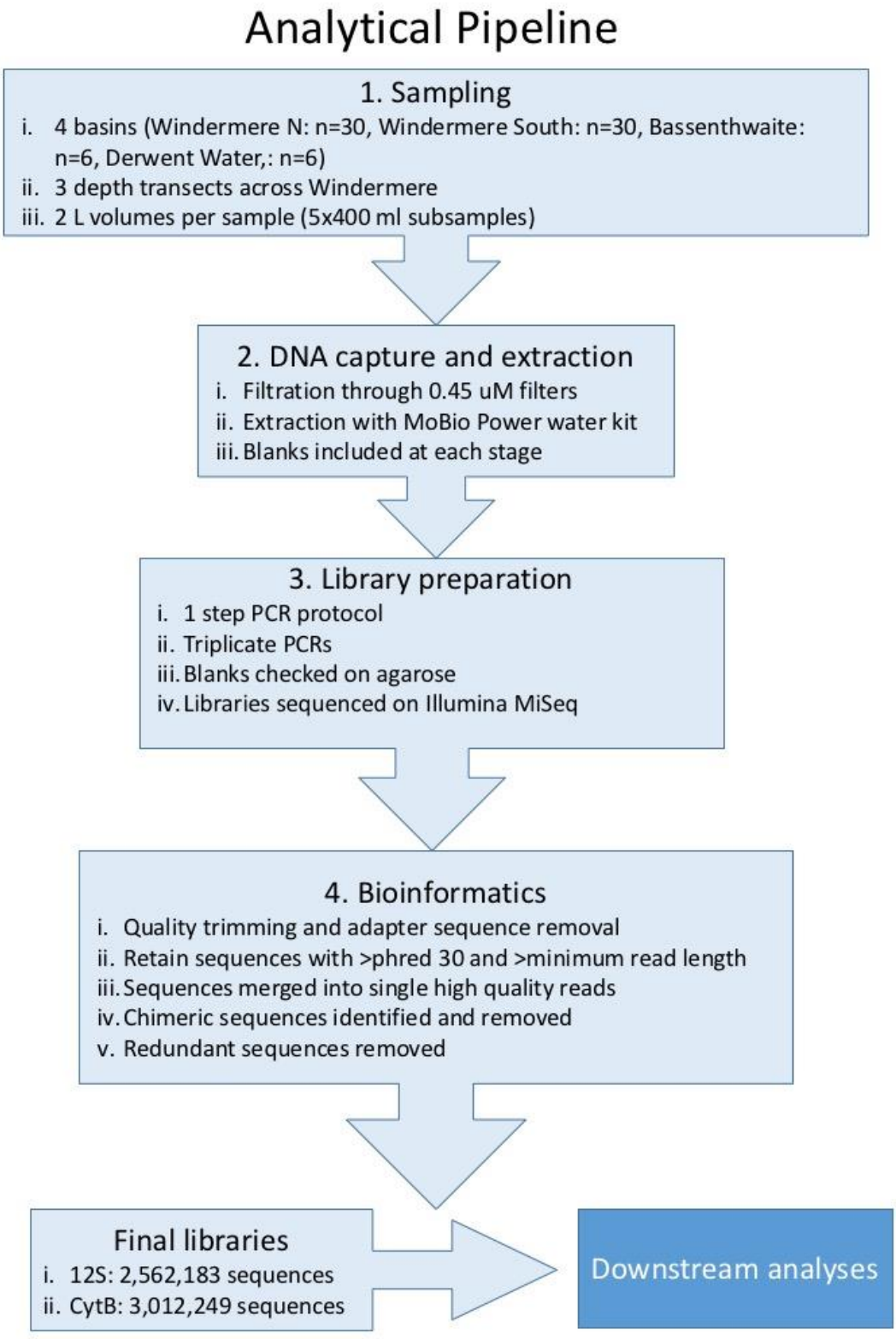


Figure 1.1 Project overview

Notes: 12S = mitochondrial 12S ribosomal RNA gene; COI = mitochondrial cytochrome oxidase 1 gene; CytB = mitochondrial cytochrome b gene; PCR = polymerase chain reaction

2 Material and methods

2.1 Method development

2.1.1 Compilation of reference databases

GenBank®[2], the genetic sequence database maintained by the US National Institutes of Health, was searched for reference sequences for 3 mitochondrial marker genes (12S, COI, CytB) for 67 fish species (Appendix A) using E-utilities (Sayers 2008) and the records downloaded in GenBank format. These species include all taxa previously recorded in British freshwaters and a range of species which could potentially be introduced in the near future.

Fresh tissue samples were collected for 24 species for which either full 12S gene sequences were missing on GenBank or which were used as positive controls (Appendix A).

A set of novel primers (12S_30F: CACTGAAGMTGYTAAGAYG and 12S_1380R: CTKGCTAAATCATGATGC) was designed from an alignment of whole mitochondrial fish genomes so as to amplify the entire 12S region.

Polymerase chain reactions (PCR) were undertaken in 25µl reaction volumes using Bioline's BIOTAQ™ DNA polymerase kit[3] containing 2.5µl 10× NH_4 reaction buffer, 1.0µl magnesium chloride ($MgCl_2$) (50mM), 2.5µl deoxynucleotide (dNTP) mix (10mM), 2.0µl of each primer (10mM), >100ng DNA template and 1 unit (U) BIOTAQ DNA polymerase, made up to 25µl with double-distilled water (ddH_2O). The PCR was conducted using the following thermal profile: one denaturation step at 95°C for 2 minutes, 30 cycles of 95°C for 30 seconds, 50°C for 30 seconds and 72°C for 50 seconds, followed by a final elongation step at 72°C for 10 minutes. Purified PCR products were Sanger sequenced directly (Macrogen Inc., Republic of Korea) in both directions using the PCR primers. The software CodonCode Aligner was used to quality check and edit individual sequences, and to align forward and reverse sequences of each individual. The sequences were subsequently converted into GenBank format and added to the reference database.

GenBank records and novel sequences were further processed in the ReproPhylo environment (Szitenberg et al. 2015). Records containing full mitochondrial genomes were cropped to retain exclusively 12S, CytB or COI, respectively. The remaining sequences were extracted in FASTA format and clustered at 100% identity to remove redundancy using the program CD-hit-est (Li and Godzik 2006).

As a final quality control, phylogenetic trees were inferred from the non-redundant sets of reference sequences for each marker gene in ReproPhylo (Szitenberg et al. 2015) as follows.

1. Sequences shorter than 400 base pairs (bp) were removed and the remaining sequences were aligned using MAFFT (Katoh and Standley 2013).

For COI and CytB records, nucleotide sequences were translated to protein sequences prior to alignment and aligned protein sequences were converted back to nucleotide sequences using Pal2Nal (Suyama et al. 2006).

[2] www.ncbi.nlm.nih.gov/genbank/
[3] www.bioline.com/uk/biotaq-dna-polymerase.html

2. Alignments were trimmed using trimAl.
3. Maximum likelihood trees were inferred with RAxML 8.0.2 (Stamatakis 2006) using the GTR+gamma model of substitutions.
4. The resulting trees were investigated manually to identify any sequence records that were obviously misplaced in the phylogenetic trees, that is, records, which were likely mislabelled. Such sequences were removed from the database as they were likely to cause conflicts in downstream analyses.
5. The remaining sequences, that is, the curated non-redundant reference databases, were used in all downstream analyses.

2.1.2 *In silico* testing of alternative minibarcoding primers

The curated non-redundant sets of reference sequences (12S, COI, CytB) were each converted into suitable format for subsequent *in silico* primer testing using the ECOPCRformat.py script from the ecoPCR package (Ficetola et al. 2010). The program EcoPrimers (Riaz et al. 2011) was used to identify novel primers for the targeted sequence length of 50–120 bp, based on the COI and CytB set of reference sequences. ecoPCR software was used for *in silico* evaluation of the performance of suitable novel primer pairs and 4 pairs of previously published primers (SPYGEN, France; Kelly et al. 2014, Thomsen et al. 2012a).

The primer pairs where tested specifically for conservation of the primer binding site (reflected in the parameter potential amplification success) and whether the variability of the amplified region is high enough to distinguish the target species. The primer pair L14841 and H15149 (Kocher et al. 1989) was also tested; this amplifies a 460 bp fragment of the CytB gene and has been used commonly for standard DNA barcoding of fishes. ecoPCR could not be applied to the CytB primers because the reference database did not cover the location of the forward primer. A maximum likelihood (ML) phylogeny of the non-redundant CytB database was therefore constructed which was cropped to the 460 bp using the methodology outlined above. This was used to evaluate the performance of the CytB primer pair using visual inspection of the ML tree.

2.1.3 *In vitro* and *in situ* testing of minibarcoding primers

A subset of 22 (33%) of the species from Appendix A was chosen to test the consistency of PCR amplification across taxa. Tissues were sourced from the existing collection at the University of Hull, or in some cases were specifically collected for this project.

Fish DNA was extracted from fin clips and muscle tissues using a DNeasy Blood & Tissue kit (Qiagen) and the DNA concentration was measured using a ND-1000 spectrophotometer (NanoDrop). PCR amplifications were carried out in 25µl reaction volumes using Bioline's BIOTAQ™ DNA polymerase kit containing 2.5µl 10× NH_4 reaction buffer, 1.0µl $MgCl_2$ (50mM), 2.5µl dNTP mix (10mM), 2.0µl of each primer (10mM), 2.0µl DNA template (5ng/µl) and 1U BIOTAQ DNA polymerase, made up to 25µl using ddH_2O. Thermal cycling conditions consisted of an initial denaturation (2 minutes at 95°C) followed by 30 cycles with 15 seconds at 95°C, 15 seconds at the optimal annealing temperature of the PCR primer pair and 20 seconds at 72°C, and a final extension step of 5 minutes at 72°C.

2.1.4 Testing of DNA capture and extraction method using samples from Thwaite Lake

Three 2.5 litre water samples were collected at separate locations at Thwaite Lake in Cottingham, East Yorkshire. Each water sample consisted of 5 subsamples of 500ml, which were pooled for the analysis. A 15ml aliquot from each sample was used for the precipitation method and 2 litres for the filtration method.

For filtration, all samples were filtered through a 0.45µm cellulose nitrate filter using Nalgene filtration units in combination with a vacuum pump. For precipitation, the method described by Ficetola et al. (2008) was used.

Both Qiagen's DNeasy Blood & Tissue kit and MoBio's PowerWater® kit for DNA extraction were tested. DNA quantity was measured using a ND-1000 spectrophotometer (NanoDrop).

Because this work had to be completed before the testing of a suitable minibarcoding primer (see Sections 2.1.1 and 2.1.2), 2 pairs of established barcoding primers (Folmer et al. 1994, Ivanova et al. 2007), which amplify a 650 bp region of COI, were used to evaluate the DNA quality obtained from the 2 different methods. The Folmer primers are generic and amplify a wide range of taxonomic groups, whereas the Ivanova primers are fairly fish specific.

2.2 Application of eDNA based fish metabarcoding in Cumbrian lakes

2.2.1 Water sampling and capture of eDNA

Sampling was carried out in 3 lakes (Windermere, Bassenthwaite Lake and Derwent Water) largely by boat (Figure 2.1a) using the principal method for water collection and the sampling design described below.

Principal method for water collection

At each site five samples were collected within proximity of 100 m using a Friedinger (Windermere) or Ruttner (Bassenthwaite Lake and Derwent Water) sampler at a specified depth (Figure 2.1b). The global positioning system (GPS) location was recorded at the sampling midpoint.

A 400ml aliquot was taken from each of the 5 samples (Figure 2.1c); these subsamples were pooled in a 2 litre sterile plastic bottle (Figure 2.1d) and stored in a Coleman cooler. Four 15ml subsamples were taken from each sample and transferred to 35ml 95% ethanol in 50ml Falcon tubes for precipitation.

Between samples, the equipment was sterilised by washing in 10% of a commercial bleach followed by 10% microsol (detergent) and rinsed with purified water. The sampler was rinsed again in lake water at the next sampling location.

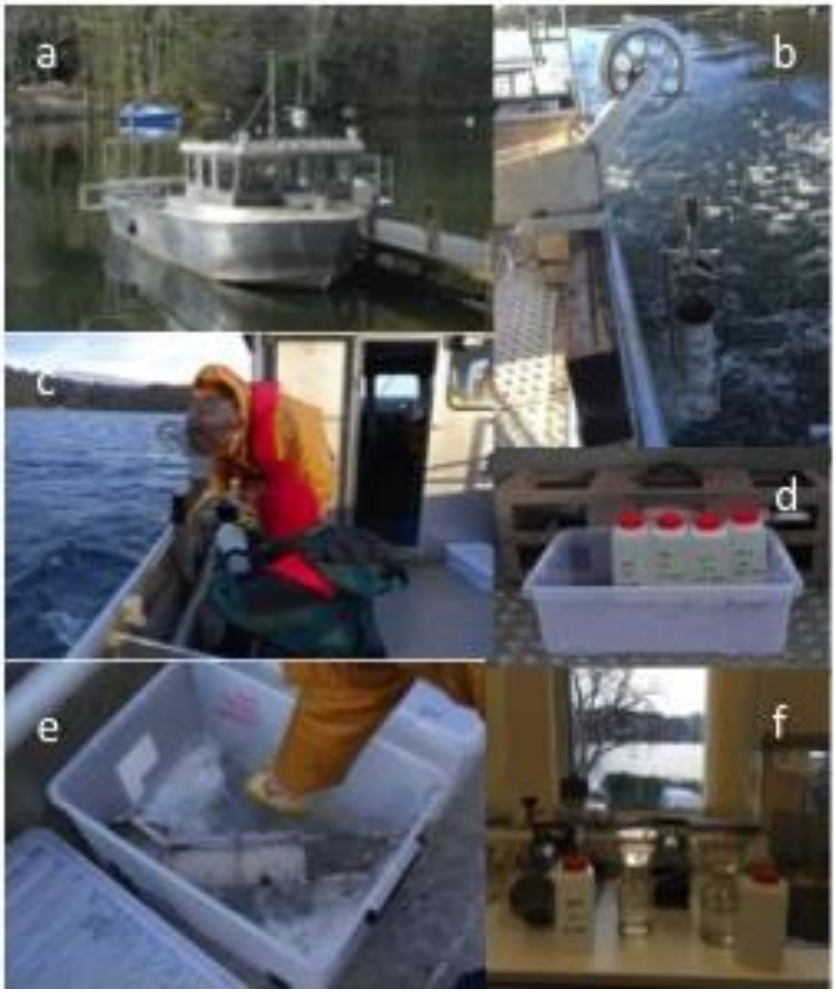


Figure 2.1 Overview of sampling methods: (a) boat; (b) Ruttner samples; (c) pooling subsamples; (d) sterile collection bottles; (e) treatment of equipment with bleach; and (f) water filtration units at the FBA's laboratory

Sampling design

Windermere

A total of 66 samples were collected from Windermere during a 3-day period from 28 to 30 January 2015. Most of these were collected along 3 transects with an approximately 1km sampling interval between sites.

The three transects ran along the 5m depth contour, the 20m depth contour and the lake midline respectively (Figure 2.2a: red, green and blue dots respectively). The sampling depth for these 3 transects was 2, 10 and 20m respectively.

This sampling scheme covered 7 of the 10 sites used for annual gill net surveys. A hydroacoustic survey was run during sampling of transects.

Water samples were also collected at the 3 remaining gill net sites (Figure 2.2a, orange dots) and 6 samples were collected from the shoreline near the Ferry Landing estate of the Freshwater Biological Association (FBA) on the western side of Windermere (Figure 2.2a, yellow dots).

At the deepest point along the midline transect in both North (63m) and South basin (45m) a depth profile was collected. The North basin transect was collected at 0-10-20-30-40-50-60m depth and the South basin transect was collected at 0-10-20-30-40m.

Bassenthwaite Lake and Derwent Water

Water sampling at 5 routine gill net survey sites and one additional shoreline site per lake was carried out at Bassenthwaite Lake (Figure 2.2b) and Derwent Water (Figure 2.2c) on 10 February 2015.

DNA capture and extraction

All samples were filtered through a 0.45μm cellulose nitrate filter using Nalgene filtration units in combination with a vacuum pump. All filtration equipment was sterilised in 10% bleach solution for 10 minutes after each filtration. Filtration blanks were run before the first filtration and then approximately after each sixth sample to test for possible contamination.

All 66 Windermere samples were collected and processed in CEH facilities at the FBA's Ferry Landing estate on Lake Windermere within 8 hours of collection. All samples from Bassenthwaite Lake (6 samples) and Derwent Water (6 samples) were collected on 10 February and processed at the laboratory at the University of Hull within 12 hours of collection.

