Article (refereed) - postprint

Hänfling, Bernd; Lawson Handley, Lori; Read, Daniel S.; Hahn, Christoph; Li, Jianlong; Nichols, Paul; Blackman, Rosetta C.; Oliver, Anna; Winfield, Ian J. 2016. **Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods**. *Molecular Ecology*, 25 (13). 3101-3119. 10.1111/mec.13660

Received date: 15-Dec-2015

Revised date: 14-Mar-2016

Accepted date: 29-Mar-2016
Article type: Original Article

Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods

Bernd Hänfling[1*], Lori Lawson Handley[1*], Daniel S. Read[2], Christoph Hahn[1], Jianlong Li[1], Paul Nichols[1], Rosetta C. Blackman[1], Anna Oliver[2] and Ian J. Winfield[3].

[1]Evolutionary and Environmental Genomics Group (@EvoHull), School of Biological, Biomedical and Environmental Sciences, University of Hull (**UoH**), Cottingham Road, Hull, HU6 7RX

[2] Centre for Ecology & Hydrology (**CEH**), Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10 8BB

[3] Lake Ecosystems Group, Centre for Ecology & Hydrology (**CEH**), Lancaster Environment Centre, Library Avenue, Bailrigg, Lancaster, LA1 4AP

*Joint first authors

Corresponding author: Bernd Hänfling

Email: b.haenfling@hull.ac.uk

Keywords: eDNA, environmental DNA, metabarcoding, fish monitoring, lakes, lentic systems, EC Water Framework Directive.

Running title: eDNA metabarcoding of lake fish

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/mec.13660

Abstract

Organisms continuously release DNA into their environments via shed cells, excreta, gametes and decaying material. Analysis of this "environmental DNA" (eDNA) is revolutionising biodiversity monitoring. eDNA outperforms many established survey methods for targeted detection of single species, but few studies have investigated how well eDNA reflects whole communities of organisms in natural environments. We investigated whether eDNA can recover accurate qualitative and quantitative information about fish communities in large lakes, by comparison to the most comprehensive long-term gill-net dataset available in the UK. Seventy eight 2L water samples were collected along depth profile transects, gill-net sites and from the shoreline in three large, deep lakes (Windermere, Bassenthwaite Lake and Derwent Water) in the English Lake District. Water samples were assayed by eDNA metabarcoding of the mitochondrial 12S and cytochrome *b* regions. Fourteen of the 16 species historically recorded in Windermere were detected using eDNA, compared to four species in the most recent gill-net survey, demonstrating eDNA is extremely sensitive for detecting species. A key question for biodiversity monitoring is whether eDNA can accurately estimate abundance. To test this, we used the number of sequence reads per species and the proportion of sampling sites in which a species was detected with eDNA (i.e. site occupancy) as proxies for abundance. eDNA abundance data consistently correlated with rank abundance estimates from established surveys. These results demonstrate that eDNA metabarcoding can describe fish communities in large lakes, both qualitatively and quantitatively, and has great potential as a complementary tool to established monitoring methods.

INTRODUCTION

Rapid monitoring of changes in biodiversity in response to climate change or other anthropogenic pressures is imperative, but the time and resources required to generate the necessary data are a major constraint in conservation management and ecological research. This is particularly relevant in large lake ecosystems, where for a number of taxa, established methods currently struggle to deliver the required data to fulfil legislative obligations such as the EC Water Framework (European_Communities 2000) and corresponding legislation elsewhere in the word. This difficulty is particularly marked for fish, for which all

established sampling methods have various forms of bias (e.g. (Kubečka *et al.* 2009) and for which biological sampling is typically laborious and destructive (e.g. (Argillier *et al.* 2013). Arguably the biggest recent development in biodiversity monitoring is analysis of environmental DNA (eDNA), which refers to DNA released by organisms into their environment for example in the form of shed cells, excreta or decaying matter. eDNA has great potential for biodiversity monitoring since it is non-invasive, can detect rare or elusive species that are difficult to detect using established methods, and can distinguish cryptic species or juvenile stages that are difficult to identify taxonomically (as reviewed in (Bohmann *et al.* 2014; Lawson Handley 2015; Rees *et al.* 2015). Aquatic environments are particularly suited to eDNA analysis as DNA disperses rapidly in the water column and is more homogeneously distributed than in soil or other sediments.

The application of eDNA has so far largely focused on targeted detection of one or a few species using standard or quantitative Polymerase Chain Reaction (qPCR). Such targeted eDNA assays have proven highly successful for detecting individual species from a wide range of taxonomic groups in aquatic environments (see Table 1 in (Lawson Handley 2015) for a summary). For example, a recent eDNA study targeting great crested newts, *Triturus cristatus,* demonstrated high repeatability and substantially higher detection rates for eDNA compared to established survey methods (Biggs *et al.* 2015). The characterisation of entire communities is not feasible using such species-specific approaches due to the complexity of most ecosystems. An alternative approach is to simultaneously screen whole communities of organisms using eDNA metabarcoding. Here, community DNA is PCR-amplified using broad range primers, and sequenced on a High Throughput Sequencing (HTS) platform (reviewed by Lawson Handley 2015). Direct metabarcoding of homogenized community samples is revolutionising our understanding of the diversity of microscopic eukaryotes (Bik *et al.* 2012) in environments that are notoriously difficult to study, such as soil (Creer *et al.* 2010), and the deep sea (Fonseca *et al.* 2010). Metabarcoding of macrobial eDNA is still in its infancy, but the field is moving forward at a fast pace. The first studies focussed on describing fish communities in tanks or aquaria (Evans *et al.* 2015; Kelly *et al.* 2014; Mahon *et al.* 2014; Miya *et al.* 2015) or on a small scale in natural settings (Thomsen *et al.* 2012a; Thomsen *et al.* 2012b). Recent refinements of the method, including more rigorous testing in aquaria (Miya *et al.* 2015) and in marine (Miya *et al.* 2015; Valentini *et al.* 2015), and freshwater habitats (Valentini *et al.* 2015) have confirmed the method is extremely sensitive

for detecting rare species, and describing presence/absence. Important questions remain though about the efficacy of eDNA metabarcoding for obtaining accurate estimates of species abundance and biomass. 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, PCR), sequencing and bioinformatics stages (Ficetola *et al.* 2015; Yu *et al.* 2012). In metabarcoding studies, in principle, the number of sequences per taxon (or "operational taxonomic unit") could be taken as an estimator of species biomass, but unfortunately in practice, this relationship is not a simple one. For example, (Kelly *et al.* 2014) demonstrated a perfect correlation between rank abundance of eDNA sequences representing four fish genera and rank biomass in a large aquarium, but the actual number of sequence reads was not correlated to biomass. Similarly, Evans *et al.* (2015) found only a modest positive relationship between the number of sequence reads and abundance of eight fish and one amphibian species in mesocosm experiments. A second approach that may be more promising for estimating abundance is to carry out comprehensive spatial and temporal sampling of a given environment and calculate the proportion of sites in which a species is detected with eDNA. Such "site occupancy" data is often collected in ecological studies and can be used as a proxy for abundance (MacKenzie & Nichols 2004; MacKenzie *et al.* 2002). Recent studies indicate this approach could be very promising for analysing eDNA data from both targeted assays (Hunter *et al.* 2015; Pilliod *et al.* 2013; Schmidt *et al.* 2013), and metabarcoding data (Valentini *et al.* 2015).

