Development of an eDNA method for monitoring fish communities: ground truthing in diverse lakes with characterised fish faunas

Running title: eDNA method for lake fish monitoring

Jianlong Li[1]*, Tristan W. Hatton-Ellis[2], Lori-Jayne Lawson Handley[1], Helen S. Kimbell[1, 3], Marco Benucci[1], Graeme Peirson[4], Bernd Hänfling[1]

1 Evolutionary and Environmental Genomics Group (@EvoHull), School of Environmental Sciences, University of Hull (UoH), Cottingham Road, Hull, HU6 7RX, UK

2 Natural Resources Wales (NRW), Maes-y-Ffynnon, Penrhosgarnedd, Bangor, Gwynedd, LL57 2DW, UK

3 Frontiers, EPFL Innovation Park, Building I, Lausanne, CH-1015, Switzerland

4 Environment Agency (EA), Mance House, Hoo Farm Industrial Estate, Kidderminster, DY11 7RA, UK

*Corresponding author:

haikuilee@gmail.com (J.L.)

Abstract

1. Accurate, cost-effective monitoring of fish is required to assess the quality of lakes under the European Water Framework Directive (WFD). Recent studies have shown that eDNA metabarcoding is an effective and non-invasive method, which can provide semi-quantitative information on fish communities in large lakes.

2. This study further investigated the potential of eDNA metabarcoding as a tool for WFD status assessment by collecting and analysing water samples from eight Welsh lakes and six meres in Cheshire, England, with well described fish·faunas. Water samples (N = 252) were assayed using two mitochondrial DNA regions (Cytb and 12S rRNA).

3. eDNA sampling indicated very similar species to be present in the lakes compared to those expected on the basis of existing and historical information. In total, 24 species with 111 species occurrences by lake were detected using eDNA. Secondly, there was a significant positive correlation between expected faunas and eDNA data in terms of confidence of species occurrence. Thirdly, eDNA data can estimate relative abundance with the standard five-level classification scale ("DAFOR"). Lastly, four ecological fish communities were characterised using eDNA data which were agreed with the pre-defined lake types according to environmental characteristics.

4. *Synthesis and applications*. This study provides further evidence that eDNA metabarcoding could be a powerful and non-invasive monitoring tool for WFD purpose in a wide range of lake types, considerably outperforming other methods for community level analysis.

Keywords

eDNA, metabarcoding, fish monitoring, community ecology, EC Water Framework Directive

Introduction

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

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

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

Materials and methods

Study sites

The distribution of the 14 sampling lakes including eight Welsh lakes and six Cheshire meres are shown in Fig. 1, and the general characteristics are outlined in Appendix S1 Table S1.1 based on data from the UK Lakes Portal (Hughes et al. 2004). In brief, the 14 lakes can be divided into three types according to environmental characteristics. These three types were: Type 1: three low alkalinity lakes that are broadly upland in character (Llyn Cwellyn, Llyn Ogwen and Llyn Padarn); Type 2: two high alkalinity but shallow lakes (Llyn Traffwll and Llyn Penrhyn) on the west coast of Anglesey (North Wales) which are close to sea and accessible for migratory fish and Type 3: nine high alkalinity but shallow lakes that are dominated by coarse fish (Kenfig Pool, Llan Bwch-llyn, Llangorse Lake, Maer Pool, Chapel Mere, Oss Mere, Fenemere, Watch Lane Flash and Betley Mere) (Appendix S2).