DNA was extracted using the PowerWater DNA Isolation Kit (MoBio Laboratories, Inc. Carlsbad, USA)according to the manufacturer's instructions

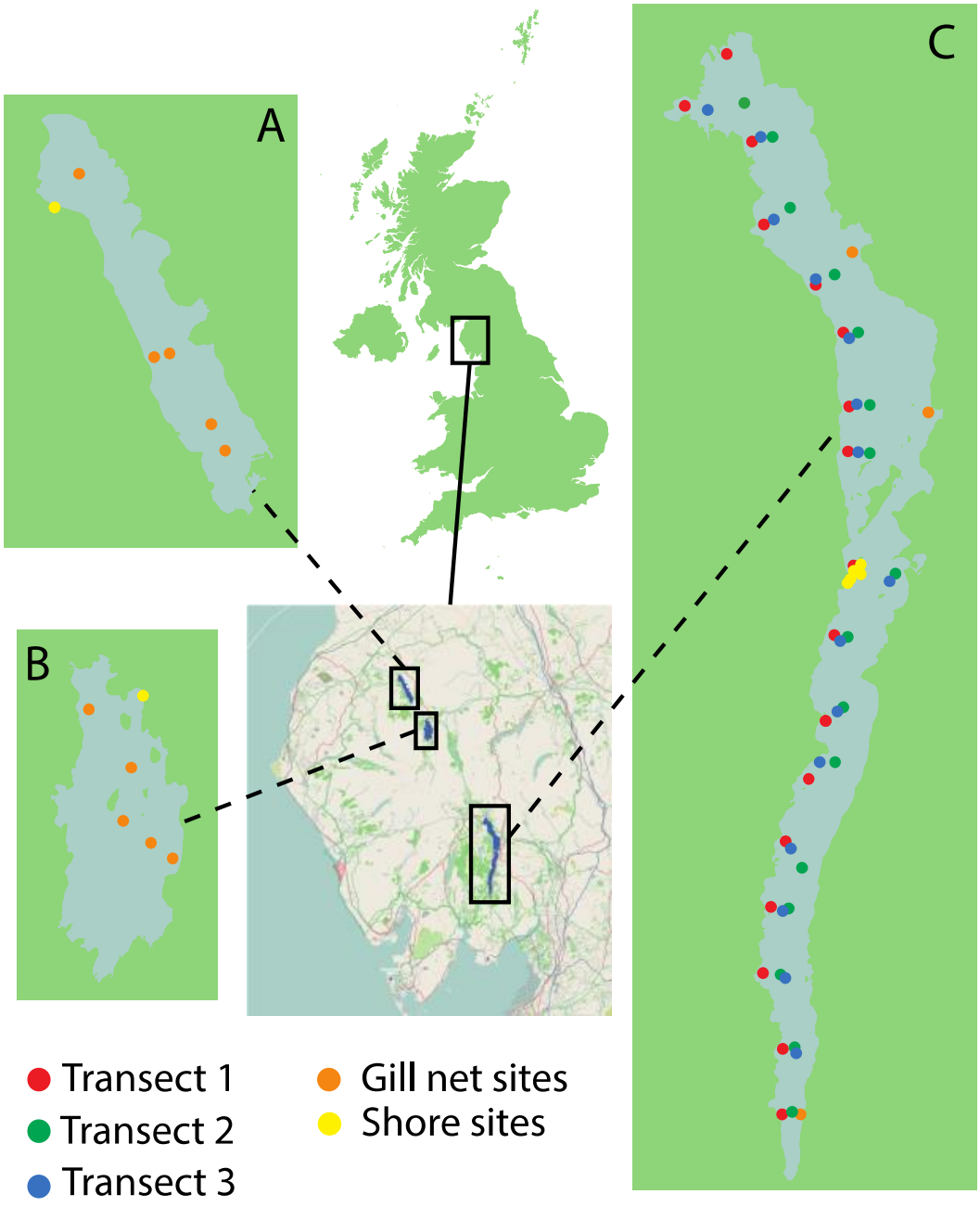


Figure 2.2 Overview of sampling sites: distribution of sampling sites at (a) Bassenthwaite Lake, (b) Derwent Water, and (c) Windermere,

Notes: Orange dots represent gill net sites.
Yellow dots represent shoreline samples.
Red, green and blue dots represent sites along transect 1, 2 and 3 respectively.

2.2.2 PCR using labelled metabarcoding primers and Illumina sequencing

For each of two loci which were chosen for the analysis of lake samples, 8 individually tagged forward primers and 12 individually tagged reverse primers were designed, allowing for 96 individually tagged combinations (one for each sample).

PCR reactions were carried out with all samples collected at Windermere, Bassenthwaite Lake and Derwent Water using the optimal conditions determined during *in vitro* testing (Section 2.1.3).

In addition, 10 mock communities were created from diluted DNA extractions from tissue samples and used as positive controls in the PCR. PCR of a set of negative controls for each primer pair (that is, all 96 combinations of tagged primers) and all collection and extraction blanks was also carried out.

PCRs were replicated 3 times for each sample and blank, and pooled to minimise bias in individual PCR reactions. The success of PCR reactions was checked through visual inspection of ethidium bromide stained agarose gels.

Two sequencing libraries (one for each locus) were prepared from all successfully amplified samples and run on the Illumina MiSeq desktop sequencer[4] using V3 2 × 300 bp chemistry. Each set of samples were normalised for concentration across the samples using the Life Technologies[5] SequalPrep™ Normalization Plate Kit and subsequently pooled to make a single sequencing library for each assay.

Each library was quantified by qPCR (average of 3 replicate quantifications) using the KAPA Library Quantification Kit on a Roche LightCycler Real-Time PCR machine according to the manufacturer's guidelines. In order improve clustering during the initial sequencing cycles 10% of PhiX[4] genomic library was added.

2.2.3 Bioinformatics analysis

The software program, Trimmomatic 0.32 (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 was above a Phred quality score of 30. All reads shorter than a defined minimum read length (12S: 90 bp; CytB: 100 bp) were discarded.

Sequence pairs were subsequently merged into single, high quality reads using the program FLASH 1.2.11 (Magoč and Salzberg 2011). To remove redundancy, sequences were clustered at 100% identity using VSEARCH 1.1.[6] Any singletons (that is, sequences occurring in only a single copy) were considered sequencing error and were omitted from further analyses. The remaining reads were screened for chimeric sequences against the curated reference databases using the 'uchime_ref' function implemented in vsearch 1.1 (https://github.com/torognes/vsearch). To remove redundancy, sequences were clustered at 100% identity using vsearch 1.1 (https://github.com/torognes/vsearch). Clusters represented by less than 3 sequences were considered sequencing error and were omitted from further analyses.

[4] www.illumina.com/systems/miseq.html
[5] Now part of Thermo Fisher Scientific.
[6] https://github.com/torognes/vsearch

Non-redundant sets of query sequences were then compared with the respective curated non-redundant reference database using the BLAST® (Basic Local Assignment Search Tool) developed by the National Center for Biotechnology Information (NCBI) in the USA (Zhang et al. 2000). BLAST output was analysed using a custom Python script, which implements a lowest common ancestor approach for taxonomic assignment similar to the strategy used by MEGAN (MEtaGenome ANalyzer) (Huson et al. 2007).

In brief, after the BLAST search the most significant matches to the reference database (yielding the top 10% bit-scores) for each of the query sequences were recorded. 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. Sequences for which the best BLAST hit had a bit score below 80 or had less than 95% identity to any sequence in the curated database, were considered non-target sequences. These were subjected to a separate BLAST search against the complete nucleotide database on GenBank.

Filtered data were summarised in two ways for downstream analyses: 1) the number of sequence reads per species at each site (hereon referred to as read counts) and 2) the proportion of sampling sites in which a given species was detected (hereon referred to as the site occupancy). To reduce the possibility of false positives, we only regarded a species as present at a given site if its sequence frequency exceeded a certain threshold level (proportion of all sequence reads in the sample). The choice of threshold level was guided by the analysis of sequence data from the mock communities. This analysis revealed that threshold levels of 0.3% and 1% were required for 12S and CytB respectively to omit all false positives in the mock communities (hereon referred to as Th100. At Th100 sequences of rare expected species were also lost from the mock community data and the lake samples. We therefore decided to apply slightly less conservative values of 0.1% and 0.2% for 12S and CytB respectively, at which over 90% of false positives were omitted in the mock communities to the main analysis of lake samples (Th90). We also investigated the potential extent of contamination from tag jumping in our libraries by exploring the distribution of PhiX assigned to target samples. The level of PhiX contamination in our samples also indicated that our thresholds were appropriate to eliminate most of false positives created during the sequencing process. In 95% of the 12S and CytB libraries the proportion of PhiX did not exceed 0.0015 and 0.001 respectively (with a corresponding maximum of 0.0023 and 0.0201).

All downstream analyses were performed in R v.3.1.3. (RCoreTeam 2015).

2.2.4 Estimating required sampling effort in Lake Windermere based on rarefaction

A sample-based rarefaction (Gotelli and Colwell 2010) of sequence read data was used to determine the number of samples needed to accurately represent the species assemblage. Rarefaction was performed with 499 randomisations in the R package Vegan (Oksanen et al. 2015) for the North and South Basins of Windermere separately and combined. Only sequences corresponding to the 16 species previously recorded in Windermere were included in these analyses.

2.2.5 Preparation of data from established surveys method

Hydroacoustics and survey gill netting

Data previously collected in September 2014 at Bassenthwaite Lake, Derwent Water and Windermere using the hydroacoustics and survey gill netting techniques described in detail in 2 NERC/CEH reports prepared for the Environment Agency (Winfield et al. 2014a, Winfield et al. 2014b) were assembled and analysed to produce simple summaries by lake. These activities consisted of:

- 6 gill netting sites (including a surface site directly above a deep water bottom site) at each lake or lake basin (thus 12 sites for Windermere)
- 10 night-time hydroacoustics transects at Bassenthwaite Lake
- 5 night-time hydroacoustics transects at Derwent Water
- 3 night-time hydroacoustics transects at Windermere North Basin
- 5 night-time hydroacoustics transects at Windermere South Basin

Additional information on species presence and approximate abundance

Given that the survey gill netting licensed at the 3 lakes is relatively limited, relevant other information on local fish species presence and abundance was assembled for Bassenthwaite Lake, Derwent Water and Windermere using the review data sources of Pickering (2001), Winfield et al. (1996), Winfield and Durie (2004), Winfield et al. (2010) and Maberly et al. (2011). This included the use of the expert opinion of CEH freshwater ecologist, Dr Ian Winfield, to assign each recorded species to an approximate abundance band ranging in units from 1 (very scarce) to 5 (very common).

2.2.6 Comparisons between data from established surveys and eDNA

A series of correlations was performed to:

- compare the data from established surveys and eDNA metabarcode data
- investigate whether eDNA data are sufficiently quantitative to provide estimates of relative abundance

Specifically, the relationship between eDNA data (the average number of sequence reads per species or proportion of sites occupied by a species) and data from established surveys (rank abundance or biomass based on long-term expert opinion or actual numbers from September 2014 gill net surveys) was investigated by calculating Spearman's rho (for rank correlations) and Pearson's product–moment correlation coefficient (for actual numbers) in R v3.1.3. The analyses were repeated for both loci and all 4 sampled basins.

3 Results

3.1 Method development

3.1.1 Compilation of reference databases

The complete reference database included a total of 747, 3034 and 4813 sequences (partial or complete) for the 3 markers 12S, COI and CytB respectively and covered all 67 target species for COI and CytB and 59 species for 12S. 12S sequences were not obtained for 7 species (*Aspius aspius, Coregonus autumnalis, Lampetra planeri, Misgurnus fossilis, Neogobius melanostomus, Proterorhinus semilunaris, Vimba vimba*). None of these species has previously been recorded in the target lakes and only one species (*L. planeri*) is confirmed to occur in the UK.

After curating, that is, removing redundant (that is, identical haplotypes) and likely mislabelled records (based on phylogenetic tree inference), the database contained 268 sequences for 12S, 687 sequences for COI and 2155 sequences for CytB. These were used as curated non-redundant reference databases in subsequent steps. The complete list of retained reference sequences for 12S and CytB is provided in a Microsoft® Excel spreadsheet (Appendix D).

3.1.2 *In silico* testing of alternative minibarcoding primers

A total of 10 primer pairs with suitable target length (40–110 bp) were identified for the COI and CytB regions. The *in silico* testing results of the 4 most promising combinations for each region are shown in Table 3.1.

All suitable novel primer pairs within the COI and CytB regions performed better in terms of potential amplification success and compared with the published 12S and CytB primers (Table 3.1).

None of the primer pairs could fully resolve all species pairs in the reference database. Some species pairs were poorly resolved by all or most primer pairs such as:

- 3 *Coregonus* species
- 3 Asian carp species of the genera *Hypophthalmichthys* and *Ctenopharyngodon*
- the American cat fish species of the genus *Ameiurus*
- *Lampetra planeri and L. fuviatilis*
- 2 of the native *Leuciscus* species

In addition, neither 12S primer combinations could resolve the 2 *Salvelinus* species and the Kelly primers could not distinguish between *Perca fluviatilis* and *Sander lucioperca*. This lack of resolution reflects the fact that the species pairs are very closely related and cannot therefore be clearly distinguished by short mitochondrial DNA (mtDNA) markers (for example, *Coregonus*) or that the 2 taxa are probably not reproductively isolated (for example, *Lampetra*).

Table 3.1 Summary of *in silico* testing results for published primers and primers designed in this study

Target region	Forward primer	Reverse primer	Fragment length (bp)	Species amplified (%)	Unresolved species pairs	Reference
12S	12S_F	12S_R	~70	74	1, 2, 4, 5, 6	SPYGEN
12S	12S_510F	12S_655R	~106	77	1, 2, 3, 6, 8	Kelly et al. 2014
CytB	Fish2bCBR	Fish2CBL	40	16	1, 2, 4, 6, 9	Thompson et al. 2012a
CytB	Fish2degCBL	Fish2CBR	40	23	1, 2, 4, 6, 9	Thompson et al. 2012a
CytB	CytB_F1	CytB_R1	110	91	1, 2, 9	This study
CytB	CytB_F1	CytB_R2	110	94	1, 2, 9	This study
CytB	CytB_F2	CytB_R3	50	95	1, 2, 5, 9	This study
CytB	CytB_F2	CytB_R4	50	95	1, 2, 5, 9	This study
COI	COI_F1	COI_R1	100	94	1, 4, 5, 7, 9	This study
COI	COI_F1	COI_R2	100	94	1, 4, 5, 7, 9	This study
COI	COI_F2	COI_R1	100	94	1, 4, 5, 7, 9	This study
COI	COI_F2	COI_R2	100	94	1, 4, 5, 7, 9	This study

Notes: Full primer sequences are given in Appendix B.
Unresolved species pairs: 1 = *Coregonus*;* 2 = *Hypophthalmichthys nobilis, H. molitrix*; 3 = *Ctenopharyngodon idella*, *H. molitrix*; 4 = *Ameiurus melas, A. nebulosus*; 5 = *Leuciscus idus, L. leuciscus*; 6 = *Salvelinus. alpinus, S. fontinalis*; 7 = *Alosa fallax, A. alosa*; 8 = *Perca fluviatilis, Sander lucioperca*; 9 = *Lampetra planeri, L. fluviatilis*
* Could not be tested in EcoPCR because the forward primer lies outside the reference sequences.

3.1.3 *In vitro* and *in situ* testing of minibarcoding primers

The published 12S minibarcoding primers worked reliably on all test samples under a range of annealing temperatures; the PCR products obtained under an annealing temperature (T_A) of 50°C are shown in Figure 3.1a and Figure 3.1b.

The CytB and COI minibarcoding primers designed in the present study only worked consistently for a subset of species. None of the newly designed primers amplified all species. Among those, the best results were obtained using the primer combination COInew_F2 and COInew_R2, which amplified all but one of the 22 species (Figure 3.1c). The published CytB barcoding primers from Kocher et al. (1989) amplified all 22 test samples successfully (Figure 3.2).

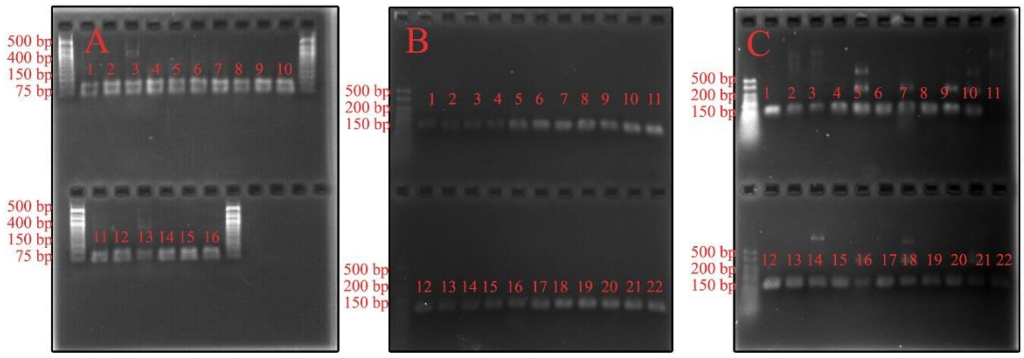


Figure 3.1 Results of *in vitro* testing of 12S and COI minibarcoding primers. Electrophoretic separation of PCR products (T_A 50°C) from 3 minibarcodes on 2.5% agarose gels, stained with ethidium bromide. Primer combinations: (A) 12S (SPYGEN patent); (B) 12S (Kelly et al 2014); and (C) COI (this study, COInew_F2 and COInew_R2)

Notes: The numbers in red on the gels indicate different species and correspond to those given in Appendix A in the column headed 'Species number in positive controls'.

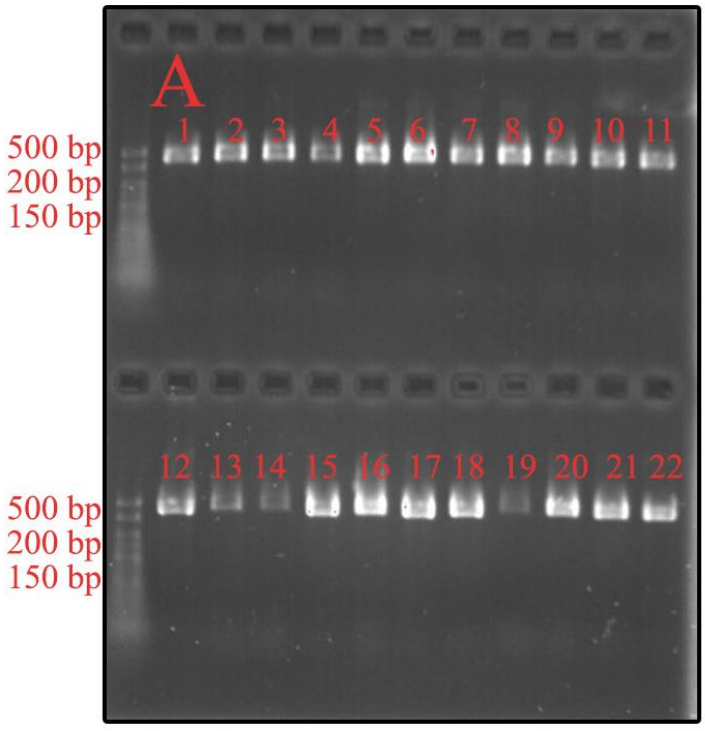


Figure 3.2 Results of *in vitro* testing of cytochrome b primers

Notes: Electrophoretic separation of PCR products from a 460 bp fragment of Cytb on 2.5% agarose gels, stained with ethidium bromide.
Primer combination from Kocher et al. (1989).
Numbers indicate different species and correspond to those given in Appendix A in the column headed 'Species number in positive controls'.

3.1.4 Testing of DNA capture and extraction method using samples from Thwaite Lake

DNA concentrations were, on average, higher in the filtrate extractions than in the precipitation extracts.