How well eDNA metabarcoding performs compared to established survey methods for generating both qualitative (presence/absence) and quantitative (abundance/biomass) data remains a key question in the development of the technology for biodiversity monitoring. Here, we addressed this question by comparing eDNA metabarcoding data to the most comprehensive long-term data available for lake fish populations in the UK. We carried out rigorous spatial sampling in three large, deep lakes (Windermere, Bassenthwaite Lake and Derwent Water) in the English Lake District, which are the best-studied lakes in the UK in terms of their fish fauna. Firstly, we developed a workflow for lake fish eDNA metabarcoding, which included building an appropriate reference database of mitochondrial 12S and cytochrome *b* (CytB) genes, testing primer combinations, and developing pipelines for eDNA analyses from sampling to bioinformatics. Second, we carried out water sampling

along depth-profile transects, at gill-net survey sites and at shoreline locations within the lakes. Finally we compared the qualitative and quantitative results from eDNA metabarcoding with long-term and recent gill-net survey datasets to investigate the performance of eDNA against established methods.

MATERIAL AND METHODS

Sampling

Sampling was carried out in three natural lakes (Bassenthwaite Lake, Derwent Water and Windermere) in the English Lake District, UK, that have been intensively studied in terms of their fish populations, physio-chemical and other biological properties for many years (Maberly *et al.* 2011, Fig. 1). Fish populations in these three lakes have been monitored since the early 1990s (Bassenthwaite Lake and Derwent Water, e.g. (Winfield *et al.* 2012a; Winfield *et al.* 2015b) or early1940s (Windermere, e.g. (Winfield *et al.* 2008a; Winfield *et al.* 2015b). This monitoring has been performed using gill netting, trapping, hydroacoustics or analysis of recreational anglers' catches and constitutes the best long-term lake fish datasets in the UK. Windermere, England's largest natural lake (surface area 1480 ha, maximum depth 64 m), is composed of two distinct basins with different physical, chemical and ecological characteristics (North Basin: surface area of 810 ha, maximum depth 64 m, mesotrophic; South Basin: surface area 670 ha, maximum depth 44 m, eutrophic). Bassenthwaite Lake (surface area 528 ha, maximum depth 19 m, eutrophic) and Derwent Water (surface area 535 ha, maximum depth 22 m, mesotrophic) are also among the largest lakes in England and are linked by the River Derwent.

In total 30 offshore samples were collected from each of the two Windermere basins. Additionally, six samples were collected opportunistically from a small area of the shoreline at the Northern end of the South Basin. Water samples were collected from Windermere during 28th – 30th January 2015. Most offshore samples were collected along three transects with approximately 1 km sampling interval between sites. Transects 1, 2 and 3 run along the 5m, 20m depth contour and the lake midline respectively (Fig. 1). The sampling depth for transect 1, 2 and 3 was 2 m, 10 m and 20 m respectively. This sampling scheme covered 7 of the 10 sites that are used for annual gill net surveys (Winfield *et al.* 2015b). Water samples

were also collected at the 3 remaining gill net sites (Fig. 1). At the deepest point along the midline transect in both North (approximate depth 64 m) and South Basin (approximate depth 44 m) a depth profile was collected. The North Basin depth transect was collected at 0-10-20-30-40-50-60 m depth and the South Basin depth transect was collected at 0-10-20-30-40 m. (Fig. 1). Water samples were also collected at 5 gill net sites (Winfield *et al.* 2015a) and one shore site per lake at both Bassenthwaite Lake and Derwent Water (Fig. 1) on 10th February 2015. The total number of samples (excluding blanks) was therefore *N*=78.

Offshore water sampling was carried out by boat using a Friedinger (Windermere) or Ruttner (Bassenthwaite Lake and Derwent Water) sampler (Fig. S1) deployed at a specified depth. For each 2 L water sample, five 400 ml subsamples were collected in proximity of 100 m around the sampling point, and pooled in a sterile plastic bottle (Fig. S1). The GPS location was recorded at the sampling midpoint (Appendix 1 and 2). Between samples, sampling equipment was sterilised by washing in 10% of a commercial bleach solution (containing <3% sodium hypochlorite) followed by 10% microsol detergent (Anachem, UK) and rinsed with purified water (Fig. S1). The sampler was then rinsed again in lake water at the next sampling location. 2 L of purified water was rinsed through the sampler following decontamination after every 5 samples, and the water retained as a sampling blank to allow us to check for contamination during sampling. Shoreline samples were collected by immersing a sterile 2 L plastic bottle by hand. For each sample, five 400 ml samples were collected from within a 100 m stretch of shoreline and pooled. All samples were stored in an insulated box at approximately 4 °C until filtration.

eDNA capture, extraction, amplification, library preparation and sequencing

The full 2 L of each sample was filtered through sterile 0.45 µm cellulose nitrate membrane filters and pads (47 mm diameter; Whatman, GE Healthcare, UK) using Nalgene filtration units in combination with a vacuum pump (Fig. S1). Most samples required one filter and filtered in less than an hour. For more turbid and thus slow to filter samples, a second filter was used. Filtration equipment was sterilized in 10% commercial bleach solution for 10 minutes then rinsed with 10% microsol and purified water after each filtration. Filtration blanks (2 L purified water) were run before the first filtration and then approximately after

every sixth sample, in order to test for contamination at the filtration stage. Windermere samples were filtered within 8 hours of collection in a lakeside laboratory (within the facilities of the Freshwater Biological Association, Windermere) that is not used for handling fish or DNA and was decontaminated before use by bleaching floors and surfaces. Samples from Bassenthwaite Lake and Derwent Water were filtered in a dedicated eDNA facility at the University of Hull within 12 hours of collection. Detailed operating procedures are in place in our eDNA laboratory which are aimed at avoiding contamination and access to the laboratory is strictly limited to staff who are familiar with these procedures. DNA was extracted from filters using the PowerWater DNA Isolation Kit (MoBio Laboratories, Inc. Carlsbad, USA) using the manufacturer's instructions.

Full details of the steps involved in reference database construction, *in silico* and *in vitro* primer testing, including PCR conditions, are given in the Supplementary Text. Briefly, we compiled custom, phylogenetically curated reference databases (Supplementary Text and Fig. S2) for standard mitochondrial fish DNA barcoding genes (12S and cytochrome *b*) for 67 freshwater fish species including all those recorded in the UK and additional non-native species that could potentially be present (Table S1). A number of published primers (Table S2) were evaluated against these databases *in silico* for conservation of primer binding sites and species resolution of the resulting PCR amplicons (Table S3) using the program EcoPCR (Ficetola *et al.* 2010). Two previously published primer pairs, which amplify fragments of contrasting length, from two different mtDNA regions, were selected for metabarcoding, since no single primer pair resolved all species (Table S3). The primer pair 12S_F1 and 12S_R1 (Table S2) amplifies a ~106 bp fragment of the mitochondrial 12S gene. These primers were designed and tested *in silico* (Riaz *et al.* (2011) and used in a large marine mesocosm eDNA metabarcoding study of bony fish communities (Kelly *et al.* 2014). The second selected primer pair, CytB_L14841 and CytB_H15149 (Table S2) amplifies a 460bp fragment of the cytochrome *b* gene (CytB) gene and has been used commonly for standard DNA barcoding of fishes (Kocher *et al.* 1989). Selected primer pairs were then tested *in vitro* on 22 species, firstly in individual reactions (Fig. S3) to check consistency of amplification across taxa, and secondly in 10 mock communities to evaluate whether all species amplified in competitive mixed assemblages. Mock communities were generated from spectrophotometer-quantified DNA extractions of same 22 species (Supplementary Text and Table S4) and community samples were sequenced via metabarcoding as detailed below.