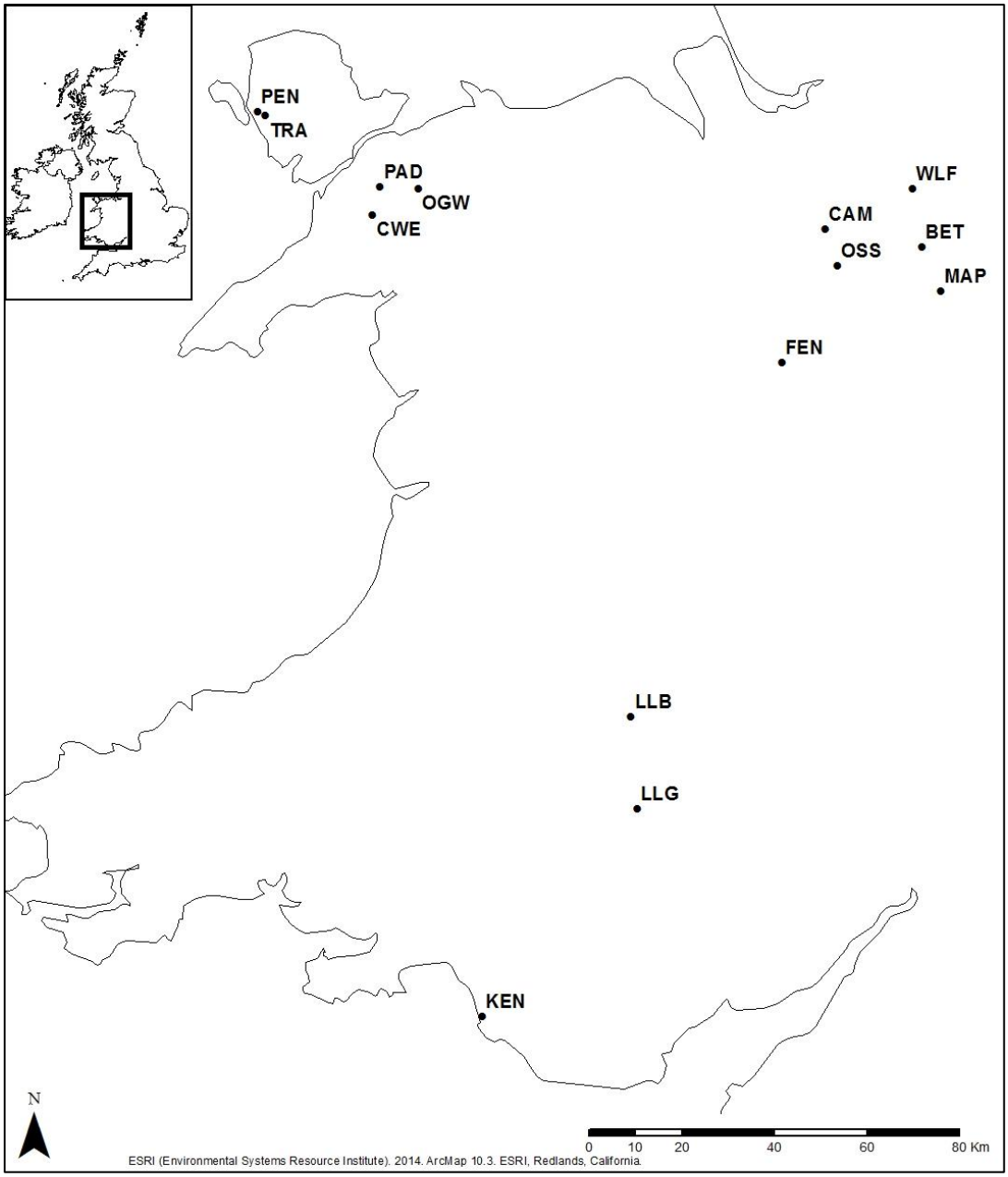

Fig. 1. Distribution of sampling lakes. Sampling lake codes: (CWE) Llyn Cwellyn; (PAD) Llyn Padarn; (OGW) Llyn Ogwen; (PEN) Llyn Penrhyn; (TRA) Llyn Traffwll; (KEN) Kenfig Pool; (LLB) Llan Bwch-llyn; (LLG) Llangorse Lake; (MAP) Maer Pool; (CAM) Chapel Mere; (OSS) Oss Mere; (FEN) Fenemere; (WLF) Watch Lane Flash; (BET) Betley Mere. The GPS coordinates of sampling locations in each lake are listed in Appendices S3 & S4.

eDNA capture, extraction, library preparation and sequencing

eDNA capture and extraction

In total, 252 samples were collected from the 14 lakes between 01/12/2015 to 12/07/2016 (Appendix S1 Table S1.2). The GPS coordinates of sampling locations in each lake are listed in Appendices S3 & S4. The detail of sampling strategy can be found in Appendix S5.1. All samples were filtered within 24 hrs of collection. Two litres of each water sample were filtered through a 47 mm diameter 0.45-µm mixed cellulose acetate and nitrate filter (Whatman) using Nalgene filtration units in combination with a vacuum pump (15–20 in. Hg; Pall Corporation). Our previous study demonstrated that the 0.45-µm filters are suitable for fish metabarcoding, with low variation and high repeatability between the filtration replicates (Li et al. 2018a) compared to other filtration methods. The filtration units were cleaned with 10% v/v commercial bleach solution (containing ~3% sodium hypochlorite) and 5% v/v microsol detergent (Anachem, UK), and then rinsed thoroughly with de-ionised water after each filtration run to prevent cross-contamination. For each sampling lake, one sampling blank and one filtration blank (2 L deionised water) were filtered before sample filtration in order to test for possible contamination at the sampling and filtration stages. After filtration, all filters were placed into 50 mm sterile petri dishes using sterile tweezers, sealed with parafilm and then immediately stored at –20 °C until extraction. DNA extraction was carried out using the PowerWater DNA Isolation Kit (Qiagen) following the manufacturer's protocol.

Library preparation and sequencing

Sequencing libraries were generated from PCR amplicons targeting two mitochondrial loci: Cytochrome b (Cytb) and 12S rRNA (12S). We previously tested both fragments *in vitro* on 22 UK freshwater fish species and *in situ* on three deep lakes in the English Lake District,

and demonstrated their suitability for eDNA metabarcoding of UK lake fish communities (Hänfling et al. 2016).

To enable the detection of possible PCR contamination, we included no-template controls (NTCs) of molecular grade water (Fisher Scientific) and single-template controls (STCs) of cichlid fish DNA (the Eastern happy, *Astatotilapia calliptera*, a cichlid from Lake Malawi, which is not present in natural waters in the UK) within each library (N = 307). Three PCR technical replicates were performed for each sample then pooled to minimise bias in individual PCRs. All PCRs were set up using eight-strip PCR tubes in a PCR workstation with UV hood and HEPA filter in the eDNA laboratory of University of Hull (UoH) to minimise the risk of contamination.

The Cytb locus, targeting a 414-bp vertebrate-specific fragment, was amplified using a one-step library preparation protocol (Kozich et al. 2013). The Cytb one-step library preparation protocol for this study was described in Appendix S5.2 and our previous study (Hänfling et al. 2016). The final 10 pM denatured Cytb library mixed with 30% PhiX genomic control was sequenced with the MiSeq reagent kit v3 (2×300 cycles) at the UoH. The 12S locus targets a 106-bp vertebrate-specific fragment (Riaz et al. 2011). Following the consistently lower sequencing yield of the one-step protocol for the 12S compared to Cytb in previous studies, we decided to switch to a two-step library preparation protocol using nested tagging (Kitson et al. 2018). The detail of the full two-step 12S library preparation can be found in Appendix S5.2. To improve clustering during initial sequencing, the denatured 12S library (13 pM) was mixed with 10% PhiX genomic control. The library was sequenced with the MiSeq reagent kit v2 (2 × 250 cycles) at the UoH.

Data analysis

Collation of fish data

Existing fish data for the 14 lakes was collated using a range of data sources (see the detail in Appendix S2). These included site-specific surveys carried out by Natural Resources Wales or the Environment Agency and their predecessor bodies, third party fishery surveys, data published in the literature and *ad hoc* records stored on databases such as the National Biodiversity Network Atlas (https://nbnatlas.org). Based on existence of survey data and expert opinion (TWH-E and GP), fish species were placed into four general categories with confidence of species presence/absence for each lake (Table 1), reflecting both the available data and the uncertainty around it. Assessing relative abundance *a priori* was more challenging. For the Welsh lakes, an abundance category approach has been used, with each species being placed on the DAFOR scale (D = Dominant; A = Abundant; F = Frequent; O = Occasional; R = Rare) (Tansley 1993) using available data where possible and/or TWH-E's expert opinion. All records where no abundance was recorded were assumed to be rare. For the Cheshire meres, summaries of fish community composition and fish density per unit area (ind. ha^{-1} and kg ha^{-1}) were produced from the Ecological Consultancy Limited (ECON) fishery survey using point abundance sampling by electro-fishing (PASE) and seine netting in the same period as water samples collection with this study in 2016 (Appendix S1 Tables S1.3 & S1.4). The ECON fishery survey in 2016 was undertaken by ECON and commissioned by Natural England as part of an investigation of the impacts of fisheries on Sites of Special Scientific Interests (SSSIs) condition.