Positive amplifications for all sets of primers could be achieved for both extraction methods and all 3 primer combinations. However, the results for the filtration method combined with the MoBio PowerWater kit appeared more consistent and the bands were generally stronger; there were indications of PCR inhibition with the Qiagen extractions.

The combination of filtration and the MoBio PowerWater kit was therefore chosen for the application in the Cumbrian lakes.

3.2 Application of eDNA based fish metabarcoding in the Cumbrian lakes

3.2.1 PCR using labelled metabarcoding primers and Illumina sequencing

Amplifications of all target samples and positive controls were successful for both 12S and CytB. None of PCR negative controls, collection or filtration blanks showed any noticeable bands. Therefore sequencing libraries were created only for the target samples and positive controls.

The pooled and normalised samples were quantified by qPCR and run on an Illumina MiSeq using the V3 2 × 300 bp chemistry.

The 12S sequencing library contained 9.70 million raw reads of which 9.09 million passed initial quality filtering; 59.2% of this library was made up of Illumina's PhiX Control library[7] to help control for the expected low diversity of sequence reads. Hence the final number of sequences in the whole library was 3,708,720 across 96 samples (average 38,633 reads per sample). Although the number of reads varied across samples, even the sample with the lowest number of reads contained over 3,000 sequences.

The CytB sequencing library contained 12.57 million raw reads of which 10.99 million passed initial quality filtering. 74.1 % of this library comprised PhiX, so the final number was 2,846,410 sequences across 96 samples (average of 29,650 reads per sample).

3.2.2 Read count data overview

The bioinformatics analysis confirmed that the quality of the data was high and a large percentage of sequence reads for both CytB and 12S could be assigned either to species in the reference database (target sequences) or to non-target taxa using a BLAST search against the complete nucleotide database on GenBank.

The proportion of target sequences in the lake samples ranged from 3.4 to 33.7% (average 16.4%) for 12S and from 0 to 100% (average 49.0%) for CytB. In terms of sequence counts, this translates into a range of 3,454 to 50,322 fish sequences per site for 12S and 0 to 45,440 fish sequences per site for CytB. The full sequence count

[7] www.illumina.com/products/phix_control_v3.html

data for each primer dataset are attached to the report as spreadsheets (Appendix E for 12S and Appendix F for CytB).

Good consistency was found between 12S and Cytb in terms of both the site occupancy (SO, that is, the proportion of sites where a species was detected) and sequence read counts (RC). Data from the two loci were significantly correlated (Pearson's r consistently $P < 0.05$) for all basins, for both SO and RC (Figure 3.3). Consistent significant correlations were also found between SO and RC for each basin and locus, therefore only the results for SO are presented in the following main text. All results based on RC are provided in Appendix C (Figure S1).

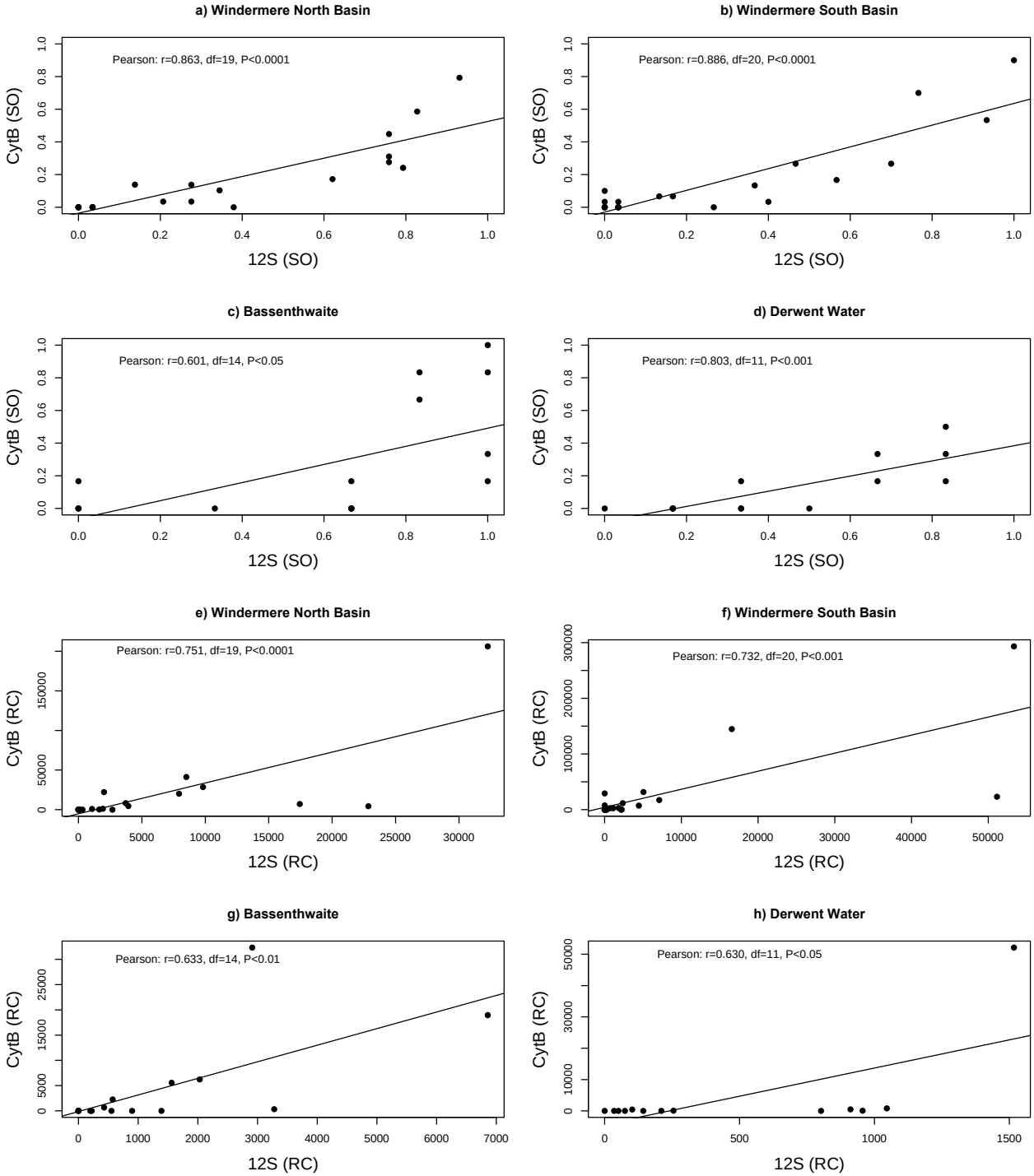


Figure 3.3 Correlations between 12S and CytB in terms of the site occupancy (SO, that is, proportion of sites where a species was detected, a-d) and read count (RC, i.e. the number of sequence reads, e-h) per species

3.2.3 Determining a threshold for defining the presence of species at individual sites

Across all sample sites within each lake, it was possible to identify 12S sequences from 20 species in Windermere, 10 species in Bassenthwaite Lake and 12 species in Derwent Water. A very similar picture emerged using CytB sequence data, although fewer species were identified across all sample sites within each basin. A total of 16 species were found in Windermere, 11 species in Bassenthwaite Lake and 6 species in Derwent Water. However, a number of species were represented with only a few sequences per site. In such cases, the possibility of a false positive identification through sequencing error or low level cross contamination pre- or post-PCR cannot be excluded.

To reduce the possibility of false positives, we only regarded a species as present at a given site if its sequence frequency exceeded a certain threshold level (proportion of all sequence reads in the sample).

Figures 3.4a and 3.4b shows the proportion of sites in Windermere where a species was identified as present based on 3 different detection thresholds for 12S and CytB sequence data respectively. For example, at a threshold of 0.001, a species was only regarded as present when the proportion of sequences exceeded 0.1%. At the highest threshold a number of species expected to be present were also lost from the data set

The choice of a suitable threshold value was therefore guided by an analysis of false positives in the mock communities (were the species composition was known). A threshold level was chosen were over 90% of false positives were omitted from the mock community data and no true positives were lost. This value was 0.001 for 12S and 0.002 for CytB. A detailed docmentation of this analysis is available in Hänfling et al. (*in press*)

The majority of potential false positives in the 12S data set were found in a single sample from Windermere North Basin, which was consequently omitted from all further analysis (sample W14, Appendix E).

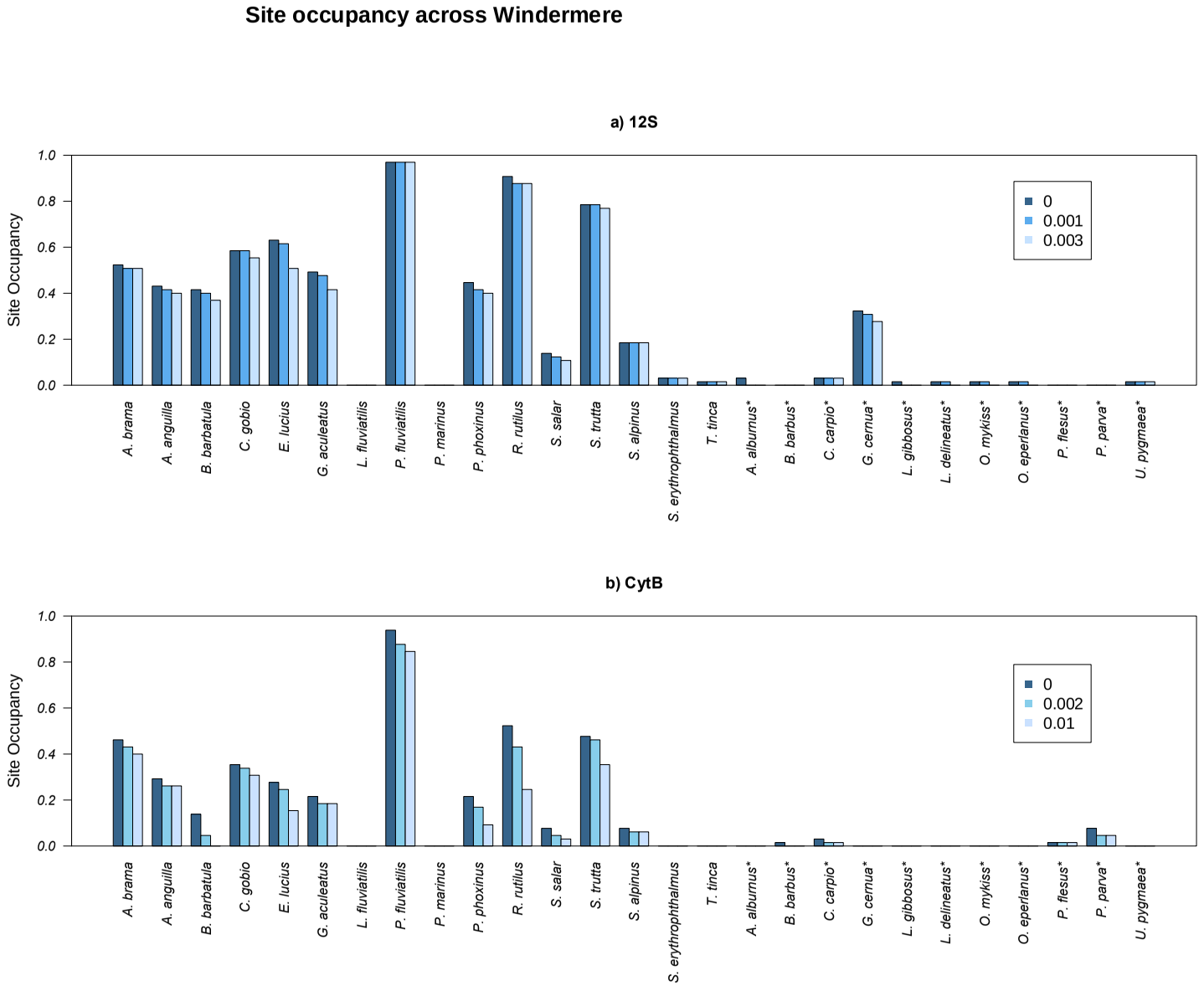


Figure 3.4.

Site occupancy across Windermere based on (a) 12S and (b) CytB sequence data using 3 different detection thresholds.

Notes: included are all species have previously been recorded in Windermere and all species which have been used in mock communities* (i.e. potential source of cross-contamination).

3.2.4 Fish abundance and distribution estimated from established surveys

Hydroacoustics

In September 2014, night-time total fish abundances were recorded as follows.

- At Bassenthwaite Lake, total fish abundance was 1,250.0 fish ha^{-1} (geometric mean with lower and upper 95% confidence limits of 831.6 and 1,879.7 fish ha^{-1} respectively).
- At Derwent Water, total fish abundance was 132.6 fish ha^{-1} (geometric mean with lower and upper 95% confidence limits of 69.6 and 252.6 fish ha^{-1} respectively).
- At Windermere North Basin, total fish abundance was 163.5 fish ha^{-1} and at Windermere South Basin it was 1741.1 fish ha^{-1}. Confidence limits were unavailable for both basins.

On 27 January 2015, total fish abundance at daytime in Windermere North Basin was 25.9 fish ha^{-1}. The corresponding figure for Windermere South Basin was 35.6 fish ha^{-1}. Confidence limits were unavailable for both basins.

At the following night-time, these figures increased to 178.9 fish ha^{-1} in Windermere North Basin and 334.5 fish ha^{-1} in Windermere South Basin. Confidence limits were unavailable for both basins.

Total fish abundances recorded during daytime on the short transects between eDNA sampling sites in Windermere on 28–30 January 2015 ranged from 0 to 661.1 fish ha^{-1}.

Survey gill netting

In September 2014, the gill netting survey produced a total of 191 individuals at Bassenthwaite Lake, 202 individuals at Derwent Water, 627 individuals at Windermere North Basin and 525 individuals at Windermere South Basin (Table 3.2).

Arctic charr were not recorded in this survey, probably because of the relatively low sampling effort in the context of this rare species. However, they were recorded when more intensive but non-destructive specialised gill netting was used on a spawning ground in Windermere North Basin in late autumn 2014 during the continuation of sampling described by Winfield et al. (2014b).

Additional information on species presence and approximate abundance

The total fish species lists contained 10 species for Bassenthwaite Lake, 9 species for Derwent Water and 16 species for Windermere. These are listed in Table 3.2 together with the approximate abundance score for each species, presented separately for the 2 basins of Windermere.

Table 3.2 Summary of species abundance data from established method survey for the 4 Cumbrian basins

Species	Bassenthwaite Lake		Derwent Water		Windermere North Basin		Windermere South Basin	
Arctic charr					3		6	
Atlantic salmon	9				9		11	
Brown trout	6	(2)	7	(1)	5	(12)	4	(6)
Bullhead					8		10	
Common bream					12		7	
Dace	7	(2)	9					
Eel	4		8		6		5	
Minnow	8		4		7		9	
Perch	1	(78)	1	(132)	1	(595)	1	(477)
Pike	5	(1)	6	(1)	4	(5)	3	(4)
River lamprey					15		15	
Roach	3	(38)	2	(30)	2	(15)	2	(38)
Rudd					14		14	
Ruffe	2	(68)	3	(22)				
Sea lamprey					16		16	
Stone loach					10		12	
Tench					13		8	
Three-spined stickleback					11		13	
Vendace	10	(2)	5	(16)				
Total number of species recorded	**10**	**(7)**	**9**	**(5)**	**16**	**(4)**	**16**	**(4)**

Notes: The relative abundance rank (1 = most abundant) is shown for each basin with the number of individuals caught in a gill netting survey in September 2014 given in brackets.

In addition to the species listed in Table 3.2 that are known to be present as native or introduced populations, a number of other species have been recorded at each lake being used as live bait prior to the local ban on the use of freshwater fish as live or dead bait in 2002 (Winfield and Durie 2004). It is possible that some of these species have subsequently established small populations yet to be detected by survey gill nets or other forms of biological sampling.

At Bassenthwaite Lake, these potential populations are common carp, rainbow trout and rudd, while at Windermere they are crucian carp, dace, grayling and rainbow trout. Furthermore, three-spined sticklebacks have been recorded at many Cumbrian lakes

and, while they have not been caught during routine surveys in Bassenthwaite Lake and Derwent Water, they are likely to be present there.

3.2.5 Demonstrating species presence through eDNA

Both the 12S and CytB assays were assessed for their ability to detect species presence when considered as binary data (presence/absence) at each site across each lake sampling campaign.

For Lake Windermere (Figure 3.5), 88% of the previously recorded species (14/16) were detected using the 12S assay and 75% (12/16) using the CytB assey. The concurrent gill net survey detected 25% (4/16) of the species known to have been recorded in Windermere.

In addition to the previously recorded species, a number of additional species were detected with both assays, including common carp (both 12S and CytB), ruffe, sunbleak, rainbow trout, smelt (12S), flounder and topmouth gudgeon (CytB). It is not known which of these detections are due to:

- genuine yet-to-be-detected species (for example common carp, see Section 3.24)
- the presence of DNA but not the species themselves ('environmental contamination' from, for example, bird faeces or wastewater)
- laboratory or sequencing contamination

Two species that have been recorded but are not present in the sequence data are river and sea lamprey. This is likely to be due to their very low abundance and, because due to their lifecycle, they are unlikely to be present at the time of sampling.

Similar patterns were observed in Derwent Water and Bassenthwaite Lake (Figure 3.4).

In Derwent Water, 88% previously recorded species (8/9) were detected with the 12S assay and 67% (6/9) with the CytB. The 12S assay detected an additional 4 species previously unrecorded – stone loach, three and nine spined sticklebacks, and rudd. Gill net sampling detected 77% (7/9) of species known to be present.

For Bassenthwaite Lake, 90% (9/10) of previously recorded species were detected using the 12S assay and 70% (7/10) with the CytB assay. Additional species not previously recorded included common bream, bullhead, common carp (CytB) and three-spined stickleback (both 12S and CytB). Gill net sampling detected 60% (6/10) of species known to be present.