Samples for metabarcoding were PCR amplified with a one-step library preparation protocol using, for each locus, 8 individually tagged forward primers and 12 individually tagged reverse primers allowing for 96 uniquely dual-indexed combinations (Kozich *et al.* 2013). All collection and extraction blanks were included in PCRs and contamination during PCR was evaluated by "amplifying" all 96 combinations of tagged primers with purified water and checking on ethidium bromide-stained agarose gels. PCRs were replicated three times for each sample, and pooled in order to minimise bias in individual PCR reactions (see Supplementary Text for full PCR conditions). Each library was normalised to approximately 1–2 ng/μl PCR product per sample using the SequalPrep Normalization Plate Kit (Invitrogen, Life Technologies) and samples subsequently pooled. Libraries were then quantified by qPCR (average of three replicate quantifications) using the KAPA Illumina Library Quantification Kit on a Roche LightCycler Real-Time PCR machine using manufacturers guidelines. Libraries were run at a 6 pM concentration on an Illumina MiSeq using the 2 x 300 bp V3 chemistry. In order improve clustering during the initial sequencing cycles 10% of PhiX genomic library was added.

Bioinformatics and data analysis

The 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 5 bp sliding windows starting from the 3'-end of the read and reads were clipped until the average quality per window was above phred 30. All reads shorter than a defined minimum read length (12S - 90bp; CytB - 100bp) were discarded. Sequence pairs were subsequently merged into single high quality reads using the program FLASH 1.2.11 (Magoč & Salzberg 2011). 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. Non-redundant sets of query sequences were then compared to the respective curated non-redundant reference database using BLAST (Zhang *et al.* 2000). BLAST output was interpreted using a custom python function, which implements a lowest common ancestor (LCA) approach for taxonomic assignment similar to the strategy used by MEGAN

(Huson *et al.* 2007). In brief, after the BLAST search we recorded the most significant matches to the reference database (yielding the top 10% bit-scores) for each of the query sequences. If only a single taxon was present in the top 10%, the query was assigned directly to this taxon. If more than one reference taxon was present in the top 10%, the query was assigned to the lowest taxonomic level that was shared by all taxa in the list of most significant hits for this query. Sequences for which the best BLAST hit had a bit score below 80 or had less than 100% / 95% identity (12S / CytB) to any sequence in the curated database, were considered non-target sequences. The custom bioinformatics pipeline used for data processing is available on Github (https://github.com/HullUni-bioinformatics/metaBEAT). To assure full reproducibility of our analyses we have deposited the entire workflow in an additional dedicated Github repository (https://github.com/HullUni-bioinformatics/Haenfling_et_al_2016). In order to obtain a qualitative assessment of the taxonomic diversity, non-target sequences were pooled across all lake samples and subjected to a separate BLAST search against NCBI's complete nucleotide (nt) database. Taxonomic assignment for non-target sequences was obtained using MEGAN 5.10.6 (Huson *et al.* 2007).

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 and is explained in full in the Supplementary Text (and corresponding Tables S4, S5 and Figs S5 and S6). 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, Tables S4, S5 and Fig. S5). At Th100 sequences of rare expected species were also lost from the mock community data (Tables S4 and S5) and the lake samples (Fig. S6). 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 (see Supplementary Text and Fig. S7 for full details). 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). Before investigating species detection and abundance estimation with eDNA, we first evaluated whether 12S and CytB datasets produced consistent results by calculating the Pearson product-moment correlation coefficient for both read count and site occupancy in R v.3.1.3. (RCoreTeam 2015).

A flow chart summarising of our analytical pipeline, from reference database compilation to data analyses is provided in Appendix 5 of the Supplementary Online Material.

Species detection using eDNA

In order to maintain a balanced sampling design, the Windermere shore sites which were only collected in a small area of the South basin, were excluded from all comparisons of species presence and abundance comparisons across basins.

First, we evaluated the performance of eDNA to detect species previously recorded in our four lake basins. Second, we used site occupancy data to investigate the spatial distribution of eDNA records within Windermere. It should be noted that full site occupancy modelling requires temporal replication to estimate the detection probability and the true proportion of occupied sites (MacKenzie *et al.* 2002). This was not possible during the current study, so our estimates of site occupancy are simply based on presence/absence, and should be treated as preliminary. We explored whether there were differences in eDNA distribution between transects, between offshore and shoreline samples, along depth profiles, and between Windermere North and South Basins. A persistent difference in species composition between the two Windermere basins has been extensively described by established sampling methods and is linked to their contrasting trophic status (Winfield *et al.* 2008a; Winfield *et al.* 2012b;

Winfield *et al.* 2008b). eDNA records from species with no preference for trophic state are consequently expected to be distributed throughout the lake, whereas eDNA from eutrophic-favouring species will be more predominant in the south than north basin and eDNA from species that prefer less eutrophic conditions will be more predominant in the north than south basin. Finally, we used sample-based rarefaction (Gotelli & Colwell 2010) to determine the number of samples needed to detect species present, focussing on Windermere, where sampling was spatially comprehensive. Rarefaction was performed with 499 randomisations in the R package Vegan (Oksanen *et al.* 2015) for CytB and 12S for the North and South Basins of Windermere combined. Only sequences corresponding to the 16 species previously recorded in Windermere were included in these analyses.

Comparison of data from eDNA and established survey methods

Summaries of fish community composition and abundance were produced for each of the four lake basins using a combination of data collected at six sites in each of our four lake basins in September 2014 using standardised survey gill-netting techniques (described in detail by (Winfield *et al.* 2015a) and (Winfield *et al.* 2015b). Gill-net survey data alone are not sufficient to describe the whole fish community since this technique under-samples or even fails to record some species, even when they are locally abundant (e.g. those with an extremely shallow distribution such as bullhead, *Cottus gobio*, or elongate morphology such as eel, *Anguilla anguilla*). Gill-net data were therefore supplemented with published information (Maberly *et al.* 2011; Pickering 2001; Winfield *et al.* 2012a; Winfield *et al.* 1996; Winfield & Durie 2004; Winfield *et al.* 2010; Winfield *et al.* 2008b) to summarise fish community compositions. This information and IJW's expert opinion developed during 25 years of sampling the four lake basins was then used to assign each recorded species to an abundance rank, with a rank of 1 given to the most abundant species by numbers. The ranking produced in this way is likely to be very robust for the most abundant species which consistently appeared in the catches of the survey gill nets, but is likely to be less so for a few species which anglers' catches indicate are present in small numbers in each lake but which are very rarely or never recorded by scientific sampling. This entire expert opinion ranking process was undertaken prior to the eDNA analysis and therefore with no knowledge of the corresponding rankings. Further details of the results from established surveys are provided in the Supplementary Text and Table S5.

A series of correlations was performed to compare the fish abundance data generated from established surveys and eDNA metabarcoding. Specifically, the relationship between eDNA data (read count and site occupancy) 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, when data was normally distributed) in R v3.1.3 (R Core team 2015). The analyses were repeated for both loci and all four sampled basins.

A work flow diagram of our entire approach is available as electronic Appendix 5.