Table 1. Criteria for assessing the confidence of species presence/absence using existing data

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

Bioinformatics analysis

Raw read data from Illumina MiSeq sequencing have been submitted to NCBI (BioProject: PRJNA454866). Bioinformatics analysis was implemented following a custom reproducible metabarcoding pipeline (*metaBEAT* v0.97.9) with custom-made reference databases (Cytb and 12S) as described in our previous study (Hänfling et al. 2016). Sequences for which the best BLAST hit had a bit score below 80 or had less than 95%/100% identity (Cytb/12S) identity to any sequence in the curated databases were considered non-target sequences. To assure full reproducibility of our bioinformatics analysis, the custom reference databases and the Jupyter notebooks for data processing have been deposited in an additional dedicated

GitHub repository (https://github.com/HullUni-bioinformatics/Li_et_al_2018_eDNA_fish_monitoring). The Jupyter notebook also performs demultiplexing of the indexed barcodes added in the first PCR reactions.

Low-frequency noise threshold

Filtered data were summarised into the number of sequence reads per species/sample for downstream analyses (Appendixes S3 & S4). After bioinformatics analysis, the low-frequency noise threshold (proportion of STC species read counts in the real sample) was set to 0.07% and 0.3% for Cytb and 12S respectively to filter high-quality annotated reads passing the previous filtering steps that have high-confidence BLAST matches but may result from contamination during the library construction process or sequencing (De Barba et al. 2014; Hänfling et al. 2016; Port et al. 2016). Filtered data were summarised in two ways for downstream analyses: (a) the number of sequence reads per species at each lake (hereafter referred to as read counts) and (b) the proportion of sampling locations in which a given species was detected (hereafter referred to as site occupancy).

Estimating abundance with eDNA

Based on our previous results (Hänfling et al. 2016), site occupancy is a better proxy for estimates of abundance than read counts. However, sequence reads should not be completely ignored because they still contain important abundance information. The maximum site occupancy across both loci was used as a score for species abundance (Table 2). In addition to site occupancy score, the maximum relative read count (i.e. proportion of read counts per species at each sampling lake) across both loci, and the number of loci with which the species was detected were used as confidence indicators to assign fish species into the same general

categories which were used for non-eDNA survey data (Table 3). Those species that were assigned the lowest confidence scores (probably absent) were excluded from downstream analysis (i.e. *Rutilus rutilus* in Llyn Cwellyn, *Abramis brama* and *Cottus gobio* in Llyn Ogwen, *Esox lucius* in Llyn Penrhyn, *Cottus gobio* in Llyn Traffwll, *Gobio gobio* in Betley Mere). Each retained species was assigned a corrected abundance score by multiplying site occupancy score × confidence score. The corrected abundance score was used to assign each species to a relative abundance DAFOR scale (Table 4).

Table 2. Site occupancy score based on maximum site occupancy across Cytb and 12S

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

Table 3. Criteria for assessing confidence of species occurrence using eDNA data

Category	**Confidence score**	**Description**		
		Locus	**Site occupancy score**	**Maximum relative read count (across Cytb and 12S)**
Probably absent (Pa)	Very low (-1)	Any	= 1	> 0 and ≤ 0.005
Possibly present (Po)	Low (1)	Any	> 1	> 0 and ≤ 0.005
		One	≥ 1	> 0.005 and ≤ 0.05
Probably present (Pr)	Moderate (3)	Both	≥ 1	> 0.005 and ≤ 0.01
		Both	= 1	> 0.01 and ≤ 0.1
		One	≥ 1	> 0.05 and ≤ 0.1

		One	= 1	> 0.1 and ≤ 0.5
Established (E)	High (5)	Both	> 1	> 0.01
		Both	= 1	> 0.1
		One	> 1	> 0.1
		One	= 1	> 0.5

Table 4. The relative abundance DAFOR scale based on eDNA data

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

Ecological and statistical analyses

All downstream analyses were performed in R v3.3.2 (R_Core_Team 2016) and graphs plotted using *ggplot2* v2.2.1 (Wickham & Chang 2016). Before investigating species detection and abundance estimate with eDNA, we first evaluated whether Cytb and 12S datasets produced consistent results by calculating the Pearson's product-moment correlation coefficient for both read counts and site occupancy. Spearman's rank correlations were performed between historical data and eDNA data – firstly in terms of confidence of species presence and secondly in terms of relative abundance. To investigate differences in fish communities between sampling lakes, hierarchical clustering dendrograms was used to assess

the existence of distinct community types using the function *fviz_dend* in *factoextra* v1.0.4 (Kassambara & Mundt 2017) to extract and visualize the results calculated from the function *hclust* using Canberra distances method based on site occupancy. Furthermore, non-metric multidimensional scaling (NMDS) in individual sampling location of lakes, allied with analysis of similarities (ANOSIM), were performed using the abundance-based Bray-Curtis dissimilarity index with the function *metaMDS* and *anosim* respectively in *Vegan* v2.4-4 (Oksanen et al. 2017). The full R script is available on the GitHub repository (https://github.com/HullUni-bioinformatics/Li_et_al_2018_eDNA_fish_monitoring/tree/master/R_script)

Results

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

Significant correlations were found between site occupancy and average read counts, for both loci and all sampling lakes apart from Llyn Ogwen (Pearson's $r = 0.81$, $df = 4$, $p = 0.05$) and Llan Bwch-llyn (Pearson's $r = 0.94$, $df = 2$, $p = 0.06$) for Cytb (Appendix S1 Fig. S1.1a, b). Data from the two loci were significantly correlated (Pearson's r consistently $p < 0.05$) for site occupancy for every lake (Appendix S1 Fig. S1.1c), and for average read counts apart from in three lakes: Llyn Padarn (Pearson's $r = 0.55$, $df = 7$, $p = 0.12$), Llyn Traffwll (Pearson's $r = 0.59$, $df = 5$, $p = 0.16$) and Llan Bwch-llyn (Pearson's $r = 0.89$, $df = 2$, $p = 0.11$) (Appendix S1 Fig. S1.1d).