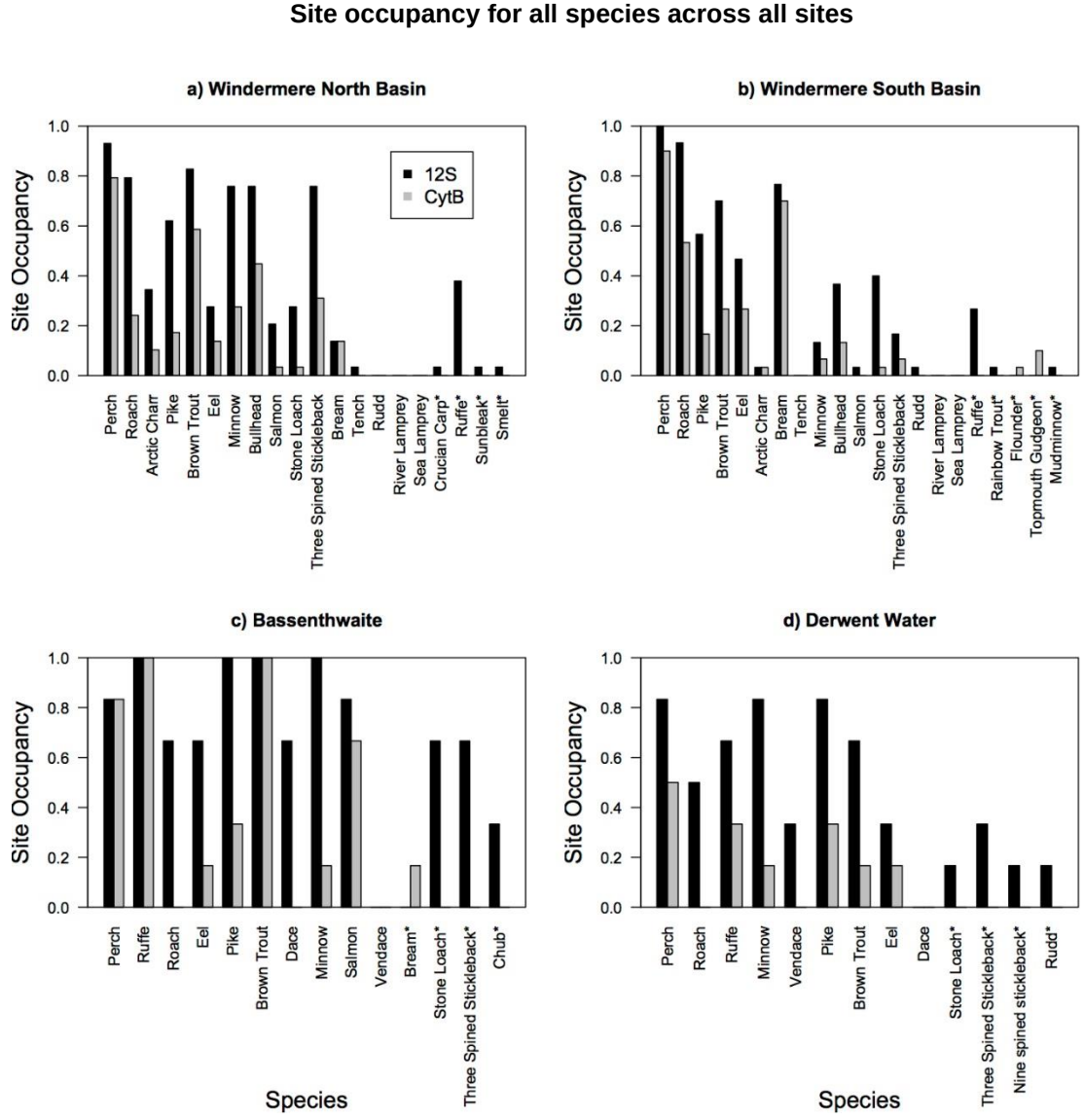


Figure 3.5 Site occupancy for all species across all sites in (a) Windermere North Basin, (b) Windermere South Basin, (c) Bassenthwaite Lake and (d) Derwent Water.

Notes: Species that had not been recorded previously are highlighted with an asterisk.

3.2.6 Estimating required sampling effort in Lake Windermere based on rarefaction

In rarefaction analysis, a good representation of the actual species assemblage has been reached when the graphs begin to plateau; in other words, when an increase in the number of samples does not substantially increase the species richness.

Sample-based rarefaction analyses indicated that between 10 and 20 samples are sufficient to accurately represent the species richness present in Windermere (Figure 3.6). The graphs plateau slightly earlier for the 12S assay (Figure 3.6, black dots) than for the CytB assay (Figure 3.6 grey dots), but this is as expected because more species are detected with the shorter 12S fragment (~100 bp compared with 460 bp for the CytB fragment).

Although this analysis indicates that 10–20 samples may be sufficient for detecting presence/absence, as discussed below more comprehensive sampling is necessary to provide the statistical power required for estimating abundance.

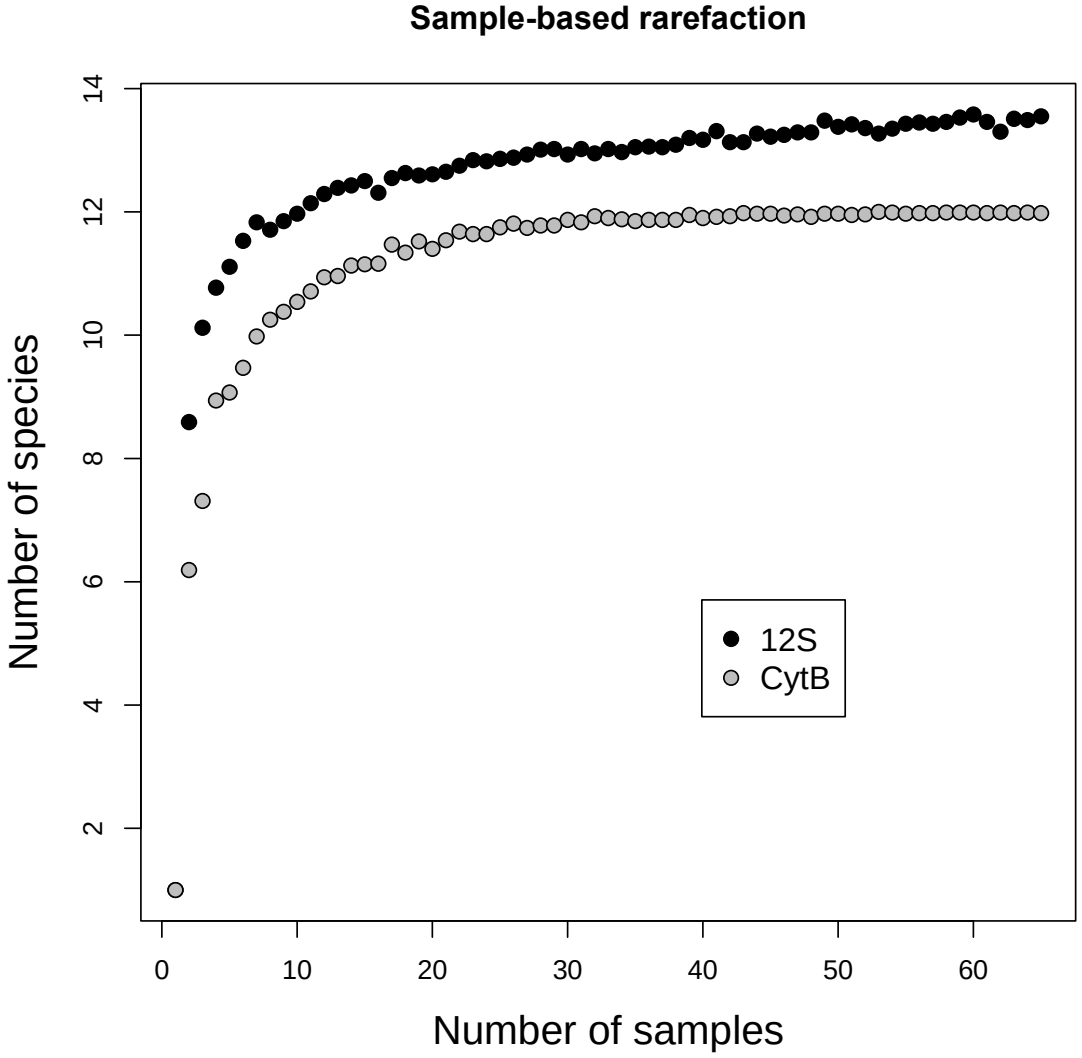


Figure 3.6 Sample-based rarefaction curves for Windermere (North and South basin combined) for 12S (black dots) cytb (grey dots).

3.2.7 Spatial distribution of eDNA records across Lake Windermere

Distribution along the north–south gradient

The presence/absence data based on 12S sequences were used to plot the spatial distribution of each species recorded at more than 2 sites around Windermere (Figure 3.7).

The general pattern emerging from this analysis is that not all species are equally distributed around the lake. Although some species such as perch, roach, pike and trout are recorded ubiquitously across the lake, other species are predominantly found in one of the 2 basins. Arctic charr, minnow and stickleback are common in the North Basin but very rare in the South Basin, whereas common bream and eel appear to prefer the more eutrophic South Basin. These results match with the ecology of the species and with long-term fish survey data.

Spatial distribution of 10 species in Windermere

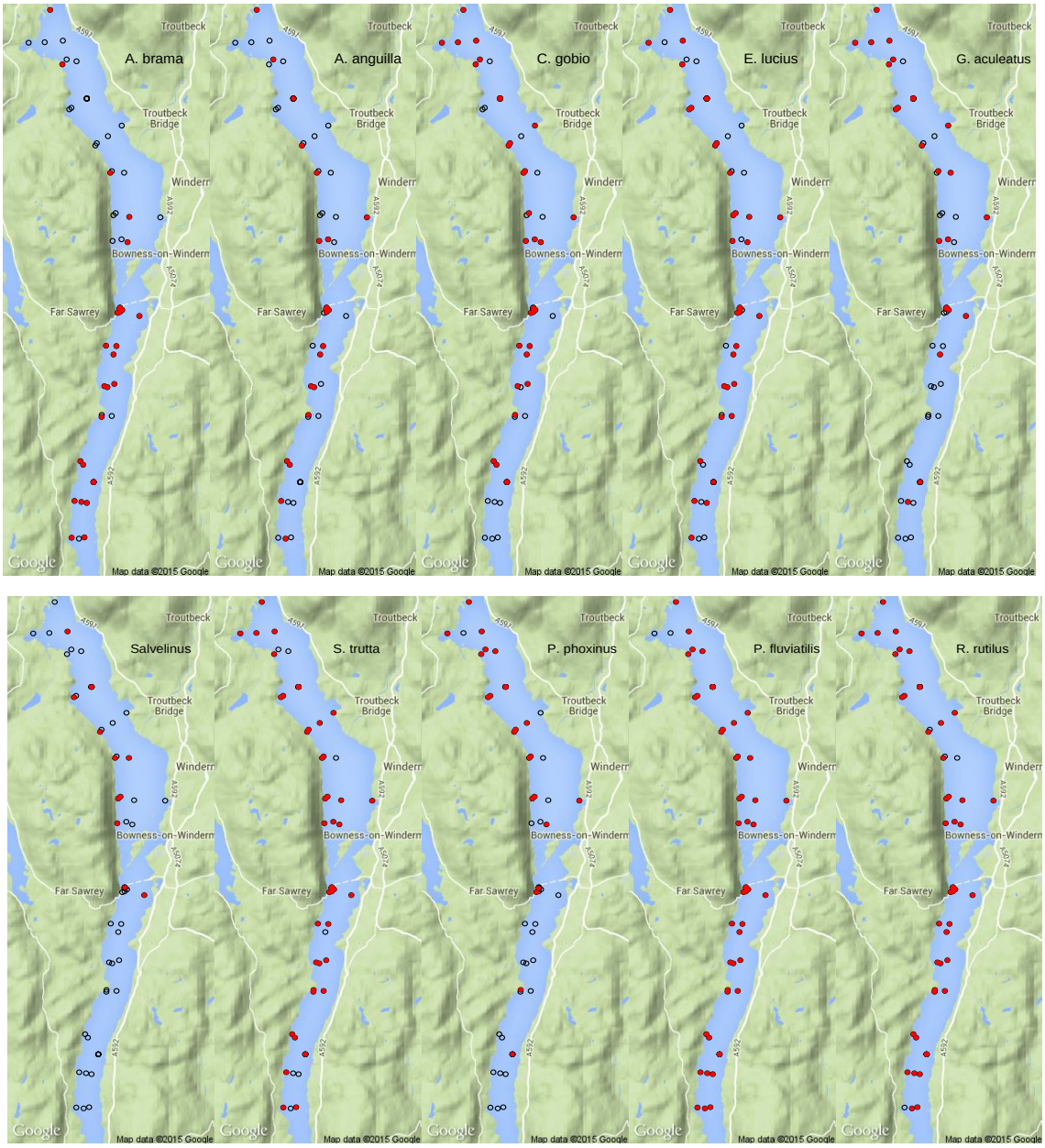


Figure 3.7 Spatial distribution of all 10 species that occurred in more than 2 sites with a 12S site occupancy frequency >0.001

Notes: Common names of species are as follows: *A. brama*: common bream; *A. anguilla*: eel; *C. gobio*: bullhead; *E. lucius*: pike, *G. aculeatus*: three-spined stickleback; *Salvelinus*: charr; *S. trutta*: brown trout; *P. phoxinus*: minnow; *R. rutilus*: roach.
Red dots refer to sites where a species was detected.
Open circles are sites where the species was not detected.

Distribution across transects

The analysis of sequence 12S count data for individual transects (Figure 3.8) revealed that the number of species identified increased from the midline transect towards the 5m transect in the North and South Basins. A total of 13 and 10 species were identified in the 5m transect in North and South Basins respectively compared with 10 and 9 species in the midline transect. Unexpectedly in the South Basin the highest number of species (12) was recorded in the shoreline samples which were collected from a very small geographical area.

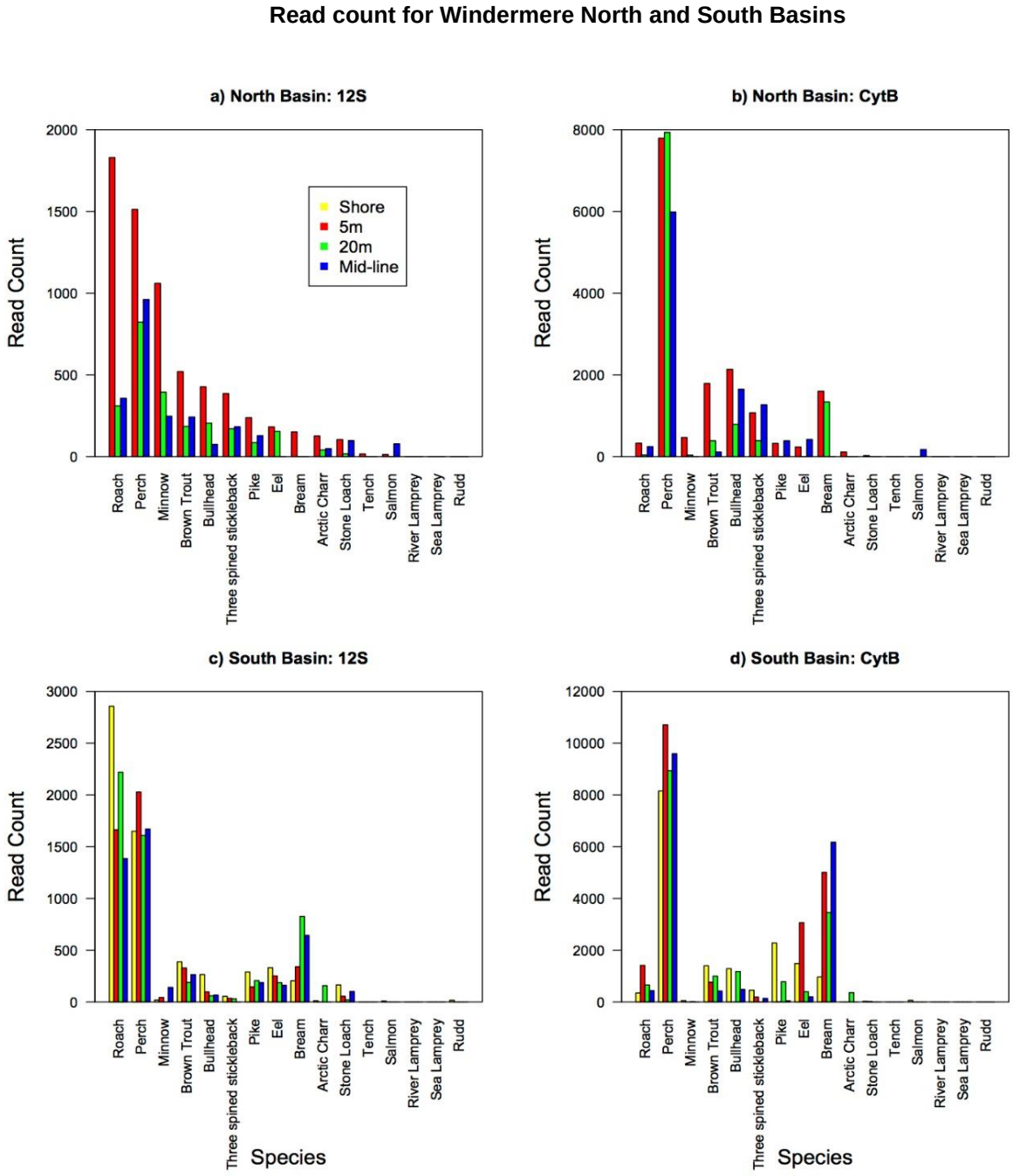


Figure 3.8 Average sequence read counts for all previously recorded species in the 3 different transects and shoreline samples

Vertical distribution in the water column

The distribution of 12S sequence count data along the depth profile (Figure 3.9) showed no marked pattern in the North Basin transect. Most species were found consistently from the surface to the bottom but 2 species, Arctic charr and eel, were only found below a depth ≥30m. In contrast there appears to be an increase of species records from surface to bottom in the South Basin transect.