RESULTS

The *in silico* testing of primer pairs showed that both of the chosen 12S and CytB fragments could unambiguously distinguish all species which could potentially occur at the study sites (Table S1 and S3). However, across the wider reference database a number of taxa could not be identified to the species level. *Lampetra planeri* and *L. fluviatilis,* which are probably not reproductively isolated, could not be resolved by either fragment. Additionally, 12S did not distinguish species of the genera S*alvelinus* and *Coregonus*, three species of non-native Asian carp *(Hypophthalmichthys nobilis, H. molitrix, Ctenopharyngodon idella)* and two species of the family Percidae (*Perca fluviatilis* and *Sander lucioperca).* However, given that Percidae and the genera *Coregonus* and *Salvelinus* are represented only a single species each (*Perca fluviatilis, Salvelinus alpinus and Coregonus albula* respectively) in the study area we have attributed sequence counts for the higher taxonomic levels to these individual species for further downstream analysis. This was also confirmed by the CytB data which showed that no other members of these taxonomic groups were present. Both loci amplified consistently well across 22 target species in *in vitro* testing in single species amplifications (Fig. S3). All 22 species were detected in the 12S mock communities (Table S4, Fig. S4 a), whereas three species were not detected in the CytB mock community data (Table S5, Fig. S4 b and Supplementary Text for full details). Observed and expected number of sequence reads were not significantly different for either locus (12S χ^2 = 0.224, df = 21, $P > 0.05$; CytB χ^2 = 0.367, df = 21, $P > 0.05$ Fig. S4). Moreover, there was a significant correlation between the

number of sequence reads/ng PCR template DNA for 12S and CytB (Pearson's $r = 0.599$, df = 20, $P = 0.01$, Fig. S4 c),

Clear PCR bands were obtained for all 78 eDNA samples at both loci. In contrast no target-sized bands were observed in the PCR negatives, collection or filtration blanks and we therefore decided not to sequence these. The total sequence read count passing quality control per library, before removal of chimeric sequences, was 6,306,326 for 12S and 4,793,108 for CytB (average read count per sample 71663 and 54467 respectively). After chimera removal, the 12S and CytB libraries contained 2,698,144 and 3,161,608 sequences respectively. This means that 43% of the raw dataset was non chimeric sequences for 12S, and 66% for CytB. The final libraries, after removal of redundant sequences, contained 2,562,183 sequences for 12S and 3,012,249 sequences for CytB, with average read counts per sample of 29,116 and 34,230 respectively. The proportion of target (fish) sequences ranging from 3.4-88.3% (average 23.5%) and 0-100% (average 49.0%) for 12S and CytB respectively. Most of the target sequence assignments in the lake samples were to species level with the exceptions mentioned above. The assignments to higher taxonomic levels were taken into account for calculation of total sequences read number per sample but otherwise not considered for further downstream analysis. For the CytB data of the mock communities some genus level sequence assignments were interpreted as belonging to specific species (for full details see Supplementary text and Table S5). The full sequence count data for each primer pair are available in the Supplementary Material Appendix 1 and 2).

High consistency was found between CytB and 12S in terms of both site occupancy (SO) and average read count (RC) (Fig. S8). Data from the two loci were significantly correlated (Pearson's r consistently $P < 0.05$) for all basins, for both SO and RC (Fig. S8). Consistent significant correlations were also found between SO and RC for each basin and locus (Fig. S9), therefore only the results for site occupancy are presented in the following main text. All results based on read count data are provided in the Supplementary Material.

Species detection using eDNA

The gill-net survey of September 2014 detected 25% (4/16) of the previously recorded species in Windermere. By contrast, 14 of the 16 previously recorded species (i.e. 88%) were detected using 12S and 75% (12/16) using CytB across the entire lake. Within each Windermere basin 13 previously-recorded species were detected with 12S whereas 12 and 11 species were detected for the North and South Basins respectively with CytB (Fig. 2 a, b; Fig. S10). A number of additional species were also detected in Windermere, including *C. carpio*, *Gymnocephalus cernuus*, *Leucaspius delineatus*, *O. mykiss*, *Osmerus eperlanus* (12S), *Platichthys flesus* and *Pseudorasbora parva* (CytB). Two species that have been recorded in Windermere but are not present in the sequence data are the two lamprey species *L. fluviatilis* and *Petromyzon marinus*. In the 12S data set the majority of potential false positives were found in a single sample from Windermere North Basin which was consequently omitted from all further analysis (sample W14). Gill-net sampling detected 60% (6/10) of the species known to be present in Bassenthwaite Lake whereas 90% (9/10) of species were detected using 12S and 70% (7/10) with CytB (Fig. 2 c; Fig. S10). Additional species not previously recorded in Bassenthwaite included *Abramis brama* (CytB), and *Barbatula barbatula*, *G. aculeatus,* and *S. erythrophthalmus* (12S, Fig. 2 c). In Derwent Water, gill-net sampling in September 2014 detected 77% (7/9) recorded species, whereas 88% (8/9) of species were detected with 12S and 67% (6/9) with CytB (Fig. 2 d; Fig. S10). The 12S assay detected an additional four species previously unrecorded, including *B. barbatula*, *G. aculeatus*, *Pungitius pungitius* and *S. erythrophthalmus*.

Sample-based rarefaction analyses on the combined Windermere data set indicated that approximately 10-25 samples captures the majority (~85%) of the taxa present in the entire sample although the number of samples required to achieve the same taxon coverage is higher for CytB (Fig. 3).

Estimating abundance with eDNA

There was a consistent, negative relationship between eDNA site occupancy and long-term rank (where rank abundance decreases from 1-16) and this correlation is highly significant for Windermere North and South Basins, for both loci (Fig. 4 a, b, e, f). Similar trends were found for Bassenthwaite Lake and Derwent Water but correlations were not significant (Fig.

4 c, d, g, h). The number of sequence reads was also significantly correlated with long-term rank in Windermere North and South Basins, for both loci (Fig. S11 a, b, e, f). Again similar trends were seen for Derwent Water and Bassenthwaite Lake but only the correlation for Derwent Water at 12S is significant (Fig. S11 c, d).

Site occupancy and number of sequence reads were also compared against actual numbers sampled in the September 2014 gill-net surveys for all four basins (Figs S12 and S13 respectively). There was a consistent positive relationship between abundance data from the recent gill-net surveys and eDNA (both read count and occupancy, and both loci), in spite of the small number of species (4-6) detected in the gill net surveys and hence low statistical power in the analyses. However only the correlations for CytB read count were consistently significant in all basins (Fig. S13 e-h), and this result may be driven by the high abundance and read count for *P. fluviatilis*.

Spatial distribution of eDNA records within Windermere

Comparing the distribution of eDNA data by transect indicates a slight trend for more species to be detected at inshore versus deeper mid-lake regions (Fig. 5). With 12S, 13 species were detected in samples from the 5 m transect compared to 10 from the mid-line. Twelve species were detected in the 6 geographically-close shore samples. A similar trend was found for CytB, with 11 species detected in both 5 m transect and shore samples, compared to 8 in the mid-line (Fig. 5). Depth profiles in the North and South Basins revealed that eDNA from the majority of detected species was distributed throughout the water column (Fig. S14). Within the depth profiles, *A. anguilla* and *S. alpinus* were only detected in deep water in the North Basin (≥60 m and 30 m respectively, Fig. S14 a and c). Similarly, in the South Basin depth profile *P. phoxinus* and *S. salar* were only detected at the deepest sampling point (40 m) (Fig. S14 b and c).

Site occupancy data based on 12S sequences were used to investigate the spatial distribution of each species recorded at more than two sites around Windermere (Fig. S15). The general pattern emerging from this analysis is that species-specific eDNA was not evenly distributed around the lake. Although some species such as *P. fluviatilis*, *R. rutilus*, *E. lucius* and *S.*

trutta, are recorded almost ubiquitously within the lake, eDNA from other species is predominantly found in one of the two basins. *S. alpinus*, *P. phoxinus* and *G. aculeatus* eDNA was common in the North Basin but very rare in the South Basin, whereas *A. brama* and *A. anguilla* eDNA is more common in South Basin (Fig. S15). Overall the relative proportion of sequence read counts for different species across sample sites was significantly different between Windermere North and South Basins (χ^2 = 47817; df = 13; $P < 0.001$ and χ^2 = 134750; df = 11; $P < 0.001$ for 12S and CytB respectively, Fig. 6 a, b). A similar pattern was observed for the relative proportion of sites occupied (χ^2 = 61.43; df = 13; $P < 0.001$ and χ^2 = 48.65; df = 11; $P < 0.001$ for 12S and CytB respectively Fig. 6 c, d). Distribution of eDNA reflected in the two Windermere Basins reflected the expected association between species and ecological condition. eDNA from species associated with eutrophic conditions (*R. rutilus*, *T. tinca, S. erythrophthalmus*, *A. brama,* and *A. anguilla*) was more abundant in the South than North Basin, while eDNA from species that prefer less eutrophic conditions (*S. salar, S. trutta, S. alpinus, P. phoxinus,* and *C. gobio*) was more abundant in the North than South Basin (Fig. 6).