Comparison of eDNA data and expected results

In total, 24 species with 111 species occurrences by lake were detected across Cytb and 12S. Perch *Perca fluviatilis* was the most common species which were detected in all lakes. The second most frequently occurring species was roach *Rutilus rutilus* with detections in 10 lakes. In addition to these two species, pike *Esox lucius* were found in all nine coarse fish lakes (Table 5; Fig. 2). There was a significant positive correlation between expected faunas and eDNA data in terms of confidence of species occurrence by lake (Spearman's $r = 0.74$, $df = 109$, $p < 0.001$). A total of 73 species occurrences have been recorded across all lakes according to historical data. Of these, 72 (98.6%) were detected with eDNA (Table 5). The only false negative in the eDNA datasets was tench *Tinca tinca* in Llyn Penrhyn. Tench were detected by 12S in Llyn Penrhyn but at a read count less than the low-frequency noise threshold (Appendix S4; site occupancy = 0.3, average read counts = 3.9). Tench are likely to be very rare in Llyn Penrhyn: only two tench were recorded by the 2016 Royal Society for the Protection of Birds survey and small numbers have also been detected in previous years (Appendix S2.4).

There were consistent positive correlations between DAFOR scale based on eDNA data and expected DAFOR scale based on both historical data for Welsh lakes and fish density per unit area from the ECON survey of Cheshire meres (Fig. 3). The correlations were significant in three out of eight Welsh lakes (Fig. 3). Significant correlations were observed in three and four Cheshire meres based on individual density (ind. ha^{-1}) and biomass density (kg ha^{-1}), respectively (Fig. 3; Appendix S1 Fig. S1.3).

Characterisation of fish assemblages using eDNA data

The hierarchical clustering dendrograms based on site occupancy for two loci indicated that there were four distinct community types compared to three pre-defined lake types according to environmental characteristics. Specifically, the clustering dendrograms accorded with the pre-defined lake Type 1 (Llyn Cwellyn, Llyn Ogwen and Llyn Padarn) and Type 2 (Llyn Traffwll and Llyn Penrhyn), but indicated that Watch Lane Flash and Betley Mere can be divided from the pre-defined lake Type 3 (Fig. 4a1, b1). The NMDS ordination allied with ANOSIM based on read counts for two loci confirmed the clustering dendrograms results that there were four distinct community types according to the predominant groups of fish (Fig. 4a2, b2). These four identified community types were: Community 1: salmonids and minnow *Phoxinus phoxinus*; Community 2: mixed diadromous fish; Community 3: coarse fish and Community 4: bream-dominated coarse fish (Fig. 4). The *R* statistic in the ANOSIM global test with these four different communities were high (Appendix S1 Table S1.5, 0.75 ± 0.02) supporting statistical differences between communities (Appendix S1 Table S1.5, $p = 0.001$). The overall distance pattern of these four fish communities was that both coarse fish communities (Community 3 and Community 4) were close to each other, the mixed diadromous fish community (Community 2) was between the coarse fish communities and the salmonids and minnow community (Community 1) (Fig. 4; Appendix S1 Table S1.5).

The salmonids and minnow community (Community 1) is characteristic of three low alkalinity upland lakes (Llyn Cwellyn, Llyn Ogwen and Llyn Padarn). These sites were generally dominated by brown trout *Salmo trutta*, rainbow trout *Oncorhynchus mykiss* or Arctic charr *Salvelinus alpinus*, but also included minnow and Atlantic salmon *Salmo salar* (Fig. 2; Appendix S1 Fig. S1.2). Llyn Cwellyn and Llyn Padarn are designated as SSSIs under the Wildlife and Countryside Act 1981 (as amended) to conserve Arctic charr, and the oligotrophic lake habitat of Llyn Cwellyn is also protected as a Special Areas of Conservation

under the EU Habitats Directive. The eDNA data showed that Llyn Padarn contained a small number of Arctic charr, whereas this species was dominant in Llyn Cwellyn (Fig. 2; Appendix S1 Fig. S1.2). Rainbow trout were only detected in Llyn Ogwen, where they are regularly stocked by the local angling club (Appendix S2.3). Some of these sites may also contain low density of other species; usually diadromous species were detected using eDNA as well such as European eel *Anguilla anguilla* in all three upland lakes and three-spined stickleback *Gasterosteus aculeatus* in Llyn Padarn. Perch have recently been recorded from Llyn Padarn for the first time using captured-based survey methods (Appendix S2.2). Surprisingly, this species was detected using eDNA (though shown to be rare), in all of these three lakes (Fig. 2; Table 5). The diadromous fish community (Community 2) is characteristic of Llyn Traffwll and Llyn Penrhyn which were dominated by European eel and three-spined stickleback. Other species included brown trout, nine-spined stickleback *Pungitius pungitius*, roach, perch and rudd *Scardinius erythrophthalmus* (Fig. 2). The coarse fish communities consisted of bream *Abramis brama*, common carp *Cyprinus carpio*, pike, perch, roach, rudd and tench, plus some additional species (e.g. European eel, bullhead *Cottus gobio*, three-spined stickleback and gudgeon *Gobio gobio*). Pike, perch, roach and tench were dominant in Kenfig Pool, Maer Pool, Chapel Mere, Llan Bwch-llyn, Llangorse Lake, Oss Mere and Fenemere (Community 3); however, the dominant species in Watch Lane Flash and Betley Mere (Community 4) was bream (Fig. 2; Appendix S1 Fig. S1.2).