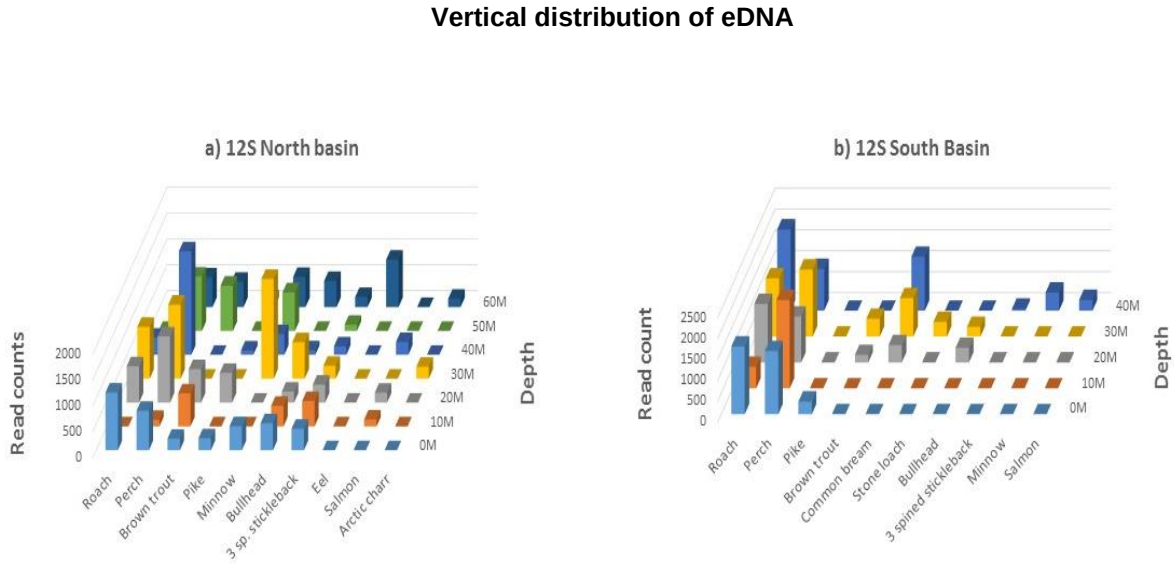


Figure 3.9 12S sequence read counts for different species along a vertical transect (10m sampling interval) in (a) Windermere North Basin and (b) Windermere South Basin

3.2.8 Comparison with data from established surveys and estimating relative abundance from eDNA

Site occupancy of eDNA in sampling sites as proxy for relative abundance

For all 4 basins and for both 12S and cytb sequence data, the proportion of sites in which a species was detected using eDNA (or 'site occupancy') was compared with rank abundance based on long-term data from established surveys and expert opinion (Figure 3.10).

There is a consistent, negative relationship between site occupancy from eDNA and long-term rank (where abundance decreases from 1 to 16). This correlation is highly significant for the Windermere North and South Basins for both loci.

Similar trends were found for Bassenthwaite Lake and Derwent Water, but the correlations are not significant, probably due to a combination of lower statistical power from fewer species and larger confidence intervals around the estimates because they are based on a smaller number of samples.

This estimate of abundance assigns a very high abundance rank to species that are spatially widely distributed compared with species which show a clustered dispersion pattern. This might overestimate the abundance of species such as pike and trout, which are widely dispersed throughout the lake but occur probably at low densities. In addition, the proportion of sites where a species is detected is a naïve estimate of the true abundance as it assumes a detection probability of 1.

Sequence count data as proxy for relative abundance

An alternative way of estimating the relative abundance of individual species is to use the sequence counts of individual species as proxy. This holds true under the assumption that no significant bias is introduced during PCR or sampling. Overall this provides a different quantitative impression of the data compared with presence/absence data. Ultimately, count data might provide a better estimate of biomass than the presence/absence data, but this requires further exploration.

Despite this difference in the expectations for the 2 proxies, results are similar with strong negative correlations between the number of sequence reads and long-term rank (in descending order of abundance) for both 12S and CytB sequence data for the North and South Basins of Windermere. Results are also similar for Derwent Water with 12S, but show non-significant trends for the other combinations (Figure 3.11).

The average number of sequence read counts was also compared against actual numbers sampled in the September 2014 gill net surveys for all 4 basins (Appendix C, Figure S2). There is a consistent trend for the number of sequence reads to increase with abundance, but these correlations are not significant. Again this could be due to low statistical power with only 4–6 species included in the analyses.

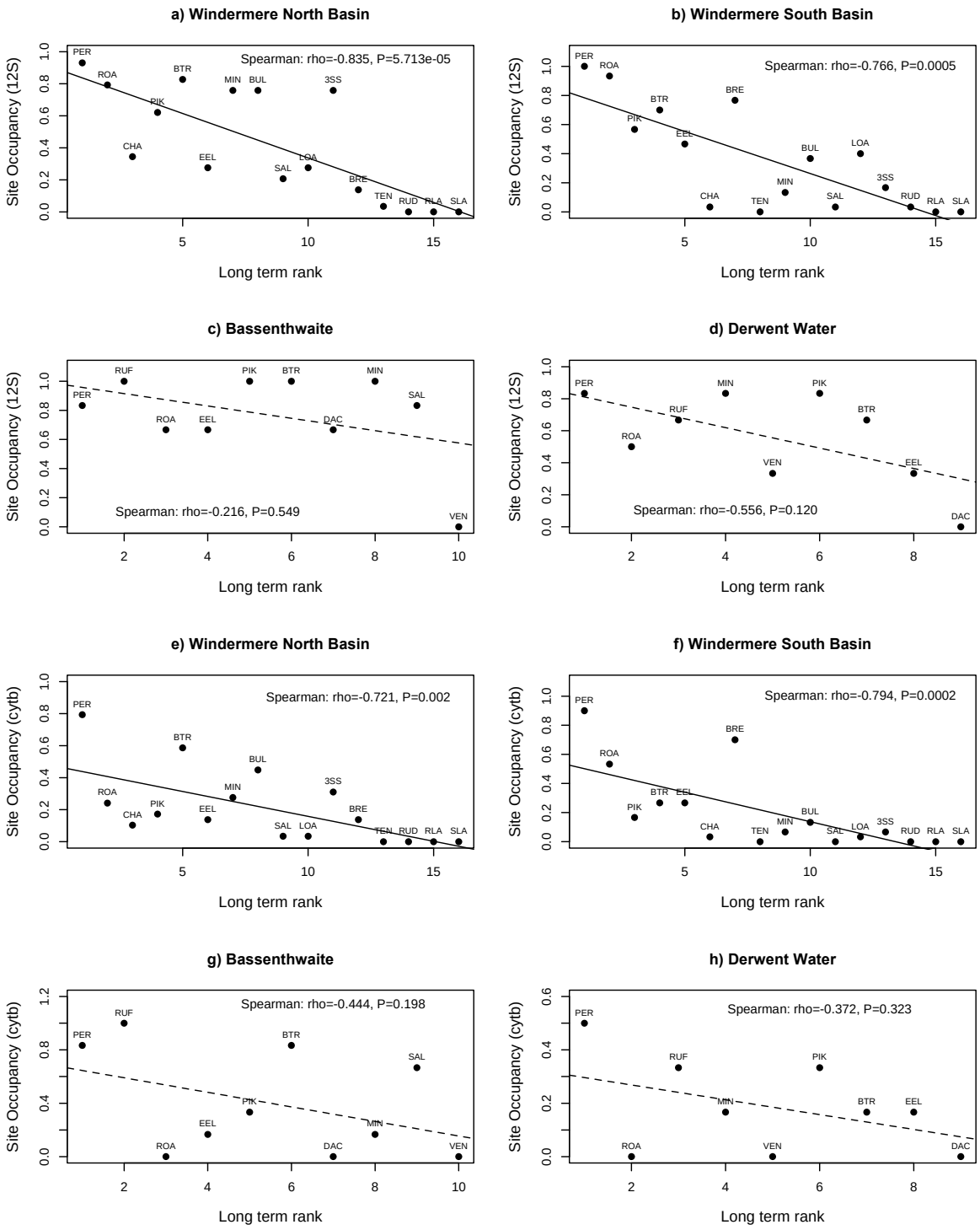


Figure 3.10 Relationship between the site occupancy and long-term rank abundance (from expert opinion) for the 4 basins for 12S (a-d) and CytB (e-h) data.

Notes: Long-term rank (based on data from established surveys and expert opinion) decreases from 1 to 16.
Spearman correlation coefficient and associated significance values are shown.
The three letter codes are species name abbreviations (see List of Abbreviations at the end of report for details.

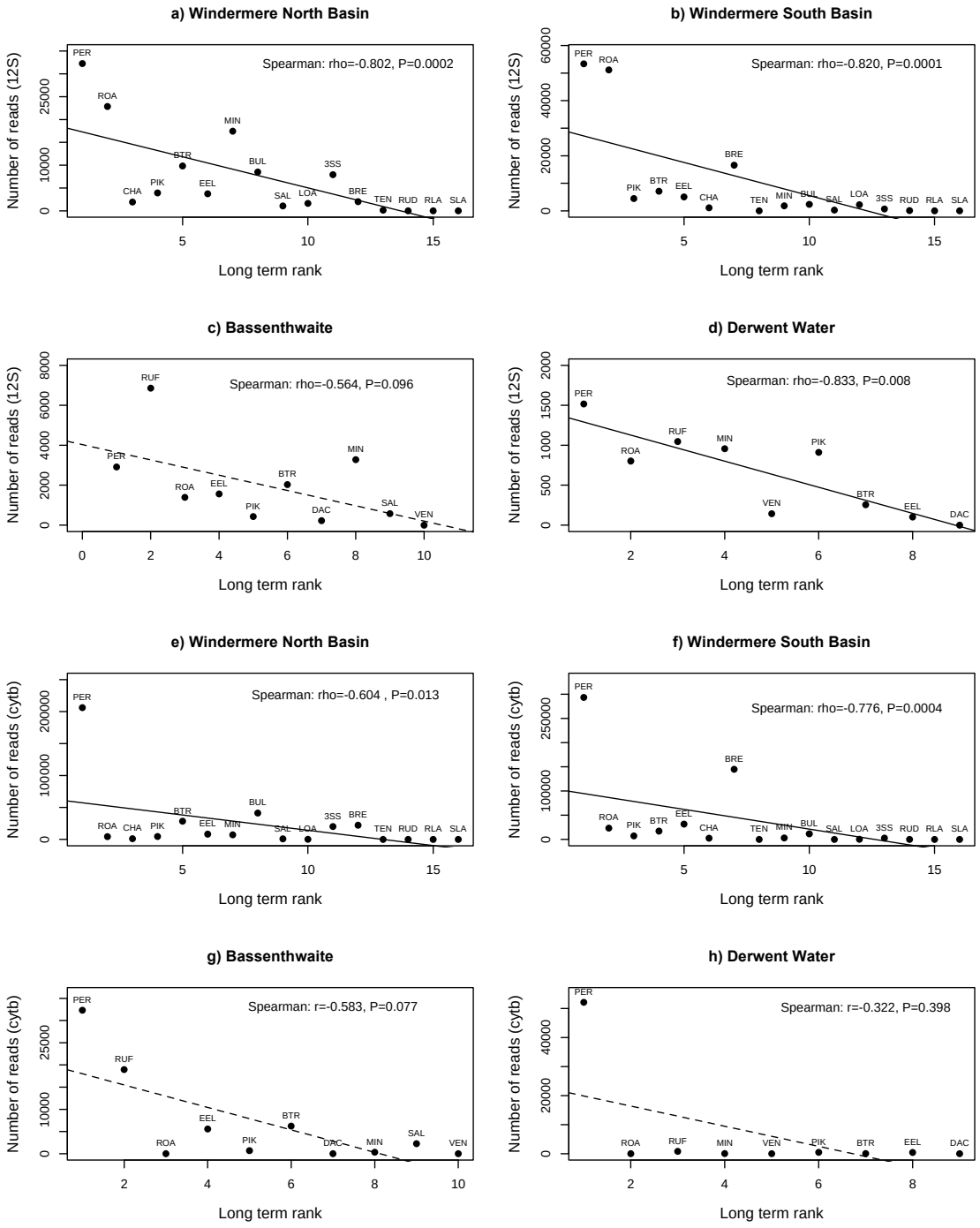


Figure 3.11 Average 12S (a-d) and CytB (e-h) sequence read counts for species previously recorded in each basin

Notes: Species are ordered according to expert abundance rank (decreasing abundance from left to right). Spearman correlation coefficient and associated significance values are shown. 3-letter codes are species name abbreviations (see 'List of Abbreviations' for details).

3.2.9 Difference in species composition between Windermere North and South Basins

Overall the relative proportion of 12S sequence counts for different species across sample sites was significantly different between the Windermere North and South Basins (Figure 3.12).

The proportion of species with a clear association with oligotrophic water bodies such as salmon, brown trout, Arctic charr, minnow and bullhead was much higher in the North Basin (34%) compared with the South Basin (9%). This trend was reversed for species with a clear eutrophic association such as roach, rudd, common bream, tench and eel where the association was 19% in the North Basin and 50% in the South Basin (Figure 3.12).

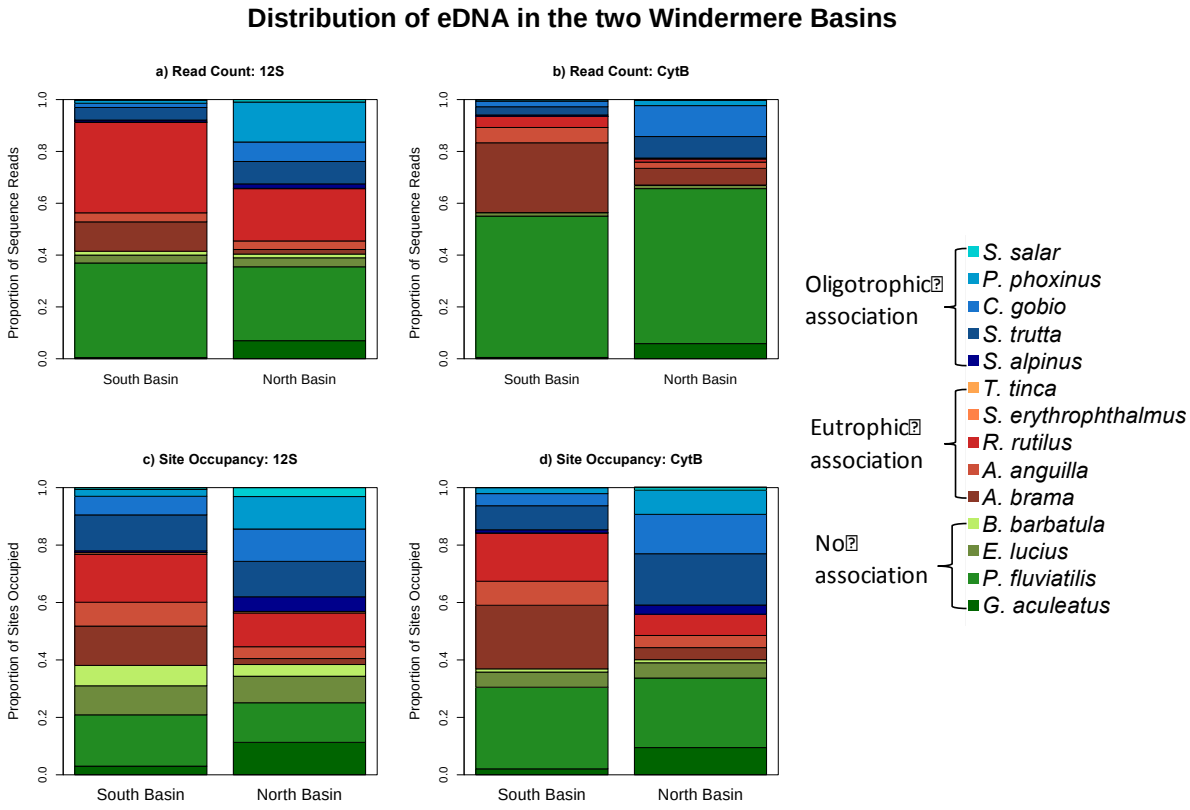


Figure 3.12 Distribution of eDNA based on Read Count (a and b) and Site Occupancy data (c and d) for both North and South Basins of Windermere. Data for both 12S (a and c) and CytB (b and d) are illustrated.

Notes: eDNA from species that have a more oligotrophic association is found at a higher frequency in the North compared to South Basin, whereas eDNA from species with a more eutrophic association is found at a higher frequency in the South compared to North Basin.

3.2.10 Taxonomic assignment of non-target reads

The majority of the 12S and CytB sequences did not significantly match the curated reference database. All of these non-target sequences were subjected to a BLAST search against the complete nucleotide database on GenBank with the NCBI tool blastn.

This revealed a positive identification of a wide range vertebrate species including mammals, birds, amphibians and some marine fish which were not included in our reference database. The list includes a wide range of species typically associated with aquatic habitats such as otter, moorhen, cormorant, various duck and geese species but also many other species potentially occurring in the wider catchment area. These include common farm animals such as cow, sheep and chicken, humans and wild animals such as red deer, red squirrel, fox and tawny owl.

The primers appear to be largely vertebrate specific, except for low-level amplification of bacterial 16S detected in the 12S dataset. No invertebrate sequences were identified. The 12S primers in particular appear potentially suitable not only to monitor fish but also amphibians, water birds and possibly even terrestrial vertebrates occurring in the catchment.

4 Discussion

4.1 Introduction

Although many investigations have examined the potential of eDNA techniques to survey aquatic and other biota (see the extensive review by Rees et al. 2014), very few of these studies have addressed lake fish populations. Moreover, similarly small numbers of these studies have utilised NGS methods to facilitate the DNA barcoding of entire communities through metabarcoding. Consequently, the present study is at the global forefront of such eDNA research through its application of NGS methods to the fish communities of several large UK lakes.

In the context of lake fish assessments to meet the requirements of the EU Water Framework Directive and other international and national environmental legislation, appropriate sampling and assessment protocols still lag behind those developed for most other biota. As reviewed by Kubečka et al. (2009), there is still a significant need for method development with respect to such fish communities, to which eDNA approaches could make a significant contribution.

Following considerable effort devoted to eDNA methodological developments, the 4 basins of Bassenthwaite Lake, Derwent Water, Windermere North Basin and Windermere South Basin proved in early 2015 to be excellent study sites for the present purposes. As reviewed in detail by Maberly et al. (2011), these Cumbrian lakes include a wide range of conditions with respect to nutrient loadings and thus lake trophic status. In turn, these diverse environmental conditions contribute to an equally diverse range of fish communities and abundances as reported in detail by Winfield et al. (2015a, 2015b) for combined gill netting and hydroacoustic surveys carried out in late 2014, with some hydroacoustic observations also made within the present project.