Non-fish sequences

A large proportion of both 12S and CytB sequences could not be assigned to UK freshwater fish from the custom database, and were compared to the NCBI database using BLAST. Non-fish sequences included a wide range of species directly associated with aquatic habitats including mammals such as otter, *Lutra lutra* and birds, including moorhen, *Gallinula chloropus*; cormorant, *Phalacrocorax carbo* and various duck and geese species found within the UK. The list also included many other vertebrate species potentially occurring in the wider catchment area (Table S6) including domesticated farm animals such as cow, *Bos taurus*; sheep, *Ovis aries* and chicken, *Gallus gallus domesticus*, and wild vertebrates such as red deer, *Cervus elaphus*; red squirrel, *Sciurus vulgaris*; red fox, *Vulpes vulpes* and tawny owl, *Strix aluco*. Sequences assigned to *Homo sapiens* were also abundant, likely present as genuine eDNA found in lake water due to the high degree of human interaction with the lakes through water sports, angling and waste water, or present as a laboratory contaminant. 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.

DISCUSSION

In this study we used high-throughput sequencing of eDNA from the mitochondrial 12S and CytB genes to characterise the fish community composition in three large lakes (Lake Windermere, Derwent Water and Bassenthwaite Lake) in the UK. eDNA data was compared to comprehensive long-term data on fish distribution and abundance from established survey methods. eDNA outperformed established methods in terms of species detection. More surprisingly, eDNA data accurately reflected the rank abundance of species within the lake fish community, suggesting eDNA methods may be more quantitative than previously thought.

Comparison of of eDNA and established methods for species detection

eDNA metabarcoding was effective in detecting fish species when compared against decades of data from established sampling techniques and other sources (as described most recently by Winfield *et al.* 2015a and Winfield *et al.* 2015b). In Windermere, 60 offshore (30 for each basin) and 6 shoreline samples were analysed and 14 of the 16 previously-recorded species were detected. The two rarest species, river lamprey, *L. fluviatilis* and sea lamprey, *P. marinus,* were not detected in the eDNA data, but these species were unlikely to be present in the lakes at the time of sampling and temporally replicated sampling is required to address this issue. Other rare species such as tench, *T. tinca* and rudd, *S. erythropthalmus* were detected at low levels with 12S in the North and South Basins respectively. The results of the rarefaction analysis on the Windermere data indicate that a detection probability of over 85% can be achieved with a substantially lower number of samples; approximately 10 for 12S and 25 for CytB. In contrast, only the four most common species were detected in the gill net survey from 2014, which is typical of surveys (4-5 species have been typically sampled each year since 2011, Winfield *et al.* 2012c; Winfield *et al.* 2013; Winfield *et al.* 2014).

The eDNA results from Bassenthwaite Lake and Derwent Water were also remarkably concordant with the fish community based on long-term gill-netting (Thackeray *et al.* 2006) given that only six samples were collected per lake. All but the rarest species were detected in Derwent Water and Bassenthwaite (dace, *L. leuciscus*, and vendace, *C. albula* respectively) using 12S. Dace was however detected in Bassenthwaite, and vendace in Derwent Water with

12S, while neither species was detected with CytB. Dace has been recorded intermittently and in low numbers in Derwent Water within the last decade (Thackeray *et al.* 2006) but was not detected by gill netting in 2014 (Winfield *et al.* 2015a). Vendace is known to occur only in a restricted deep area of Bassenthwaite Lake and only three individuals have been recorded in gill-net surveys since 2000 (Winfield *et al.* in press). In these cases DNA concentration might fall below the detection threshold of the PCR assay or those which were set for the bioinformatics analysis in order to reduce the possibility of "false positives". Roach, *R. rutilus*, on the other hand, is a common species in all four basins, but was not detected with CytB in Bassenthwaite and Derwent Water. This species was also detected in the CytB mock community at lower than expected frequency, suggesting that the CytB primers may not amplify this species well in competitive reactions.

Overall, eDNA metabarcoding data produced a more comprehensive species list than gill net surveys with a similar effort. The under-representation of species in gill-netting surveys is an acknowledged sampling artefact which has a number of causes including fish morphology (e.g. eel species are not susceptible to retention in gill nets), fine-scale spatial distribution (e.g. three-spined stickleback may be limited to the extreme inshore where nets cannot be deployed) or movement patterns (e.g. bullhead may be unlikely to be sampled by gill nets due to their relatively limited movements). This corroborates results from Thomsen *et al.* (2012a) and Valentini *et al.* (2015) who showed that eDNA metabarcoding data detected more species of marine fish than alternative surveying techniques.

Detection of previously unrecorded species with eDNA

Eight previously unrecorded species were detected in Lake Windermere, four in Bassenthwaite Lake and four in Derwent Water. In most cases these eDNA records were at very low occupancy (1 or 2 sites) and read counts (0.1%-1.0%), just above our threshold for accepting a positive record. These records could be either genuine detections of species that have been missed with established methods, false positives from sequencing error (barcode misassignment, Deakin *et al.* 2014; or "tag jumps" Schnell *et al*. 2015), laboratory or environmental contamination (i.e. the presence of DNA in the environment from, for example, the wider watershed, bird faeces, waste water or fishing bait). The unexpected records likely originate from a combination of factors, discussed below.

Only one of the eight previously unrecorded Windermere species, ruffe, *G. cernua,* was detected at high frequencies with eDNA. 12S sequences were present in 27% of the sites in the South Basin and 38% of the sites in the North Basin although the species was not detected with CytB. This species has been recently introduced to a number of Cumbrian lakes (Winfield *et al.* 2010), and is present in Rydal Water approximately 3 km upstream of Windermere. It is therefore possible that *G. cernua* has colonised Windermere and is present at very low abundance (below the detection limits of gill-netting programme), or that eDNA has been transported from the *G. cernua* populations upstream. Three kilometres is well within the range of eDNA transport distances that have previously been recorded (Deiner and Altermatt 2015). Absence of positive records with the long CytB fragment also suggests that only relatively degraded *G. cernua* DNA was present in the lake, lending further support to this hypothesis. Although this species was present in the mock communities, the high frequency of occurrence means it is unlikely that this result can be explained by sequencing errors such as barcode misassignment.

The other seven previously-unrecorded Windermere species (common carp, *C. carpio;* sunbleak, *L. delineates*; topmouth gudgeon, *P. parva*; rainbow trout, *O. mykiss*; smelt, *O. eperlanus*; flounder, *P. flesus* and mudminnow, *U. pygmea*) were detected at very low levels. The actual presence of *U. pygmea*, *L. delineates* and *P. parva*, in Windermere seems extremely unlikely since their known distribution does not overlap with the Windermere catchment. Given that all three species were included in the mock communities these records are most likely explained by low level laboratory contamination or sequencing barcode misassignment from the mock communities into the samples (Deakin *et al.* 2014). *O. mykiss, O. eperlanus* and *P. flesus*, do occur in the catchment and the former two species are also a very popular dead bait used by pike anglers. Since none of these species have been handled in the laboratory and pike anglers were active during the water sampling, it seems that such dead baiting or eDNA transport from other parts of the catchment are likely sources of eDNA for these species in the lake. *C. carpio,* was recorded with both CytB and 12S at one of the shore sites. The fact that both markers were recorded at the same site indicates that common carp DNA and individuals might have been present in the lake water but highly localised and undetected by established sampling techniques. However this species was also present in the mock communities and therefore laboratory contamination or "tag jumping" cannot be excluded.