Table 5. Correspondence of confidence of species occurrence by lake between predicted and eDNA data

Community type			**1**		**2**		**3**							**4**	
Scientific name	**Code**	**CWE**	**PAD**	**OGW**	**PEN**	**TRA**	**KEN**	**LLB**	**LLG**	**MAP**	**CAM**	**OSS**	**FEN**	**WLF**	**BET**
Abramis brama	BRE							Pr/E	E/E	Po/Po	Po/Po	E/E	E/E	E/E	E/E
Alburnus alburnus	BLE			Pa/Po			Pa/Po								
Anguilla anguilla	EEL	E/Pr	E/E	Po/Pr	E/E	E/E	E/E		E/E				E/E	Po/Po	
Carassius auratus	GOF		Pa/Pr												
Carassius carassius	CRU														Po/Pr
Cottus gobio	BUL						Pa/Po		Po/Pr						Po/Po
Cyprinus carpio	CAR									E/E	Po/Po	E/E	E/E	E/E	E/E
Esox lucius	PIK						E/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E
Gasterosteus aculeatus	3SS		E/E		E/E	E/E					Po/E		E/Pr	Po/E	Po/Pr
Gobio gobio	GUD		Pa/Po									Po/Po		E/Po	
Leucaspius deliniatus	SUN													E/E	
Oncorhynchus mykiss	RTR			E/E											
Perca fluviatilis	PER	Pa/E	Pr/Po	Pa/Pr	E/E	Pa/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E	E/E
Phoxinus phoxinus	MIN	E/E	E/E	E/E								Pa/Po			
Pseudorasbora parva	TMG														Pa/Po
Pungitius pungitius	9SS				Po/Pr	Po/Po	Po/Po								
Rhodeus amarus	BIT													Po/Po	
Rutilus rutilus	ROA				E/E	E/E	Po/Po	E/E	E/E	Pa/Po		E/E	E/E	E/E	E/E
Salmo salar	SAL	E/Po	E/E												
Salmo trutta	BTR	E/E	E/E	E/E	Po/Po	Po/E									Po/Po

Salvelinus alpinus	CHA	E/E	E/Pr										
Scardinius erythrophthalmus	RUD			E/E	Po/Pr	E/E	E/E		Pa/Po	E/E	E/E		
Squalius cephalus	CHU											Pa/Po	
Tinca tinca	TEN			E/Pa		E/E	E/E	E/E	E/E	E/E	Po/Po	Po/Pr	E/E

Notes: Categories of presence confidence are described in Table 1 showing as predicted/eDNA data. Solid vertical lines indicate community types with similar faunas. Dashed line indicates division between whether bream *Abramis brama* is dominant species in the lake. Sampling lake codes are given in Fig. 1.

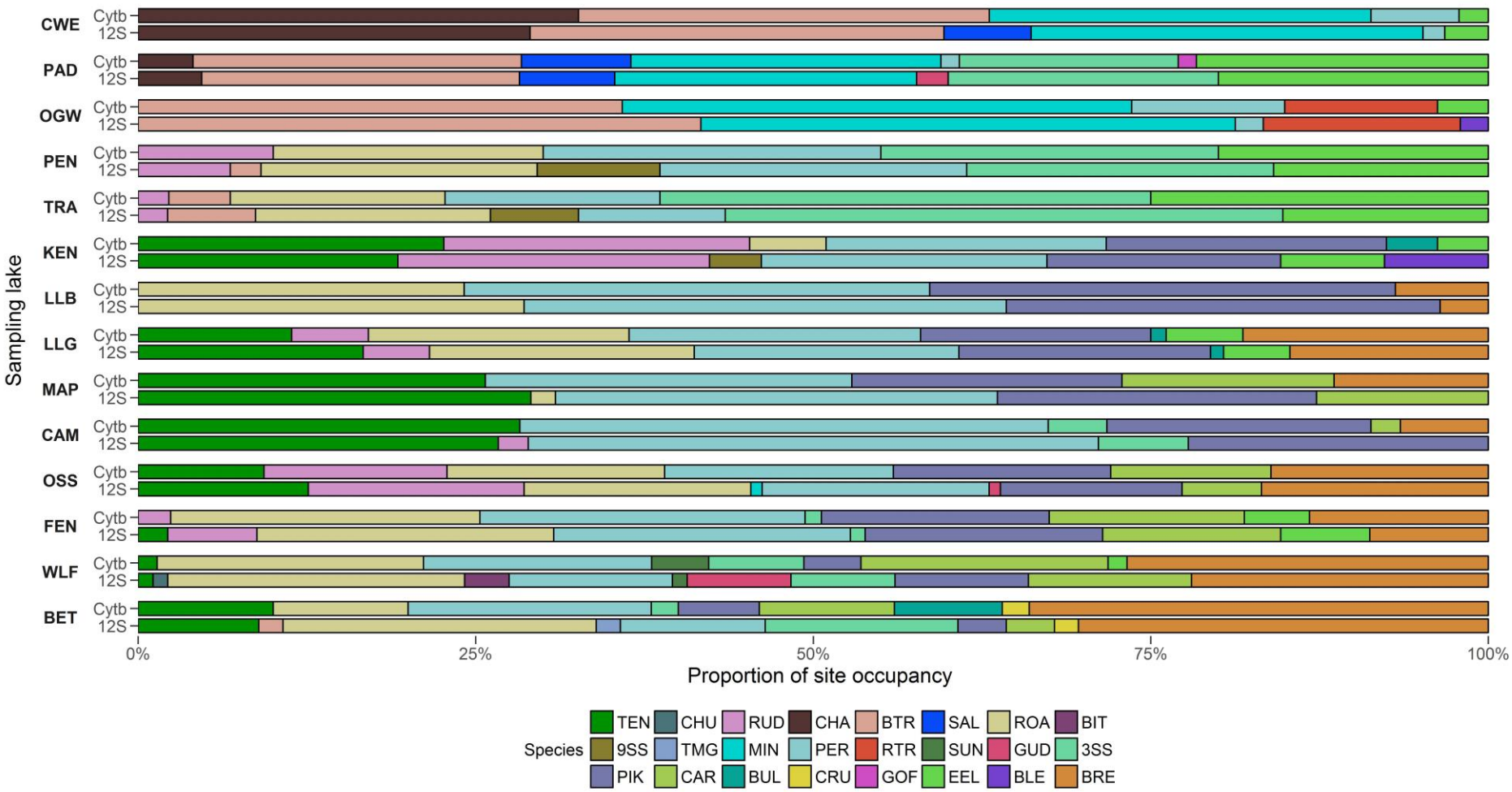

Fig. 2. Species composition of site occupancy for Cytb and 12S. Species three letter codes and sampling lake codes are given in Table 5 and Fig. 1, respectively.