This discussion first considers the methodological developments made during the present study. It then turns to the use of eDNA for the qualitative assessment of the presence/absence of lake fish and the quantitative assessment of the relative abundance of such biota. This is followed by a comparison of eDNA data with those previously generated from the study sites using the established sampling techniques of gill netting and hydroacoustics and their integration, together with other information, in the form of expert opinion. Brief comment is also made on the potential of eDNA to survey non-fish vertebrates associated with large lakes or their immediate catchments, before some closing comments are offered including suggested priorities for future research.

4.2 Methodological developments

This study developed a robust method for eDNA analysis of lake fish which proved suitable for the study and applicable to further scientific investigations. The method incorporates:

- a DNA capture and filtration method yielding high concentrations of environmental DNA
- a molecular assay to reliably amplify DNA from all target fish species and a wider range of vertebrates
- a protocol for Illumina sequencing of PCR amplicons

- a reference database and bioinformatics pipeline for the analysis of sequence reads

However, an exhaustive comparison of all potential methodological avenues was outside the scope of this study and further exploration is required to optimise the efficiency and ease of use of the method for monitoring purposes. The includes:

- development of methods that do not require immediate processing of water samples
- further experiments to better understand the relationship between fish biomass and sequence read counts

4.3 Use of eDNA for assessing the presence/absence of lake fish

Overall, and as discussed in more lake-specific detail below in the context of comparisons with data collected by established techniques, the eDNA approach was remarkably effective in detecting the local presence of fish species when judged against data from established sampling techniques and other sources from Bassenthwaite Lake, Derwent Water and Lake Windermere.

The main species of each lake were represented in the eDNA data from both the 12S and cytb markers. Generally, records were more frequent for the shorter 12S fragment (~100 bp) compared with the relatively long cytb fragment (460 bp). Given the fast degeneration rates of eDNA in water (Barnes et al. 2014), this divergence is probably a result of different degeneration rates. It is expected that a signal from a longer fragment would be lost very quickly, whereas a 100 bp fragment might persist in the environment for days allowing for dispersion across a larger geographical scale. Consequently, the CytB records may indicate that the species was present much closer to where the water sample was taken, while 12S sequences may have originated from some distance away – either within the lake or even up its tributaries. This observation in turn suggests that using a longer fragment may be useful for pinpointing the exact location of species, but that using a shorter fragment might be more useful for simply detecting the presence of a species anywhere in the water body using a limited number of subsamples.

This issue warrants further systematic exploration through experimental approaches and analysis of a wider range of eDNA fragment lengths.

4.4 Use of eDNA for assessing relative abundance of lake fish

This study attempted to assess the relative abundance of individual species by using their sequence counts as a proxy, which is a valid approach assuming that no significant bias is introduced during sampling or subsequent PCR.

At an overall level, using the average 12S sequence count per site across the 4 study basins provided a different quantitative impression of the data compared with a simple presence/absence analysis. The latter assigns a very high abundance rank to species that are spatially widely distributed compared with species that show a more clustered dispersion pattern, even though their actual abundance may be relatively low. As a result, this approach is likely to overestimate the abundance of species such as pike and brown trout which are widely dispersed throughout the lake but naturally occur at low densities.

Ultimately, count data might provide a better estimate of true relative abundance than presence/absence data, but this requires further exploration. Nevertheless, the data also showed that most of the unexpected species occurred only at a few sites and were also represented by low sequence counts. With repeated sampling, the probability of detection can be estimated, allowing for more realistic estimates of the actual species abundance via site occupancy modelling (as discussed in Section 1.1).

4.5 Comparison of results from eDNA and established methods

The results from the eDNA approach were remarkably consistent in terms of species presence and relative abundance for both native and introduced fish species when compared with results and expert opinion obtained from Bassenthwaite Lake, Derwent Water and Lake Windermere using established methods primarily based on long-term gill netting and hydroacoustics (Winfield et al. 2015a, 2015b).

Indeed, the fish species lists produced by one eDNA sampling occasion for each lake more accurately reproduced species lists, previously assembled on the basis of many years of sampling using established techniques, than did a corresponding gill netting survey. For example, in the best studied case of Windermere, the eDNA approach detected a number of species noted in the review of Pickering (2001) but not detected in the gill net survey of 2014 (Winfield et al. 2015b). Similarly, the eDNA results from Bassenthwaite Lake were remarkably concordant with the fish community of that lake as reviewed by Thackeray et al. (2006). Such under-representation of species in gill netting surveys is an acknowledged sampling artefact which has a number of causes including:

- fish morphology (for example, eel are not susceptible to retention in gill nets)
- fine-scale spatial distribution (for example, three-spined stickleback may be limited to the extreme inshore where nets cannot be deployed)
- movement patterns (for example, bullhead may be unlikely to be sampled by gill nets due to their relatively limited movements and very benthic habitat preference)

A significant increase in the level of sampling effort use for the gill nets would be likely to increase the apparent local species list. However, the deployment of this destructive sampling method is strictly limited in the UK and the degree of increased sampling efforts necessary to sample all species present is highly unlikely to be acceptable to the Environment Agency or other stakeholders. The relative inefficiency of gill nets in detecting all the fish species present in a water body under locally permissible levels of sampling effort is a significant complication in their use in any assessment protocols based on species lists. This is particularly true in the UK where relatively few lake fish species occur as a result of the last glaciation (Brucet et al. 2013) and so such effects may be proportionally more significant.

For Bassenthwaite Lake, the only notable discrepancies was the apparent recording of bleak by 12S eDNA and for common bream and bullhead by CytB eDNA, even though neither species has ever been reported from the lake by established sampling techniques. However, all of these species occurred at very low frequencies. Furthermore, none of these records could be confirmed with the other marker and therefore most likely represent low level contamination. The only consistent and robust discrepancy occurs for chub, but although never caught by established sampling or demonstrably shown to be present by anglers' catches, there have been unsubstantiated reports by anglers that the species is present (I.J.W., personal

observation) and it is known to be present in the River Derwent immediately upstream of the lake (Peter McCullough, Environment Agency, personal communication). The converse is true for the extremely rare vendace, but it must be noted that the recent return of this species to the lake from Derwent Water as described by Winfield et al. (2015a) currently only amounts to 3 individuals being recorded in the last 12 years or so.

For Derwent Water, dace failed to be detected by eDNA even though it is known to have been present until at least recently. However, it was not recorded by gill netting in 2014 and in the previous years had only been recorded intermittently and in low numbers. Conversely, rudd was recorded by eDNA but has never been seen in survey gill nets. However, this species is usually strongly associated with macrophytes and so it is possible that it occurs in the lake in just a few localised areas rich in macrophytes which have not been surveyed using gill nets. Finally it is very encouraging that eDNA detected the local presence of vendace despite the fact that, outside its short early winter spawning season, this species is restricted to deep areas which make up less than 100ha of the lake's total surface area.

The more diverse fish community of Windermere also produced very encouraging comparisons, with most of the following discussion appropriate to observations from both its basins. Species recorded by eDNA but not by established surveys included bleak, common carp, mudminnow, pumpkinseed, sunbleak, topmouth gudgeon, smelt and ruffe. The actual presence of bleak, mudminnow, pumpkinseed, topmouth gudgeon and sunbleak seems almost impossible. Again all of these species were recorded at only 1 or 2 sites at very low frequencies and only for a single marker, and are therefore most likely explained by low level contamination.

An exception is the record of ruffe, for which 12S sequences were present in 25% of the sites in the South Basin and 38% of the sites in the North Basin. Based on this high frequency it is less likely that this represents contamination, although the species was not recorded with the CytB data. It is known from anglers' catches that the upstream lake of Rydal Water contains introduced crucian carp, roach and ruffe, and possibly other species that could potentially colonise Windermere by movements along a relatively short length of connecting river. Alternatively, it is possible that only their eDNA makes such a journey, although if so, it is remarkable that this persists along much of the length of the lake. However, the records of 12S eDNA from a wide range of terrestrial vertebrates suggest that this is not inconceivable. Given the recent arrival of ruffe in a number of Cumbrian lakes as reviewed by Winfield et al. (2010), it is also possible that the species has arrived in Windermere but currently remains at very low abundance below the limits of detection of the long-term survey gill netting programme.

Although smelt has never been recorded in Windermere, this nationally scarce species is known to be present in the lower reaches of a river in a nearby catchment. Nevertheless, its otherwise undetected presence in Windermere seems unlikely, not least because of substantial barriers to the migration of this small species. However, smelt is also a very popular dead bait used by pike anglers and pike anglers were active during the sampling as they are most of the time at Windermere. It therefore seems that dead baiting is a likely source of smelt eDNA in the lake.

The Windermere results are also encouraging in that, as for the vendace in Derwent Water, the eDNA sampling recorded the known presence of the scarce Arctic charr in both basins. Moreover, the eDNA results were also consistent with the known significant difference in Arctic charr abundance in the lake's 2 basins, with it being slightly more abundant in the North Basin. These positive detections of this nationally scarce species are also remarkable given that, although unknown to the authors at the time of eDNA sampling, Windermere's recreational Arctic charr fishery had its worst season on record in 2014 (Winfield et al. 2015a) with the clear implication that population numbers are similarly currently very low.

In summary, the concordance of the eDNA results with those obtained from gill netting and hydroacoustics gives considerable promise that this new approach may be able to replace gill netting as a source of some of the biological data necessary for the assessment of lake fish communities. As such, it may be possible to incorporate it into community metrics for assessments for Water Framework Directive purposes akin to those developed by Argillier et al. (2013) and Kelly et al. (2012) for European and Irish waters, respectively, and into more species-specific assessments for Habitats Directive purposes such as those reported for Arctic charr and whitefish (*Coregonus lavaretus*) by Winfield et al. (2009) and Winfield et al. (2013), respectively.

The ability of eDNA to assess absolute, rather than relative, fish abundance has not yet been demonstrated in lakes. However, such measures can be obtained independently by non-destructive hydroacoustic techniques which are themselves rapidly becoming deployed and standardised across European lakes (Hateley et al. 2013).

4.6 Use of eDNA to survey non-fish vertebrates

Although the remit of the present project was focused on lake fish communities, its findings also offer some insights into the feasibility of using eDNA techniques for the wider assessment of non-fish vertebrates associated with lakes and their immediate catchments.

The majority of all 12S and CytB sequences did not match the extensive and effectively comprehensive reference database developed for freshwater fish taxa for this project. The obvious inference is that such sequences originate from non-fish taxa and so they were opportunistically analysed using a simple BLAST search with the NCBI tool blastn. With such investment of very little extra effort, this analysis produced positive identifications for a wide range of vertebrate species including mammals, birds, amphibians and some marine fish species (known to be used in the lakes as dead bait by anglers) which were not included in the reference database. The list included a wide range of species typically directly associated with aquatic habitats such as otter (*Lutra lutra*), moorhen (*Gallinula chloropus*), cormorant (*Phalacrocorax carbo*) and various duck and geese species. In addition, the list also included many other vertebrate species potentially occurring in the wider catchment area but less evident actually on or in the lakes. These included common domesticated farm animals such as cow (*Bos taurus*), sheep (*Ovis aries*) and chicken (*Gallus gallus domesticus*), together with wild vertebrates such as red deer (*Cervus elaphus*), red squirrel (*Sciurus vulgaris*), red fox (*Vulpes vulpes*) and tawny owl (*Strix aluco*). Moreover, the primers used appear to be vertebrate-specific since no invertebrate sequences were identified, although many such species were undoubtedly present.

Consequently, the eDNA approach employed in this study may have further applications in the qualitative but extensive high-level survey of non-fish vertebrate taxa occurring in the catchment.

4.7 Closing remarks

The present investigation was driven primarily by the need to develop reliable and cost-effective lake fish assessments to meet the requirements of the Water Framework Directive and other international and national environmental legislation. It is universally agreed that there is no single sampling method that can produce all of the kinds of information needed to make such assessments. Even the use of a combination of methods from the range of established techniques still presents an incomplete picture with varying degrees of bias and incomplete coverage (Kubečka et al. 2009).

The findings of the present study indicate that eDNA approaches can make a very significant contribution to this challenging task. The results obtained were extremely consistent with the understanding of the fish communities of the sampled lakes based on long-term monitoring using established techniques – primarily gill netting and hydroacoustics – augmented by other data sources such as anglers' catches. Moreover, this work moved beyond a simple presence/absence analysis to produce indications of the relative abundance of species, which were again consistent with earlier ecological and assessment interpretations. However, the present study was limited in its field component with sampling being restricted to just 4 lake basins during the winter. Moreover, all 3 lakes are essentially taken from a relatively small sub-set of UK lakes, that is, relatively large and essentially (with some variation) mesotrophic water bodies.

There is a clear need to explore the feasibility of using less exacting and less strenuous field sampling than used in the present study while still producing reliable results. Given the intensive and meticulous sampling used on Lake Windermere, the effects of reducing sampling effort and changing its nature, including further investigations of sampling from the shore, should be explored. The use of single-use samplers should also be included in such work which, if successful, could open up possibilities of Citizen Science programmes like the ones now being used for surveys of great crested newt (*Triturus cristatus*) following the development of appropriate methodology based on the pioneering scientific understanding reviewed by Rees et al. (2014).

In addition to the above areas of technique research, there is also a pressing need to develop and demonstrate a much wider applicability of the eDNA approach to a much greater range of types of standing waters and fish communities. In particular, expansion to cover many more oligotrophic and more eutrophic waters is highly desirable.

In terms of an initial expansion to lakes for which significant data from established techniques are already available, there is a potential for immediate expansion to around 8 more lakes in Cumbria, 7 lochs in Scotland and 2 llyns in Wales on the basis of gill netting and hydroacoustics data accessible to, or held by, CEH, Natural Resources Wales and the Environment Agency.

Expansion to further lakes is likely to have to take place in the absence of local data from established scientific sampling techniques. However, advantage could be taken of the ability of anglers' catches and background knowledge to produce similar although less quantitative and less objective forms of corroborating information. This approach would also work well at intensively managed recreational fishery waters in both rural and urban environments.

References

ARGILLIER, C., CAUSSÉ, S., GEVREY, M., PÉDRON, S., DE BORTOLI, J., BRUCET, S., EMMRICH, M., JEPPESEN, E., LAURIDSEN, T., MEHNER, T., OLIN, M., RASK, M., VOLTA, P., WINFIELD, I.J., KELLY, F., KRAUSE, T., PALM, A. AND HOLMGREN, K., 2013. Development of a fish-based index to assess the eutrophication status of European lakes. *Hydrobiologica*, 704 (1), 193-211.

BARNES, M.A., TURNER, C.R., JERDE, C.L., RENSHAW, M.A., CHADDERTON, W.L. AND LODGE, D.M., 2014. Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science & Technology*, 48 (3), 1819-1827.

BIGGS, J., EWALD, N., VALENTINI, A., GABORIAUD, C., DEJEAN, T., GRIFFITHS, R.A., FOSTER, J., WILKINSON, J.W., ARNELL, A., BROTHERTON, P., WILLIAMS, P. AND DUNN, F., 2015. Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus*). *Biological Conservation*, 183, 19-28.

BIK, H.M., PORAZINSKA, D.L., CREER, S., CAPORASO, J.G., KNIGHT, R. AND THOMAS, W.K., 2012. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, 27 (4), 233-243.

BOLGER, A.M., LOHSE, M. AND USADEL, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30 (15), 2114-2120.

BRUCET, S., PÉDRON, S., MEHNER, T., LAURIDSEN, T,L., ARGILLIER, C., WINFIELD, I.J., VOLTA, P., EMMRICH, M., HESTHAGEN, T., HOLMGREN, K., BENEJAM, L., KELLY, F., KRAUSE, T., PALM, A., RASK, M. AND JEPPESEN, E., 2013. Fish diversity in European lakes: geographical factors dominate over anthropogenic pressures. *Freshwater Biology*, 58 (9), 1779-1793.

CREER, S., FONSECA, V.G., PORAZINSKA, D.L., GIBLIN-DAVIS, R.M., SUNG, W., POWER, D.M., PACKER, M., CARVALHO, G.R., BLAXTER, M.L., LAMBSHEAD, P.J.D. AND THOMAS, W.K., 2010. Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology*, 19 (Suppl. 1), 4-20.

FICETOLA, G.F., MIAUD, C., POMPANON, F. AND TABERLET, P., 2008. Species detection using environmental DNA from water samples. *Biology Letters*, 4 (4), 423-425.

FICETOLA, G., COISSAC, E., ZUNDEL, S., RIAZ, T., SHEHZAD, W., BESSIERE, J., TABERLET, P. AND POMPANON, F., 2010. An *in silico* approach for the evaluation of DNA barcodes. *BMC Genomics*, 11 (1), 434.

FICETOLA, G.F., PANSU, J., BONIN, A., COISSAC, E., GIGUET-COVEX, C., DE BARBA, M., GIELLY, L., LOPES, C.M., BOYER, F., POMPANON, F., RAYÉ, G. AND TABERLET, P., 2015. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, 15 (3), 543-556.

FOLMER, O., BLACK, M., HOEH, W., LUTZ, R. AND VRIJENHOECK, R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoen invertebrates. *Molecular Marine Biology and Biotechnology*, 3 (5), 294-299.

FONSECA, V.G., CARVALHO, G.R., SUNG, W., JOHNSON, H.F., POWER, D.M., NEILL, S.P., PACKER, M., BLAXTER, M.L., LAMBSHEAD, P.J.D., THOMAS, W.K. AND CREER, S., 2010. Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, 1, article number 98.

GOTELLI, N.J. AND COLWELL, R.K., 2010. Estimating species richness. In: *Biological Diversity: Frontiers in Measurement and Assessment* (ed. A.E. Magurran and B.J. McGill), Chapter 4, pp. 39-54. Oxford: Oxford University Press.

HATELEY J., CLABBURN P., DRASTIKC V., GODLEWSKA M., GUILLARD J., KUBECKA J., MORRISSEY E., THACKERAY S.J. AND WINFIELDG I.J., 2013. Standardisation of hydroacoustic techniques for fish in freshwaters. In *Proceedings of the First Underwater Acoustics Conference and Exhibition* (Corfu, 23-26 June 2013), edited by J.S. Papadakis and L. Bjørnø, pp. 1595-1600. Heraklion, Crete: Institute of Applied and Computational Mathematics.