Four previously-unrecorded species were detected in each of the Bassenthwaite and Derwent Water basins. Again most of these records were based on low sequence reads and site occupancy. The records for some species (common bream, *A. brama* in Bassenthwaite Lake, nine-spined stickleback, *P. pungitius* in Derwent Water) are most likely explained by barcode misassignment because they have never been recorded in the catchment but are present in the mock communities. The presence of the remaining species (stone loach, *B. barbatula*; three-spined stickleback, *G. aculeatus*; and rudd, *S. cephalus*) in the lakes or in the catchment cannot be so easily excluded. These records therefore could either represent environmental contamination or indicate that the species are present at low numbers and have not been detected by previous long-term gill-netting (summarised by Winfield *et al.* 2012a).

We quantified the level of background contamination using sequence information from mock communities and the level of PhiX contamination in target samples, which enabled us to choose a suitable threshold level for filtering the data for false positives without losing more information than necessary. Ultimately though, it is not possible to distinguish between false positives and true positives if they occur at the same frequency, and some rare species are likely to be lost with a threshold approach. Using consistency across technical replicates as recently used by Port *et al.* (2016) might be a more suitable approach to control for false positive if rare species are of particular interest.

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 read counts or site occupancy as proxies. Using read count data is a valid approach under the assumption that no significant bias is introduced during sampling, subsequent PCR or sequencing. However, this assumption is unrealistic, and previous studies have demonstrated that the relationship between abundance and read count is complex (e.g. Ficetola *et al.* 2015; Yu *et al.* 2012; Evans *et al.* 2015; Kelly *et al.* 2014). Site occupancy models have been developed to cope with multiple levels of bias and uncertainty (e.g. imperfect detection, MacKenzie *et al.* 2002) and are therefore highly promising for eDNA (Schmidt *et al.* 2013). As discussed in the Methods, full site occupancy modelling requires estimation of detection probability from temporal sampling, which was beyond the scope of

the present study. Our site occupancy estimates should therefore be treated as preliminary. Encouragingly though, read count and site occupancy data were correlated for each basin and each locus, suggesting that both measures of abundance are informative. As we discuss below though, and not surprisingly, site occupancy relies on comprehensive spatial sampling to obtain sufficient power for estimating abundance.

We found a consistent significant relationship between rank abundance and read count or occupancy data for both basins of Lake Windermere. This indicates both read count and occupancy are equally effective at estimating relative abundance under comprehensive spatial sampling. In Derwent Water and Bassenthwaite Lake, correlations with both abundance measures are weak and not significant with one exception (number of 12S sequence reads for Derwent Water). We suggest this is related to low statistical power from analysing only six samples per lake. There was also a consistent trend between eDNA and gill-net data, but the results are less conclusive due to low statistical power from the small number of species sampled in the gill-net survey. Although these results are generally encouraging, further work is critically needed to determine how robust eDNA is for estimating abundance. Increased spatial coverage of Bassenthwaite Lake and Derwent Water, together with temporal sampling to allow estimation of detection probability and site occupancy modelling in all basins, are critical next steps.

Spatial distribution of eDNA in Lake Windermere

We investigated the spatial distribution of eDNA in Lake Windermere by comparing 1) off shore and shoreline samples, 2) three depth profile transects and 3) North and South Basins, which differ in their trophic status. Firstly, more species were detected in shallower than in deep water, with 13 species detected along the 5 m contour, compared to 9 in the mid-line transect. Interestingly, 12 of the 16 previously-recorded species were detected in the 6 shore samples, which were collected in close proximity to one another. This suggests eDNA could accumulate on the shoreline, and that shoreline sampling could be adequate for detection of most species. More rigorous sampling along the lake shore is needed to investigate this further. Second, we expected little difference along depth profile transects since our sampling was carried out in the winter, when water stratification has broken down. As predicted, within the depth transects the majority of species were detected throughout the water column but

some, including the typically deep water species Arctic charr, *S. alpinus,* were only detected at the deepest sampling points, indicating that surface water sampling might be ineffective in deeper lakes. Given the small scale of this experiment the results regarding vertical sampling should be regarded as preliminary. Thirdly, we hypothesized that eDNA from species associated with less eutrophic (i.e. mesotrophic) conditions would be more abundant in the North Basin, while eDNA from species associated with more eutrophic conditions should be more abundant in the South Basin, and species with no preference should be detected throughout the lake. We observed clear differences in the spatial distribution of eDNA, consistent with this hypothesis. These results are consistent with long-term datasets from trapping, gill-netting and recreational anglers' catches (Winfield *et al.* 2008a; Winfield *et al.* 2008b; Winfield *et al.* 2011; Craig *et al.* 2015; Winfield *et al.* 2015b). For example, established methods have found perch, *P. fluviatilis* and pike, *E. lucius* consistently in both basins (Craig *et al.* 2015; Winfield *et al.* 2008a respectively) while *S. alpinus* is much more abundant in the North than in the South Basin (Winfield *et al.* 2008b; Winfield *et al.* 2015b) and *A. brama*, although a relatively minor component of the Windermere fish community, is consistently more abundant in the South than in the North Basin (Winfield *et al.* 2011).

Technical approach and the use of 12S or CytB as a marker

In the present study we chose to validate the assays by sequencing mock communities, constructed from 22 species of fish, on the same flow cell as the eDNA samples. Although this allows for the success of the assay to be assessed within the same sequencing library as the samples, this approach may cause problems due to the low level miss-assignment of sequences from the mock community to the samples. For future studies we would recommend not including mock communities in the same library, or only including species that have no chance of being found in the eDNA samples and to sequence all negative controls and blanks.

Both markers were generally consistent in terms of the number of read counts and occupancy data generated, although clear advantages and disadvantages were associated with each marker. All species were detected in the mock communities with 12S whereas three were undetected with CytB. In the eDNA samples, site occupancy was higher, and more species were detected with 12S than CytB, as discussed earlier. Differences in amplification success

could be due to fragment size (~100bp for 12S and 460bp for CytB), mismatches in primer binding sites or both. Given that eDNA degrades rapidly in the environment (Barnes *et al.* 2014; Rees *et al.* 2014), the difference in detection is probably a result of longer persistence of the shorter 12S fragment in lake water. This may allow for dispersion of eDNA across a larger geographical scale, increasing the probability of detection at any site. Consequently, it may be that detection of the longer CytB fragment indicates the species is present closer to where the water sample was taken, while 12S fragments may have originated from some distance away either within the lake or even up its tributaries. Using a longer fragment may be useful for pinpointing the exact location of species, but 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. An additional aspect to consider is the persistence of eDNA in sediments, which has been shown to be considerably longer when compared to the water column (Turner *et al.* 2014). Differential persistence of the different sized fragments, and resuspension of eDNA during rain events could account for historical eDNA being detected. However, differences in primer specificity and efficiency between the two genes prevent conclusive answers to these issues, and this issue warrants further systematic exploration through experimental approaches and analysing a wider range of eDNA fragment lengths.