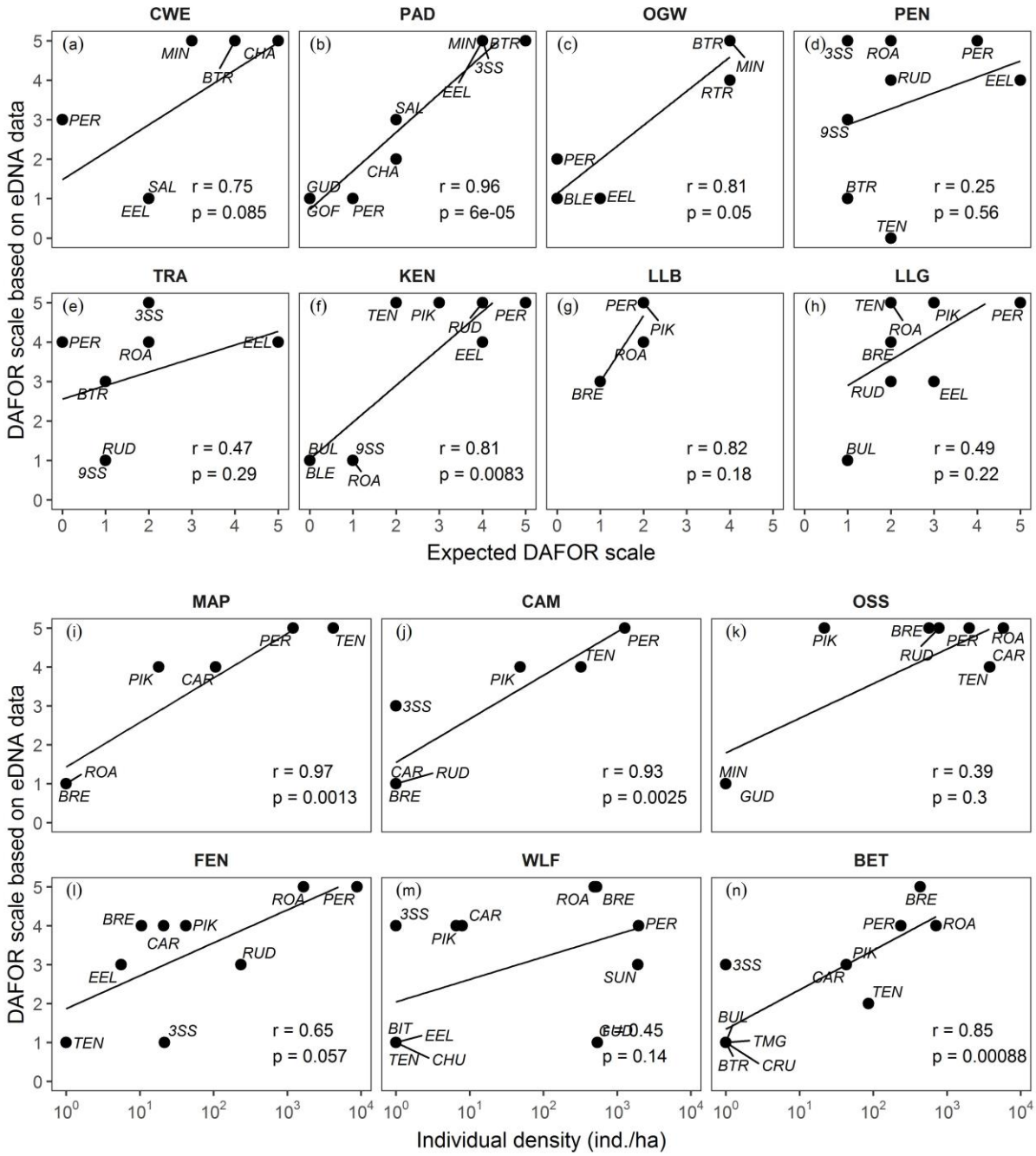

Fig. 3. Correlations between DAFOR scale based on eDNA data and expected DAFOR scale based on existing data of Welsh lakes or individual density from the ECON survey of Cheshire meres. Sampling lake codes are given in Fig. 1; Corrected abundance scores to DAFOR scale and species three letter codes are given in Table 4 and Table 5, respectively.