HUSON, D.H., AUCH, A.F., QI, J. AND SCHUSTER, S.C., 2007. MEGAN analysis of metagenomic data. *Genome Research*, 17 (3), 377-386.

IVANOVA, N.V., ZEMLAK, T.S., HANNER, R.H. AND HEBERT, P.D.N., 2007. Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes*, 7 (4), 544-548.

JI, Y., ASHTON, L., PEDLEY, S.M., EDWARDS, D.P., TANG, Y., NAKAMURA, A., KITCHING, R., DOLMAN, P.M., WOODCOCK, P., EDWARDS, F.A., LARSEN, T.H., HSU, W.W., BENEDICK, S., HAMER, K.C., WILCOVE, D.S., BRUCE, C., WANG, X., LEVI, T., LOTT, M., EMERSON, B.C. AND YU, D.W., 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16 (10), 1245-1257.

KATOH, K. AND STANDLEY, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution,* 30 (4), 772-780.

KELLY, F.L., HARRISON, A.J., ALLEN, M.S, CONNOR, L. AND ROSELL, R.S., 2012. Development and application of an ecological classification tool for fish in lakes in Ireland. *Ecological Indicators*, 18, 608-619.

KELLY, R.P., PORT, J.A., YAMAHARA, K.M. AND CROWDER, L.B., 2014. Using environmental DNA to census marine fishes in a large mesocosm. *PloS One*, 9 (1): e86175.

KOCHER, T., MEYER, A., EDWARDS, S., PAABO, S., VILLABLANCA, F. AND WILSON, A., 1989. Dynamics of mitochondrial-DNA evolution in animals – amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences of the United States of America*, 86 (16), 6196-6200.

KUBEČKA, J., HOHAUSOVÁ, E., MATĚNA, J., PETERKA, J., AMARASINGHE, U.S., BONAR, S.A., HATELEY, J., HICKLEY, P., SUURONEN, P., TERESCHENKO, V., WELCOMME, R. AND WINFIELD, I.J., 2009. The true picture of a lake or reservoir fish stock: a review of needs and progress. *Fisheries Research*, 96 (1), 1-5.

LAWSON HANDLEY, L., 2015. How will the 'molecular revolution' contribute to biological recording? *Biological Journal of the Linnean Society*, 115 (3), 750-766.

LI, W. AND GODZIK, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22 (13), 1658-1659.

MABERLY, S.C., DE VILLE, M.M., THACKERAY, S.J., FEUCHTMAYR, H., FLETCHER, J.M., JAMES, J.B., KELLY, J.L., VINCENT, C.D., WINFIELD, I.J., NEWTON, A., ATKINSON, D., CROFT, A., DREW, H., SAAG, M., TAYLOR, S. AND TITTERINGTON, H., 2011. A survey of the lakes of the English Lake District: the Lakes Tour 2010. NERC/Centre for Ecology and Hydrology report to Environment Agency, North West Region and Lake District National Park Authority. LA/NEC04357/1 [unpublished].

MACKENZIE, D.I., NICHOLS, J.D., LACHMAN, G.B., DROEGE, S., ANDREW ROYLE, J. AND LANGTIMM, C.A., 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83 (8), 2248-2255.

MACKENZIE, D.I. AND NICHOLS, J.D., 2004. Occupancy as a surrogate for abundance estimation. *Animal Biodiversity and Conservation*, 27 (1), 461-467.

MAGOČ, T. AND SALZBERG, S.L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27 (21), 2957-2963.

MAHON, A., NATHAN, L. AND JERDE, C., 2014. Meta-genomic surveillance of invasive species in the bait trade. *Conservation Genetics Resources*, 6 (3), 563-567.

OKSANEN, J., GUILLAUME BLANCHET, F., KINDT, R., LEGENDRE, P., MINCHIN, P., O'HARA, R., SIMPSON, G., SOLYMOS, P., HENRY, M., STEVENS, H. AND WAGNER, H., 2015. *Vegan: Community Ecology Package. R package version 2.2-1* [online]. Available from: http://CRAN.R-project.org/package=vegan [accessed 27 October 2015].

PICKERING, A.D., 2001. *Windermere: restoring the health of England's largest lake.* Ambleside, Cumbria: Freshwater Biological Association.

PILLIOD, D.S., GOLDBERG, C.S., ARKLE, R.S. AND 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 (8), 1123-1130.

REES, H.C., MADDISON, B.C., MIDDLEDITCH, D.J., PATMORE, J.R.M. AND 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.

RIAZ, T., SHEHZAD, W., VIARI, A., POMPANON, F., TABERLET, P. AND COISSAC, E., 2011. ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, 39 (21).

SAYERS, E., 2008. E-utilities quick start. In: Entrez Programming Utilities Help [online]. Bethesda (MD): National Center for Biotechnology Information. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25500/ [accessed 27 October 2015].

SCHMIDT, B.R., KÉRY, M., URSENBACHER, S., HYMAN, O.J. AND 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 (7), 646-653.

STAMATAKIS, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22 (21), 2688-2690.

SUYAMA, M., TORRENTS, D. AND BORK, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34 (Web Server issue), W609-W612.

SZITENBERG, A., JOHN, M., BLAXTER, M. AND LUNT, D., 2015. ReproPhylo: an environment for reproducible phylogenomics. *PLoS Computational Biology*, 11 (9), e1004447.

THACKERAY, S., MABERLY, S. AND WINFIELD, I., 2006. The ecology of Bassenthwaite Lake (English Lake District). Ambleside, Cumbria: Freshwater Biological Association.

THOMSEN, P.F., KIELGAST, J., IVERSEN, L.L., MOLLER, P.R., RASMUSSEN, M. AND WILLERSLEV, E., 2012a. Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLOS One*, 7 (8), e41732.

THOMSEN, P.F., KIELGAST, J., IVERSEN, L.L., WIUF, C., RASMUSSEN, M., GILBERT, M.T.P., ORLANDO, L. AND WILLERSLEV, E., 2012b. Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, 21 (11), 2565-2573.

WINFIELD, I.J. AND DURIE, N.C., 2004. Fish introductions and their management in the English Lake District. *Fisheries Management and Ecology,* 11 (3-4), 195-201.

WINFIELD, I.J., CRAGG-HINE, D., FLETCHER, J.M. AND CUBBY, P.R., 1996. The conservation ecology of *Coregonus albula* and *C. lavaretus* in England and Wales, U.K. In: *Conservation of Endangered Freshwater Fish in Europe* (ed. A. Kirchhofer and D. Hefti), pp. 213-223. Basel: Birhauser Verlag.

WINFIELD, I.J., FLETCHER, J.M. AND JAMES, J.B. AND 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 (2009), 30-38.

WINFIELD, I.J., FLETCHER, J.M. AND JAMES, J.B., 2010. An overview of fish species introductions to the English Lake District, UK, an area of outstanding conservation and fisheries importance. *Journal of Applied Ichthyology*, 26 (suppl. S2), 60-65.

WINFIELD, I.J., BEAN, C.W., GORST, J., GOWANS, A.R.D., ROBINSON, M. AND THOMAS, R., 2013. Assessment and conservation of whitefish (*Coregonus lavaretus* (L.)) in the U.K. Advances in Limnology, 64 (2103), 305-321.

WINFIELD, I.J., FLETCHER, J.M. AND JAMES, J.B., 2014a. *Monitoring the fish populations of Bassenthwaite Lake and Derwent Water, 2013*. NERC/Centre for Ecology and Hydrology report to Environment Agency, North West Region. LA/NEC05043/1 [unpublished].

WINFIELD, I.J., FLETCHER, J.M. AND JAMES, J.B., 2014b. Monitoring the fish populations of Windermere, 2013. NERC/Centre for Ecology and Hydrology report to Environment Agency, North West Region. LA/NEC05043/2 [unpublished].

WINFIELD, I.J., FLETCHER, J.M. AND JAMES, J.B., 2015a. *Monitoring the fish populations of Bassenthwaite Lake and Derwent Water, 2014*. NERC/Centre for Ecology and Hydrology report to United Utilities. LA/NEC05364/1 [unpublished].

WINFIELD, I.J., FLETCHER, J.M. AND JAMES, J.B., 2015b. *Monitoring the fish populations of Windermere, 2014*. NERC/Centre for Ecology and Hydrology report to United Utilities. LA/NEC05364/2 [unpublished].

YU, D.W., JI, Y.Q., EMERSON, B.C., WANG, X.Y., YE, C.X., YANG, C.Y. AND DING, Z.L., 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3 (4), 613-623.

ZHANG, Z., SCHWARTZ, S., WAGNER, L. AND MILLER, W., 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, 7 (1-2), 203-214.

List of abbreviations

12S	mitochondrial 12S ribosomal RNA gene
BLAST	Basic Local Assignment Search Tool
bp	base pair
CEH	Centre for Ecology and Hydrology
COI	mitochondrial cytochrome oxidase 1 gene
cytb	mitochondrial cytochrome b gene
ddH_2O	double-distilled water
DNA	deoxyribonucleic acid
eDNA	environmental DNA
FBA	Freshwater Biological Association
MEGAN	MEtaGenome ANalyzer [software]
ML	maximum likelihood
NCBI	National Center for Biotechnology Information [USA]
NGS	next generation sequencing
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction

List of abbreviations used in figures for species names

Scientific name	Common name	Abbreviation
Abramis brama	Common bream	BRE
Anguilla anguilla	Eel	EEL
Barbatula barbatula	Stone loach	LOA
Coregonus albula	Vendace	VEN
Cottus gobio	Bullhead	BUL
Cyprinus carpio	Common carp	CAR
Esox lucius	Pike	PIK
Gasterosteus aculeatus	Three-spined stickleback	3SS
Gymnocephalus cernua	Ruffe	RUF
Lampetra fluviatilis	River lamprey	RLA
Lepomis gibbosus	Pumkinseed	PUM
Leucaspius delineatus	Sunbleak	SUN
Leuciscus leuciscus	Dace	DAC
Oncorhynchus mykiss	Rainbow trout	RTR
Osmerus eperlanus	Smelt	SME
Perca fluviatilis	Perch	PER
Petromyzon marinus	Sea lamprey	SLA
Phoxinus phoxinus	Minnow	MIN
Platichthys flesus	Flounder	FLO
Pseudorasbora parva	Topmouth gudgeon	TMG
Pungitius pungitius	Nine-spined stickleback	9SS
Rutilus rutilus	Roach	ROA
Salmo salar	Salmon	SAL
Salmo trutta	Brown trout	BTR
Salvelinus alpinus	Arctic charr	CHA
Scardinius erythrophthalmus	Rudd	RUD
Squalius cephalus	Chub	CHU
Tinca tinca	Tench	TEN
Umbra pygmaea	Mudminnow	MUD

Glossary

Bioinformatics	Field of biology that uses computer science, statistics, mathematics and engineering to study and process biological data.
Bioinformatics pipeline	Steps involved in extracting, processing and analysing raw data generated, for example, by next generation sequencing.
BLAST®	Basic Local Assignment Search Tool – bioinformatics tool that finds regions of local similarity between DNA or protein sequences (http://blast.ncbi.nlm.nih.gov/Blast.cgi)
DNA barcoding	Identification of a species or taxon based on PCR amplification and sequencing of a standard region of DNA (often the mitochondrial cytochrome oxidase 1 gene).
GenBank®	Annotated collection of publicly available DNA sequences housed at the National Center for Biotechnology Information (USA) (www.ncbi.nlm.nih.gov/genbank/)
Hydroacoustics	Use of sonar technology for the detection and monitoring of underwater characteristics or species assemblages.
Illumina sequencing	Next generation sequencing on an platform developed by the company Illumina, such as a MiSeq (www.illumina.com/systems/miseq.html) used in the current study.
MEGAN	MEtaGenome Analyzer – a computer program for analysis of large metabarcoding datasets.
Metabarcoding	A rapid method of biodiversity assessment that combines 2 technologies: • DNA based taxon identification (DNA barcoding) • high-throughput DNA sequencing (NGS) It uses universal PCR primers to mass-amplify DNA barcodes from mass collections of organisms or from environmental DNA.
Minibarcode	DNA barcode region designed to be shorter than the standard barcodes so as to amplify degraded DNA.
Mitochondrial 12S and cytochrome b genes	Regions of mitochondrial DNA (mtDNA). 12S is a ribosomal RNA (rRNA) gene, whereas cytochrome b is a protein coding gene involved in the process of oxidative phosphorylation, in which oxygen and sugars are used to create energy in the form of adenosine triphosphate (ATP). Both gene regions have been used widely for species identification and phylogenetic placement.
Next generation DNA sequencing (NGS)	Also known as high-throughput sequencing, 'next generation sequencing' is the catch-all term used to describe a number of different modern sequencing technologies, including Illumina (Solexa). These recent technologies allow sequence DNA to much quicker and cheaper than the previously used Sanger

	sequencing and as such have revolutionised the study of genomics and molecular biology.
Polymerase chain reaction (PCR)	A method of amplifying the number of copies of a target region of DNA using oligonucleotide primers which permits downstream analysis such as DNA sequencing.
Primer	A short single-stranded stretch of DNA that is complementary to the DNA sequence of a target region. A pair of primers, flanking the target region, is required for PCR amplification. The primers bind to the target DNA during PCR and prime the addition of nucleotides, generating millions of copies of the target sequence.
Rarefaction	A technique used to estimate species richness for a given number of samples based on the construction of rarefaction curves, which plot the number of species as a function of the number of samples. If the curve reaches a plateau, it indicates that the samples accurately reflect the diversity present and more intensive sampling is unlikely to yield additional species
Ruttner sampler	Standard water sampler that is closed at a certain sampling depth by a falling weight.
Site occupancy modelling (SOM)	A statistical modelling approach for estimating the abundance of a species based on the proportion of sites in which the species is detected when the probability of detection at a site is <1. Implementation of the method requires comprehensive spatial and repeated sampling. It is widely used in ecological studies and has recently been advocated for estimating abundance from eDNA (for example, Pilliod et al. 2013, Schmidt et al. 2013).

Appendix A: List of target species and associated information

Scientific Name	Common Name	Previously recorded in study lakes	Species number in positive controls	12S sequenced during current project
Abramis brama	Common bream	x	5	yes
Acipenser sturio	Common sturgeon			
Alburnoides bipunctatus	Schneider			
Alburnus alburnus	Bleak		20	yes
Alosa alosa	Allis shad			
Alosa fallax	Twaite shad			
Ambloplites rupestris	Rock bass			
Ameiurus melas	Black bullhead			
Ameiurus nebulosus	Brown bullhead		17	yes
Anguilla anguilla	European eel	x		
Aspius aspius	Asp			
Barbatula barbatula	Stone loach	x		yes
Barbus barbus	Barbel		21	yes
Blicca bjoerkna (=Abramis bjorkna)				
Carassius auratus	Goldfish		18	
Carassius carassius	Crucian carp			
Chondrostoma nasus	Nase			
Cobitis taenia	Spined loach			
Coregonus albula	Vendace	x	4	yes
Coregonus autumnalis	Pollan			yes
Coregonus lavaretus	Whitefish			
Coregonus oxyrinchus	Houting			
Cottus gobio	Bullhead	x	23	yes
Ctenopharyngodon idella	Grass carp			
Cyprinus carpio	Common carp	x	10	yes
Esox lucius	Pike	x	1	yes
Gasterosteus aculeatus	Three-spined stickleback	x		
Gobio gobio	Gudgeon		19	yes

Scientific Name	Common Name	Previously recorded in study lakes	Species number in positive controls	12S sequenced during current project
Gymnocephalus cernuea	Ruffe	x	2	yes
Hypophthalmichthys molitrix	Silver carp			
Hypophthalmichthys nobilis	Bighead carp			
Lampetra fluviatilis	River lamprey	x		
Lampetra planeri	Brook lamprey			
Lepomis gibbosus	Pumpkinseed		11	yes
Leucaspius deliniatus	Sunbleak		12	
Leuciscus idus	Orfe			yes
Leuciscus leuciscus	Dace	x	22	yes
Lota lota	Burbot			
Micropterus salmoides	Largemouth bass			
Misgurnus fossilis	Weather loach			
Neogobius kessleri	Bigheadgoby			
Neogobius melanostomus	Round goby			
Oncorhynchus gorbuscha	Pink salmon			
Oncorhynchus mykiss	Rainbow trout			
Osmerus eperlanus	Smelt			
Perca fluviatilis	Perch	x	3	yes
Petromyzon marinus	Sea lamprey	x		
Phoxinus phoxinus	Minnow	x	8	yes
Pimephales promelas	Fathead minnow			
Platichthys flesus	Flounder			yes
Proterorhinus semilunaris	Western tubenose goby			
Pseudorasbora parva	Topmouth gudgeon		13	yes
Pungitius pungitius	Nine-spined stickleback		14	yes
Rhodeus sericeus	Bitterling			
Rutilus rutilus	Roach	x	6	
Salmo salar	Atlantic salmon	x		
Salmo trutta	Trout	x	7	yes
Salvelinus alpinus	Arctic charr	x		
Salvelinus fontinalis	Brook charr			
Sander lucioperca	Pikeperch (zander)			
Scardinius erythrophthalmus	Rudd	x	15	yes

Scientific Name	Common Name	Previously recorded in study lakes	Species number in positive controls	12S sequenced during current project
Siluris glanis	Wels catfish			
Squalius cephalus (=Leuciscus cephalus)	Chub	x		yes
Thymallus thymallus	Grayling			
Tinca tinca	Tench	x	9	yes
Umbra pygmaea	Mudminnow		16	
Vimba vimba	Vimba bream			