Use of eDNA to survey non-fish vertebrates

This study also offers some insights into the feasibility of eDNA techniques for the wider assessment of non-fish vertebrates associated with lakes and their immediate catchments. The majority of the 12S and CytB sequences generated did not match the comprehensive UK fish reference database used and non-fish sequences could be assigned to 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 our reference data base. Moreover, the primers used appear to be largely vertebrate-specific since no invertebrate sequences were identified, although many such species are 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 lake catchments.

Conclusions

The present investigation was driven primarily by the need to develop robust and cost-effective lake fish assessments to meet the requirements of the EC 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, but 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 indicated that eDNA approaches can make a very significant contribution to this challenging task. The results were consistent with our understanding of the fish communities of three large, deep lakes based on long-term monitoring using established techniques. 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 assessments and ecological interpretations. Although the eDNA approach cannot produce information on individual condition or population characteristics such as growth curves, it proved to be very effective at producing robust data at the community level which is undoubtedly the most challenging task for established sampling methods.

eDNA is arguably one of the most rapidly expanding areas of research in molecular ecology but there is much to learn before methods such as the one described here can be deployed for biological monitoring; particularly under legislative or sensitive circumstances. Temporal sampling is an essential next step from the current study, to account for imperfect detection and fully test the site occupancy modelling approach, and to investigate the effects of water stratification on the spatial distribution of eDNA. More generally, there is a pressing need to develop and demonstrate the wider applicability of eDNA to a greater range of water bodies (such as those with varied chemical and physical properties) as well as other animal and plant communities.

Acknowledgments This work was funded by a UK Environment Agency contract (SC140018) awarded to BH, LLH, DR and IJW. We are particularly grateful to Drs Kerry Walsh and Graeme Peirson for initiating the study and for support throughout. We gratefully acknowledge the Freshwater Biological Association for providing access to their laboratory facilities and United Utilities for use of gill-netting data. Ben James and Janice Fletcher

provided invaluable help during field work, while Drs Tony Dejean and Joachim Mergeay contributed to helpful discussions on eDNA approaches and Dave Lunt provided excellent advice on the bioinformatics analysis. We would like to thank Drs Holly Bik, Kristy Deiner and Cameron Truner for constructive criticism on the initial submission which helped to strengthen the manuscript

References

Argillier C, Caussé S, Gevrey M, *et al.* (2013) Development of a fish-based index to assess the eutrophication status of European lakes. *Hydrobiologia* **704**, 193-211.

Barnes MA, Turner CR, Jerde CL, *et al.* (2014) Environmental Conditions Influence eDNA Persistence in Aquatic Systems. *Environmental Science & Technology* **48**, 1819-1827.

Biggs J, Ewald N, Valentini A, *et al.* (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 HM, Porazinska DL, Creer S, *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution* **27**, 233-243.

Bohmann K, Evans A, Gilbert MTP, *et al.* (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution* **29**, 358-367.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*.

Deakin CT, Deakin JJ, Ginn SL, *et al.* (2014) Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. . *Nucleic Acids Research* **42**, e129.

European_Communities (2000) Directive 2000/60/EC, Establishing a framework for community action in the field of water policy. *Official Journal of the European Communities* **L 327**, 1-71.

Evans NT, Olds BP, Renshaw MA, *et al.* (2015) Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, n/a-n/a.

Ficetola G, Coissac E, Zundel S, *et al.* (2010) An In silico approach for the evaluation of DNA barcodes. *Bmc Genomics* **11**, 434.

Ficetola GF, Pansu J, Bonin A, *et al.* (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources* **15**, 543-556.

Fonseca VG, Carvalho GR, Sung W, *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications* **1**, 98.

Gotelli NJ, Colwell RK (2010) Estimating species richness. In: *Biological Diversity: Frontiers In Measurement And Assessment.* (eds. Magurran AE, McGill BJ), p. 345. Oxford University Press, Oxford.

Hunter ME, Oyler-McCance SJ, Dorazio RM, *et al.* (2015) Environmental DNA (eDNA) Sampling Improves Occurrence and Detection Estimates of Invasive Burmese Pythons. *Plos One* **10**, e0121655.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research* **17**, 377-386.

Kelly RP, Port JA, Yamahara KM, Crowder LB (2014) Using Environmental DNA to Census Marine Fishes in a Large Mesocosm. *Plos One* **9**.

Kocher T, Meyer A, Edwards S, *et al.* (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**, 6196-6200.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Applied and Environmental Microbiology* **79**, 5112-5120.

Kubečka J, Hohausová E, Matěna J*, et al.* (2009) The true picture of a lake or reservoir fish stock: A review of needs and progress. *Fisheries Research* **96**, 1-5.

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

Maberly SC, De Ville MM, Thackeray SJ*, et al.* (2011) A survey of the lakes of the English Lake District: the Lakes Tour 2010. *Report to Environment Agency, North West Region and Lake District National Park Authority. LA/NEC04357/1. 223 pp.*

MacKenzie DI, Nichols JD (2004) Occupancy as a surrogate for abundance estimation. . *Animal: an international journal of animal bioscience. Available at: http://abc.museucienciesjournals.cat/files/ABC-27-1-pp-461-467.pdf.*

MacKenzie DI, Nichols JD, Lachman GB*, et al.* (2002) ESTIMATING SITE OCCUPANCY RATES WHEN DETECTION PROBABILITIES ARE LESS THAN ONE. *Ecology* **83**, 2248-2255.

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

Mahon A, Nathan L, Jerde C (2014) Meta-genomic surveillance of invasive species in the bait trade. *Conservation Genetics Resources* **6**, 563-567.

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

Oksanen J, Guillaume Blanchet F, Kindt R*, et al.* (2015) Vegan: Community Ecology Package. R package version 2.2-1. http://CRAN.R-project.org/package=vegan.

Pickering AD (2001) *Restoring the Health of England's Largest Lake* Freshwater Biological Association, Ambleside, U.K.

Pilliod DS, Goldberg CS, Arkle RS, Waits LP (2013) Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences* **70**, 1123-1130.

Port JA, O'Donnell JL, Romero-Maraccini OC, Leary PR, Litvin SY, Nickols KJ, Yamahara KM, Kelly RP (2016) Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology* **25**, 527-541

RCoreTeam (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rees HC, Gough KC, Middleditch DJ, Patmore JRM, Maddison BC (2015) Applications and limitations of measuring environmental DNA as indicators of the presence of aquatic animals. *Journal of Applied Ecology* **52**, 827-831.

Rees HC, Maddison BC, Middleditch DJ, Patmore JRM, Gough KC (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*, et al.* (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research* **39**.

Schmidt BR, Kery M, Ursenbacher S, Hyman OJ, Collins JP (2013) Site occupancymodels in the analysis of environmental DNA presence/absence surveys: a case study of an emerging amphibian pathogen. *Methods in Ecology and Evolution* **4**, 4646–4653.

Thackeray SJ, Maberly SC, Winfield IJ (2006) The ecology of Bassenthwaite Lake (English Lake District). *Freshwater Forum* **25**, 1-80.

Thomsen PF, Kielgast J, Iversen LL*, et al.* (2012a) Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples. *Plos One* **7**.

Thomsen PF, Kielgast J, Iversen LL*, et al.* (2012b) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology* **21**, 2565-2573.

Turner CR, Uy KL, Everhart RC (2014) Fish environmental DNA is more concentrated in aquatic sediments than surface water. *Biological Conservation* **183**, 93-102.

Valentini A, Taberlet P, Miaud C*, et al.* (2015) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, n/a-n/a.

Winfield I, Fletcher J, James JB (2008a) The Arctic charr (Salvelinus alpinus) populations of Windermere, UK: population trends associated with eutrophication, climate change and increased abundance of roach (Rutilus rutilus). *Environmental Biology of Fishes* **83**, 25-35.