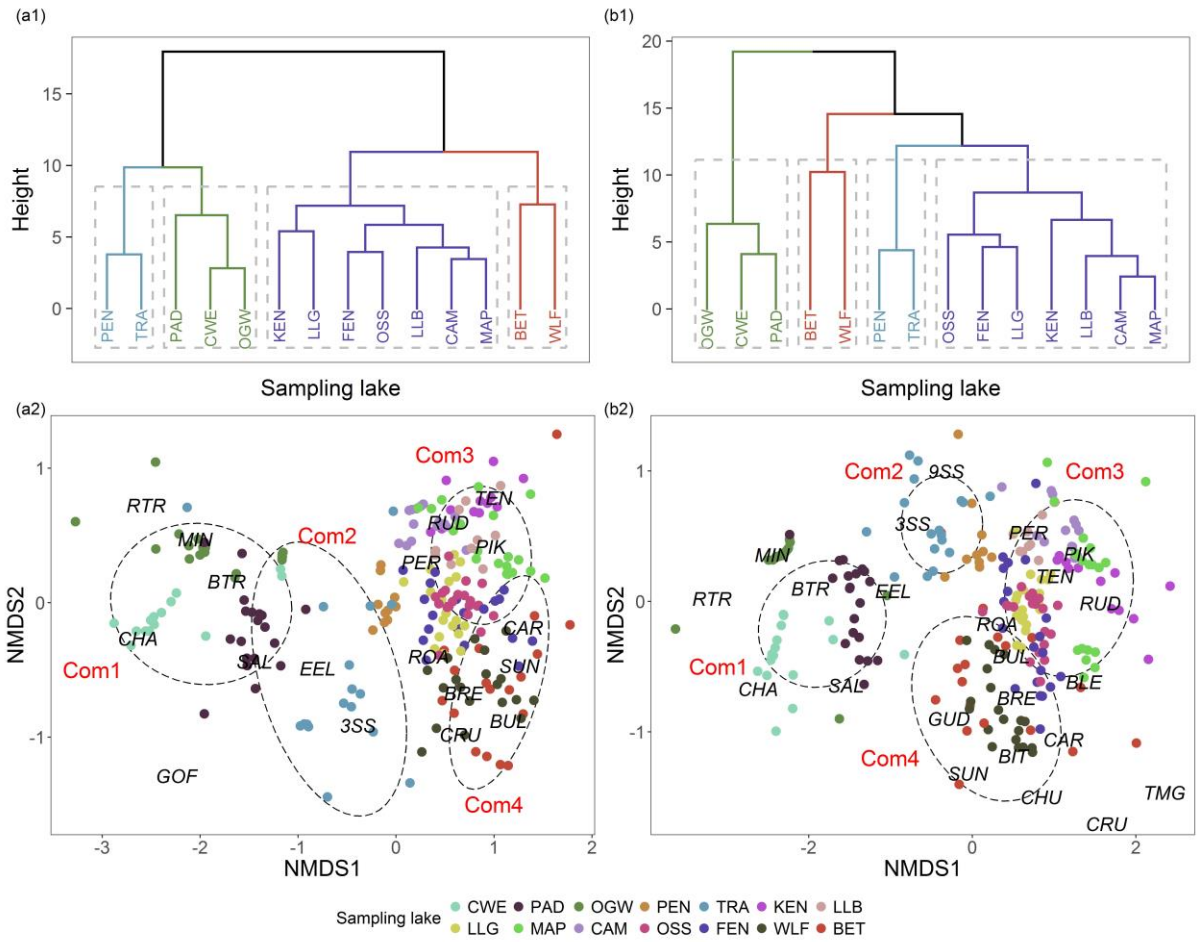

Fig. 4. Hierarchical clustering dendrograms and non-metric multidimensional scaling (NMDS) ordination of the sampling lakes. Dendrograms using Canberra distances method based on site occupancy for (a1) Cytb and (b1) 12S. NMDS in individual sampling location of lakes using Bray-Curtis dissimilarity index based on read counts for (a2) Cytb and (b2) 12S. The dashed frames are drawn around each cluster in dendrograms. The ellipse indicates 70% similarity level within each community type in ordinations. Scores for species taxa are plotted on the same axes to better visualize the ordination in space between species and samples. Species three letter codes and community type (Com 1–4) with individual lake codes are given in Table 5 and Fig. 1, respectively.

Discussion

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

Comparison of eDNA and existing data for species detection

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

Moreover, there was a significant positive correlation between historical data and eDNA data in terms of confidence of species presence. Species occurrences that were assessed as

“probably or possibly present” (25/111) and “probably absent” (13/111) based on distribution data or anecdotal evidence also fitted the eDNA criteria for lowered confidence levels. In most cases, the “probably absent” eDNA occurrences were at very low site occupancy (≤ 0.1) and read counts (~0.5%), just above our threshold for accepting a positive record. These records could be genuine detections of species that have previously been missed. For example, perch were recorded with both Cytb and 12S in Llyn Cwellyn, Llyn Ogwen and Llyn Traffwll. The confidence of species present was either “established present” or “probably present” according to eDNA data. The recent appearance of this species in nearby lakes Llyn Padarn and Llyn Penrhyn (see detail in Appendix S2) suggest that the species is spreading in North Wales either by natural means or as a result of illegal introductions. In addition, the other “probably absent” eDNA occurrences could be either false positives from sequencing error, laboratory or environmental contamination. For instance, bleak *Alburnus alburnus* in Llyn Ogwen, goldfish *Carassius auratus* and Gudgeon *Gobio gobio* in Llyn Padarn, rudd in Chapel Mere, roach in Maer Pool, topmouth gudgeon *Pseudorasbora parva* in Betley Mere and chub *Squalius cephalus* in Watch Lane Flash are most likely explained by low-level cross contamination or sequencing barcode misassignment since these species were only detected by either Cytb or 12S in single sample of the lake.

Barcode misassignment and tag jumps have been proved in metabarcoding studies (e.g. Deakin et al. 2014; Schnell et al. 2015). To minimise contamination or false assignments, further work needs to be done to optimise each step, as well as the incorporation of robust controls to identify contamination when it occurs. Together with a growing number of studies (Stat et al. 2017; Li et al. 2018b), the results demonstrate the importance of using two (or more) markers for metabarcoding to avoid problems due to primer bias (Clarke et al. 2014), gene copy number, PCR or sequencing artefacts (Schloss et al. 2011) and/or contamination. Other strategies such as using consistency of presence across technical replicates as used by

Port et al. (2016) might be a more suitable approach to control for false positives if rare species are of particular interest. Where verification of rare species detections is considered a high priority, additional confirmation from targeted species approaches (e.g. quantitative PCR) and/or field surveys may be required.

Use of eDNA for assessing relative abundance of fish

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

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

also important to investigate taxon-specific detection probabilities and abundance estimates so that a pressure-sensitive tool can be developed.

Using eDNA to describe ecological communities

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

Specifically, our results showed that the low alkalinity lakes within a predominantly upland catchment were mainly dominated by salmonids reflecting their relatively deep and oligotrophic nature. Except from salmonids, all these sites contained minnow, most likely introduced by anglers using them as live bait (Hatton-Ellis 2005). Compared to the salmonids and minnow lakes (Community 1), perch, pike and cyprinids (such as bream, common carp, roach and tench) were prevalent in the coarse fish lakes (Communities 3 & 4) reflecting their relatively shallow and eutrophic nature. These findings support that eDNA profiles are suitable to reflect the eutrophic conditions of Lake Windermere in which species that prefer less eutrophic conditions (Atlantic salmon, brown trout, Arctic charr, minnow and bullhead) were more abundant in the mesotrophic North Basin, and species associated with eutrophic conditions (roach, tench, rudd, bream and European eel) were more common in the eutrophic South Basin (Hänfling et al. 2016). Furthermore, both of the diadromous fish lakes (Community 2) contained a varied fauna including various combinations of mostly

diadromous species such as European eel, three-spined stickleback, nine-spined stickleback and brown trout. Alongside these, several coarse fish species such as roach, perch and rudd also occurred in these lakes. However, of greater conservation interest is the distribution of nine-spined stickleback. This species occurs only locally in Wales, mainly at lowland sites close to the sea (Davies et al. 2004; Hatton-Ellis 2005).

Authors' contributions

BH, TWH-E, and GP conceived and designed the study; TWH-E and GP prepared historical fish data; HSK, MB, and JL carried out the fieldwork; MB prepared the sampling map; JL conducted laboratory work and performed bioinformatics analyses; JL, LLH, and BH performed the statistical analyses; JL wrote the manuscript; all authors commented the final manuscript.