Appendix B: Sequences of primers tested in WP 2 and 4

Primer	Sequence 5′–3′	Reference
12S_F	ACACCGCCCGTCACTCT	SPYGEN
12S_R	CTTCCGGTACACTTACCRTG	SPYGEN
12S_510F	ACTGGGATTAGATACCCC	Kelly et al. 2014
12S_655R	TAGAACAGGCTCCTCTAG	Kelly et al. 2014
Fish2bCBR	GATGGCGTAGGCAAACAAGA	Thompson et al. 2012a
Fish2CBL	ACAACTTCACCCCTGCAAAC	Thompson et al. 2012a
Fish2degCBL	ACAACTTCACCCCTGCRAAY	Thompson et al. 2012a
Fish2CBR	GATGGCGTAGGCAAATAGGA	Thompson et al. 2012a
CytB_14735F	AAAAACCACCGTTGTTATTCAACTA	Kocher et al. 1989
CytB_15149R	GCDCCTCARAATGAYATTTGTCCTCA	Kocher et al. 1989
CytB_F1	CACATCTGCCGAGAYGT	this study
CytB_F2	AGAAACCTGAAAYATTGG	this study
COI_F1	GGTGCCTGAGCCGGAATAGT	this study
COI_F2	TGCCTGAGCCGGAATAGT	this study
CytB_R1	GTTTCAGGTTTCTTTGTA	this study
CytB_R2	CCRATGTTTCAGGTTTCT	this study
CytB_R3	GATATTTGTCCTCATGGAAG	this study
CytB_R4	TATTTGTCCTCATGGAAG	this study
COI_R1	GAAAATTATTACRAAGGC	this study
COI_R2	ATTATTACRAAGGCGTGGGC	this study
COI_R1	GAAAATTATTACRAAGGC	this study
COI_R2	ATTATTACRAAGGCGTGGGC	this study

Appendix C: Supplementary figures

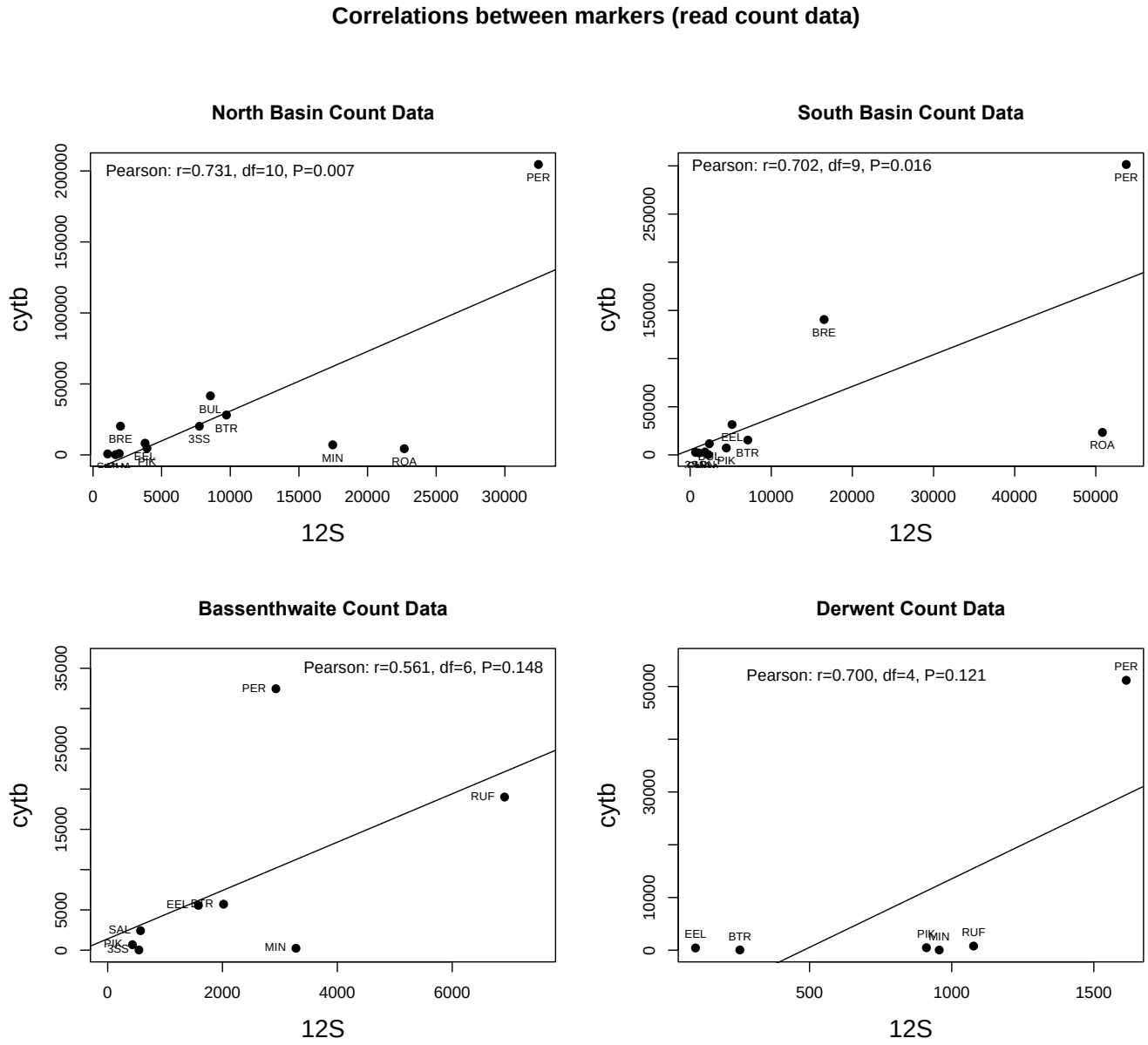


Figure S1 Correlations between 12S and cytb for read count data (RC, that is, number of sequence reads per species)

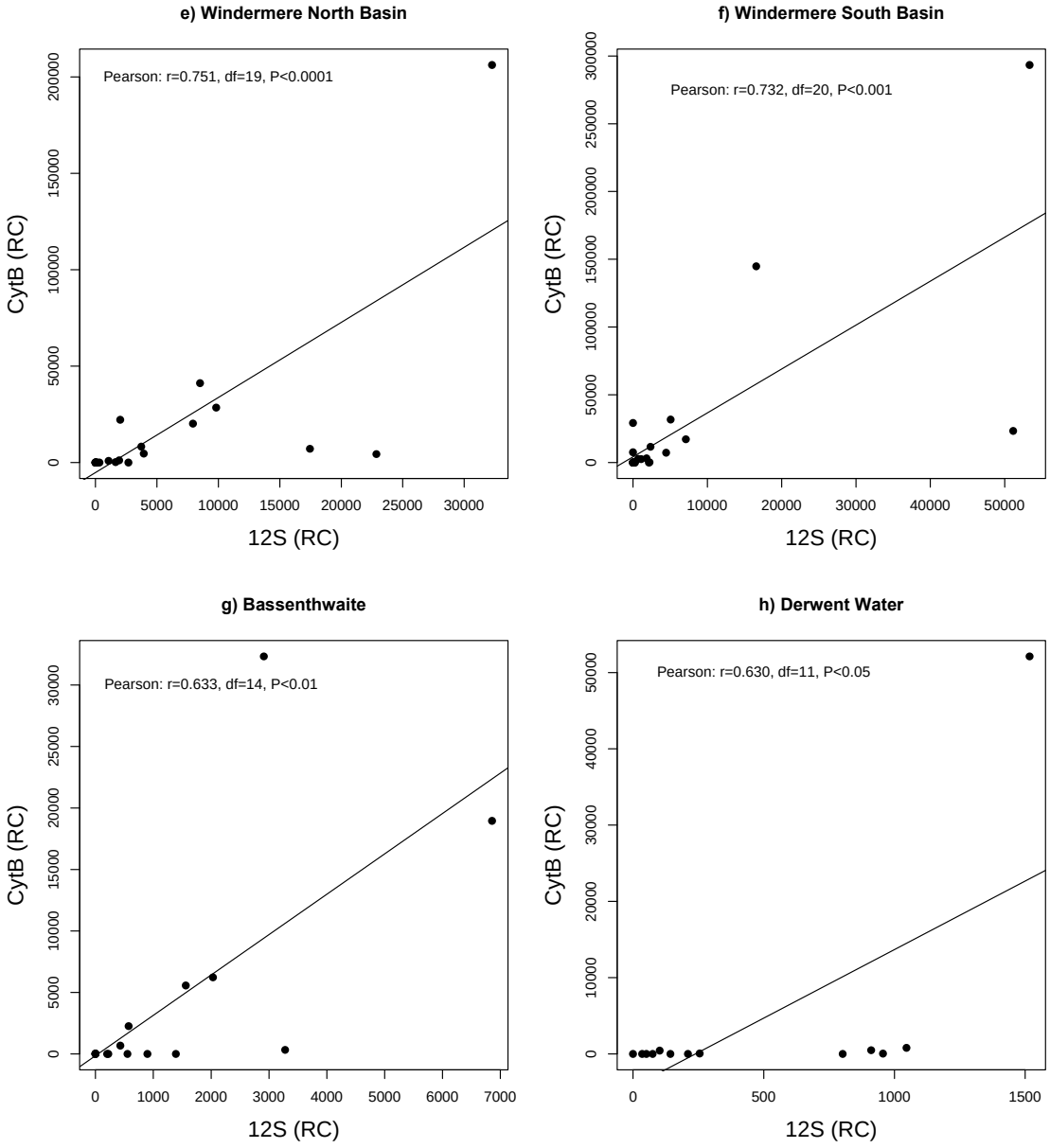


Figure S1 (continued) Correlations between 12S and cytb for read count data (RC, that is, number of sequence reads per species)

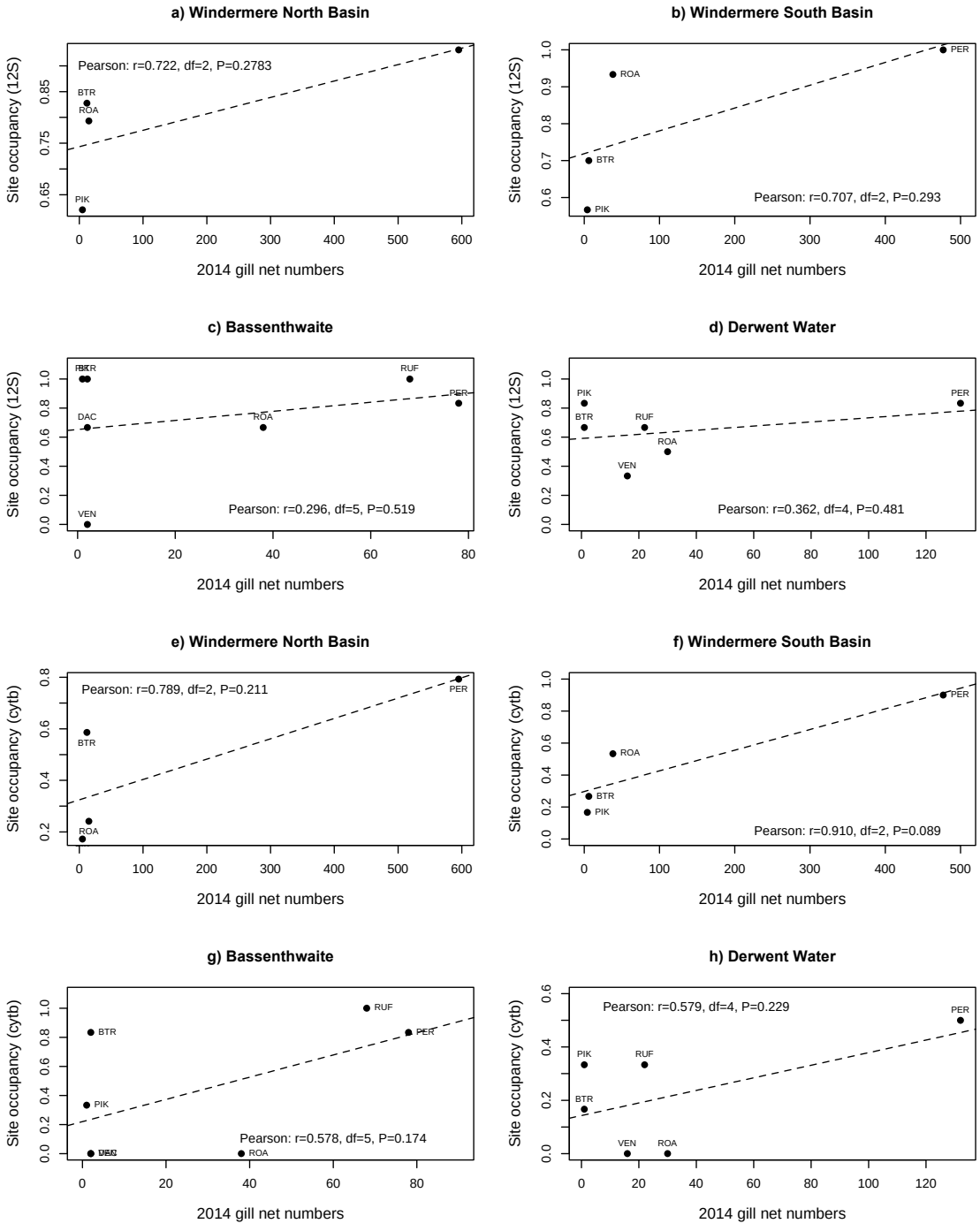
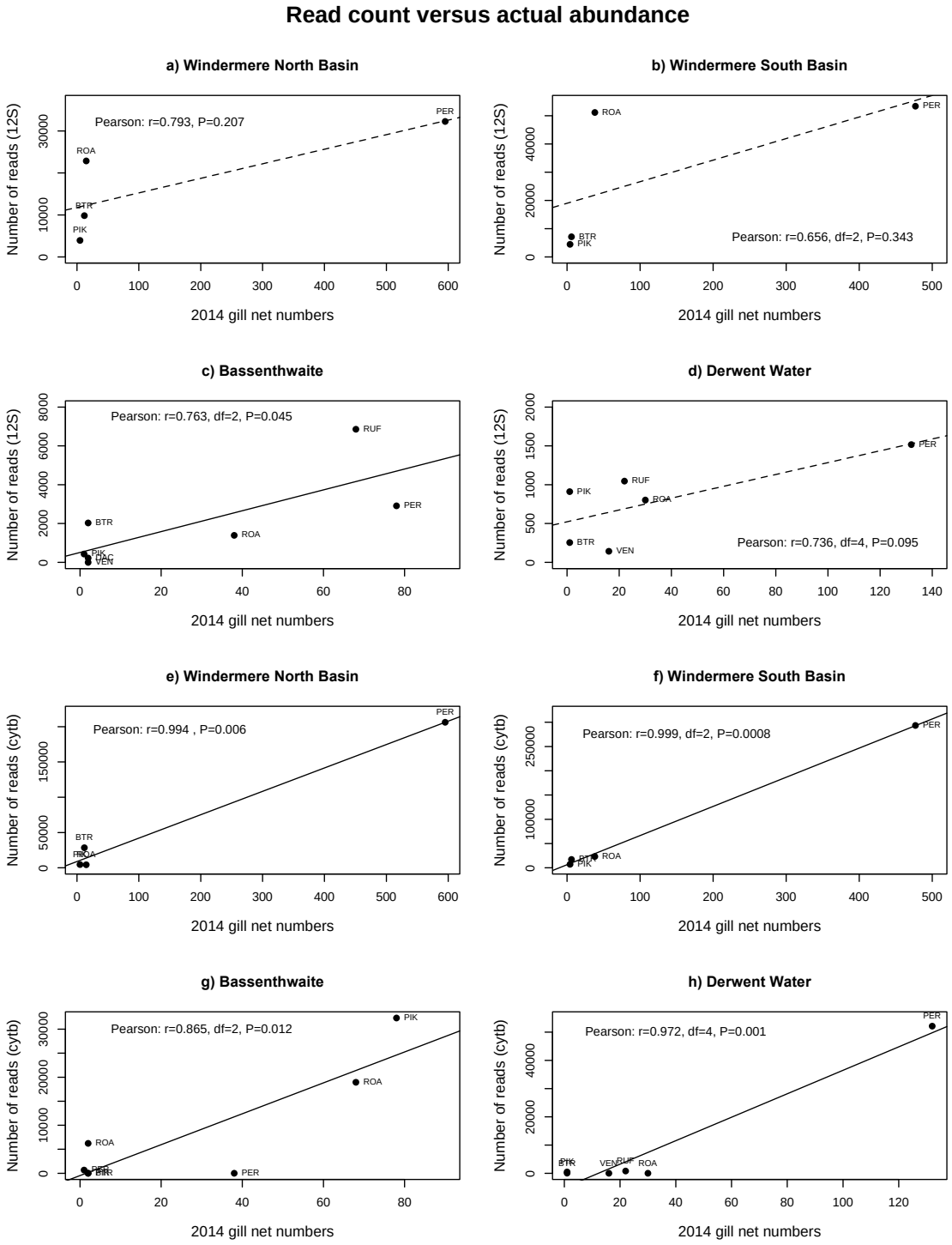


Figure S2. Correlations between eDNA site occupancy (proportion of site where a species is present, y axis) and actual abundance (numbers of each species) from 2014 gill net surveys for 12S (a-d) and CytB (e-h) in four different basins Windermere North Basin (a, e), Windermere South Basin (b, f), Bassenthwaite (c, g) and Derwent Water (d, h)



Notes: There is a trend for increasing number of sequence reads with increasing rank, but in the majority of cases, insufficient data points for correlations to be significant.

Figure S3 Correlations between number of sequence reads (count data, y axis) and actual abundance (numbers of each species) from 2014 gill net surveys for for 12S (a-d) and CytB (e-h) in four different basins; Windermere North Basin (a, e), Windermere South Basin (b, f), Bassenthwaite (c, g) and Derwent Water (d, h).

Notes: There is a trend for increasing number of sequence reads with increasing rank, but in the majority of cases, insufficient data points for correlations to be significant.

Appendix D: Complete list of retained reference sequences for 12S and CytB

See Excel spreadsheet (reference_dbs_summary)

Appendix E: Read counts for the 12S dataset

See Excel spreadsheet (12S_counts_summary)

Appendix F: Read counts for the CytB dataset

See Excel spreadsheet (CytB_counts_summary)

Would you like to find out more about us or about your environment?

Then call us on
03708 506 506 (Monday to Friday, 8am to 6pm)

email
enquiries@environment-agency.gov.uk

or visit our website
www.gov.uk/environment-agency

incident hotline 0800 807060 (24 hours)
floodline 0345 988 1188 / 0845 988 1188 (24 hours)

Find out about call charges: www.gov.uk/call-charges

Environment first: Are you viewing this on screen? Please consider the environment and only print if absolutely necessary. If you are reading a paper copy, please don't forget to reuse and recycle if possible.