Winfield IJ, Adams CE, Bean CW*, et al.* (2012a) Conservation of the vendace (*Coregonus albula*), the U.K.'s rarest freshwater fish. . *Advances in Limnology* **63**, 547-559.

Winfield IJ, Cragg-Hine D, Fletcher JM, Cubby PR (1996) The conservation ecology of *Coregonus albula* and *C. lavaretus* in England and Wales, U.K. . In: *Conservation of Endangered Freshwater Fish in Europe* (eds. Kirchhofer A, Hefti D), pp. 213-223. Birhauser Verlag, Basel.

Winfield IJ, Durie NC (2004) Fish introductions and their management in the English Lake District. *Fisheries Management and Ecology* **11**, 195-201.

Winfield IJ, Fletcher JM, Ben James J (2012b) Long-term changes in the diet of pike (Esox lucius), the top aquatic predator in a changing Windermere. *Freshwater Biology* **57**, 373-383.

Winfield IJ, Fletcher JM, James JB (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**, 60-65.

Winfield IJ, Fletcher JM, James JB (2011) Invasive fish species in the largest lakes of Scotland, Northern Ireland, Wales and England: the collective UK experience. *Hydrobiologia* **660**, 93-103.

Winfield IJ, Fletcher JM, James JB (2012c) *Monitoring the fish populations of Windermere, 2011*.

Winfield IJ, Fletcher JM, James JB (2014) Monitoring the fish populations of Windermere, 2013. . *Report to Environment Agency, North West Region. LA/NEC05043/2. 74 pp*.

Winfield IJ, Fletcher JM, James JB (2015a) *Monitoring the fish populations of Bassenthwaite Lake and Derwent Water, 2014.* .

Winfield IJ, Fletcher JM, James JB (2015b) *Monitoring the fish populations of Windermere, 2014*. Report to United Utilities. LA/NEC05364/2. 66pp.

Winfield IJ, Fletcher JM, James JB (in press) The 'reappearance' of vendace (Coregonus albula) in the face of multiple stressors in Bassenthwaite Lake, U.K. *Fundamental & Applied Limnology*.

Winfield IJ, James JB, Fletcher JM (2008b) Northern pike (*Esox lucius*) in a warming lake: changes in population size and individual condition in relation to prey abundance. . *Hydrobiologia* **601**, 29-40.

Yu DW, Ji YQ, Emerson BC*, et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution* **3**, 613-623.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology* **7**, 203-214.

Data accessibility: All de novo sequences generated through Sanger sequencing made available directed through our archived analysis pipeline on Github (see below). Accession numbers and taxon affiliations of all curated sequences are available as electronic Appendices. Raw Illumina read data has been submitted to NCBI (BioProject: PRJNA313432; BioSample accessions: SAMN04530423-SAMN04530510; SRA accessions: SRR3359939-SRR3360124). To assure full reproducibility of our analyses we have deposited the entire bioinformatics workflow in a dedicated Github repository, which also contains the

curated reference databases and further supplementary data, such as taxon specific read counts for each sample as tables (https://github.com/HullUni-bioinformatics/Haenfling_et_al_2016; the repository is permanently archived with Zenodo (DOI 10.5281/zenodo.49823). Our custom data processing pipeline is available on Github (https://github.com/HullUni-bioinformatics/metaBEAT).

Author contributions: B.H., L.L.H. and I.J.W., conceived the study; B.H., L.L.H., I.J.W., J.L. and R.B.; carried out the field work. I.J.W. prepared fish abundance data from established method surveys. P.N., J.L and R.B. carried out all pre-sequencing laboratory work. D.R. assisted in the design of the molecular assays and carried out Illumina sequencing and the initial steps of the raw data analysis; A.O. assisted with the Illumina sequencing; C.H. assembled the bioinformatics pipeline and reference data base and wrote the relevant sections of the manuscript. B.H., and L.L.H. performed the statistical analyses. B.H., L.L.H., I.J.W. and D.R. wrote the paper; all authors commented on the final draft.

Tables

Table 1: Species previously recorded in the study lakes or recorded with eDNA. Full scientific, common names and three letter codes used in figures are given.

Scientific Name	Common Name	Code	Previously recorded in study lakes
Abramis brama	Common bream	BRE	Yes
Anguilla anguilla	European eel	EEL	Yes
Barbatula barbatula	Stone loach	LOA	Yes
Coregonus albula	Vendace	VEN	Yes
Cottus gobio	Bullhead	BUL	Yes
Cyprinus carpio	Common carp	CAR	No
Esox lucius	Pike	PIK	Yes
Gasterosteus aculeatus	Three-spined stickleback	3SS	Yes
Gymnocephalus cernua (=cernuus)	Ruffe	RUF	Yes
Lampetra fluviatilis	River lamprey	RLA	Yes
Leucaspius deliniatus	Sunbleak	SUN	No
Leuciscus leuciscus	Dace	DAC	Yes
Oncorhynchus mykiss	Rainbow trout	RTR	No
Osmerus eperlanus	Smelt	SME	No
Perca fluviatilis	Perch	PER	Yes
Petromyzon marinus	Sea lamprey	SLA	Yes

Phoxinus phoxinus	Minnow	MIN	Yes
Platichthys flesus	Flounder	FLO	No
Pseudorasbora parva	Topmouth gudgeon	TMG	No
Pungitius pungitius	Nine-spined stickleback	9SS	No
Rutilus rutilus	Roach	ROA	Yes
Salmo salar	Atlantic salmon	SAL	Yes
Salmo trutta	Brown trout	BTR	Yes
Salvelinus alpinus	Arctic charr	CHA	Yes
Scardinius erythrophthalmus	Rudd	RUD	Yes
Squalius cephalus (=Leuciscus cephalus)	Chub	CHU	Yes
Tinca tinca	Tench	TEN	Yes
Umbra pygmaea	Mudminnow	MUD	No

Figure legends

Figure 1: Sampling sites in the three study lakes a) Bassenthwaite Lake, b) Derwent Water, and c) Windermere in the English Lake District (UK). Samples were collected from gill net sites (orange circles) and single shoreline sites (yellow circles) in Bassenthwaite Lake and Derwent Water. In Windermere, samples were collected along transects following the 5 m (red circles), 20 m (green circles) and mid line (blue circles) depth profiles, as well as additional gill net and shoreline sites.

Figure 2: Site occupancy for 12S and CytB data from a) offshore sites Windermere North Basin, b) offshore sites Windermere South Basin, c) Bassenthwaite Lake and d) Derwent Water. All species recorded previously are included. Previously-recorded species are ordered according to their rank abundance within basin from established survey methods. Species that have not been recorded previously are indicated with an asterisk and are ordered alphabetically. Full species names are given in Table 1.

Figure 3: Sample based rarefaction analyses for Lake Windermere. Only offshore samples and species recorded previously in Lake Windermere are included in the analyses.

Figure 4: Correlations between site occupancy data and long-term rank based on established surveys and expert opinion for all four basins and both 12S (a-d) and CytB (e-h), where 1 is the highest and 16 the lowest rank abundance. Species three letter codes are given in Table 1.

Figure 5: Average number of sequence reads obtained per transect for Lake Windermere North Basin (a,b,) and South Basin (c,d) for both 12S (a,c) and CytB (b,d). Only species that have been recorded previously are included. Species are ordered according to their rank abundance within basin from established survey methods.

Figure 6: Relative distribution of fish species and their ecological preferences in Windermere North Basin (mesotrophic) and South Basin (eutrophic) based on the proportion from the total number of sequence reads (a, b) and the relative proportion of sites occupied (c,d) reflecting the trophic status of the two basin**.**