Acknowledgments

This work was part of PhD project of JL, who was supported by University of Hull and China Scholarship Council. This work was funded by UK Environment Agency and Natural Resources Wales. We are particularly grateful to all the site owners and managers for granting access and the Bangor University for providing access to their laboratory facilities. Ecological Consultancy Limited provided fishery survey data and cooperation in data collation of Cheshire meres. Robert Jaques, Hayley Watson, Mags Cousins, Peter Clabburn, Rob Evans, Sophie Gott, Emma Keenan, Ian Sims, Emma Brown, Jason Jones, Dawn Parry, Peter Shum, Harriet Johnson, Rob Donnelly and Christoph Hahn provided invaluable help during fieldwork and lab work.

Data accessibility

Data is available from the GitHub repository (https://github.com/HullUni-bioinformatics/Li_et_al_2018_eDNA_fish_monitoring). The repository will be archived with Zenodo after the manuscript accepted.

References

Argillier, C., Causse, S., Gevrey, M., Pedron, S., De Bortoli, J., Brucet, S., ... Holmgren, K. (2013) Development of a fish-based index to assess the eutrophication status of European lakes. *Hydrobiologia,* **704,** 193-211. https://doi.org/10.1007/s10750-012-1282-y

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

Clarke, L.J., Soubrier, J., Weyrich, L.S. & Cooper, A. (2014) Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources,* **14,** 1160-1170. https://doi.org/10.1111/1755-0998.12265

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

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E. & Taberlet, P. (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources,* **14,** 306-323. https://doi.org/10.1111/1755-0998.12188

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

Hänfling, B., Lawson Handley, L., Read, D.S., Hahn, C., Li, J., Nichols, P., ... Winfield, I.J. (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology,* **25,** 3101-3119. https://doi.org/10.1111/mec.13660

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

Hering, D., Borja, A., Jones, J.I., Pont, D., Boets, P., Bouchez, A., ... Kelly, M. (2018) Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research,* **138,** 192-205. https://doi.org/10.1016/j.watres.2018.03.003

Hughes, M., Hornby, D.D., Bennion, H., Kernan, M., Hilton, J., Phillips, G. & Thomas, R. (2004) The development of a GIS-based inventory of standing waters in Great Britain together with a risk-based prioritisation protocol. *Water, Air and Soil Pollution: Focus,* **4,** 73-84. https://doi.org/10.1023/b:wafo.0000028346.27904.83

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

Kelly, F.L., Harrison, A.J., Allen, M., Connor, L. & Rosell, R. (2012) Development and application of an ecological classification tool for fish in lakes in Ireland. *Ecological Indicators,* **18,** 608-619. https://doi.org/10.1016/j.ecolind.2012.01.028

Kitson, J.J.N., Hahn, C., Sands, R.J., Straw, N.A., Evans, D.M. & Lunt, D.H. (2018) Detecting host-parasitoid interactions in an invasive Lepidopteran using nested tagging DNA-metabarcoding. *Molecular Ecology,* **n/a**. https://doi.org/10.1111/mec.14518

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K. & Schloss, P.D. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology,* **79,** 5112-5120. https://doi.org/10.1128/AEM.01043-13

Kubecka, J., Hohausova, E., Matena, J., Peterka, J., Amarasinghe, U.S., Bonar, S.A., ... Winfield, I.J. (2009) The true picture of a lake or reservoir fish stock: A review of needs and progress. *Fisheries Research,* **96,** 1-5. https://doi.org/10.1016/j.fishres.2008.09.021

Lawson Handley, L. (2015) How will the 'molecular revolution' contribute to biological recording? *Biological Journal of the Linnean Society,* **115,** 750-766. https://doi.org/10.1111/bij.12516

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

Li, Y., Evans, N.T., Renshaw, M.A., Jerde, C.L., Olds, B.P., Shogren, A.J., ... Pfrender, M.E. (2018b) Estimating fish alpha- and beta-diversity along a small stream with environmental DNA metabarcoding. *Metabarcoding and Metagenomics,* **2,** e24262. https://doi.org/10.3897/mbmg.2.24262

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

Pall, K. & Moser, V. (2009) Austrian Index Macrophytes (AIM-Module 1) for lakes: a Water Framework Directive compliant assessment system for lakes using aquatic macrophytes. *Hydrobiologia,* **633,** 83-104. https://doi.org/10.1007/s10750-009-9871-0

Port, J.A., O'Donnell, J.L., Romero-Maraccini, O.C., Leary, P.R., Litvin, S.Y., Nickols, K.J., ... Kelly, R.P. (2016) Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology,* **25,** 527-541. https://doi.org/10.1111/mec.13481

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

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P. & Coissac, E. (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research,* **39,** e145. https://doi.org/10.1093/nar/gkr732

Scheffer, M., Carpenter, S., Foley, J.A., Folke, C. & Walker, B. (2001) Catastrophic shifts in ecosystems. *Nature,* **413,** 591-596. https://doi.org/10.1038/35098000

Schloss, P.D., Gevers, D. & Westcott, S.L. (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS ONE,* **6,** e27310. https://doi.org/10.1371/journal.pone.0027310

Schnell, I.B., Bohmann, K. & Gilbert, M.T.P. (2015) Tag jumps illuminated – reducing sequence - to - sample misidentifications in metabarcoding studies. *Molecular Ecology Resources,* **15,** 1289-1303. https://doi.org/10.1111/1755-0998.12402

Stat, M., Huggett, M.J., Bernasconi, R., DiBattista, J.D., Berry, T.E., Newman, S.J., ... Bunce, M. (2017) Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports,* **7,** 12240. https://doi.org/10.1038/s41598-017-12501-5

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

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

Thomsen, P.F. & Willerslev, E. (2015) Environmental DNA–an emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation,* **183,** 4-18. https://doi.org/10.1016/j.biocon.2014.11.019

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P.F., ... Dejean, T. (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology,* **25,** 929-942. https://doi.org/10.1111/mec.13428

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

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

Winfield, I.J., Fletcher, J.M., James, J.B. & Bean, C.W. (2009) Assessment of fish populations in still waters using hydroacoustics and survey gill netting: experiences with Arctic charr (*Salvelinus alpinus*) in the UK. *Fisheries Research,* **96,** 30-38. https://doi.org/10.1016/j.fishres.2008.09.013