Natural England Commissioned Report NECR325

# Meta-analysis of eDNA metabarcoding data from UK lakes: optimising species detection probability and sampling effort

First published November 2020

www.gov.uk/natural-england

# Foreword

Natural England commission a range of reports from external contractors to provide evidence and advice to assist us in delivering our duties. The views in this report are those of the authors and do not necessarily represent those of Natural England.

## Background

DNA – based methods offer a significant opportunity to change how we monitor and assess biodiversity. These techniques may provide cheaper alternatives to existing species monitoring or an ability to detect species that we cannot currently detect reliably.

However, for most species, there is still much development required before they can be used in routine monitoring. Natural England has been exploring the further use of these methods for environmental monitoring for several years, delivering a series of reports which focus on the development of DNA-based methods with potential in a particular area.

This report focusses on the development of methods for environmental DNA (eDNA) based monitoring of lake fish communities. New protocols using eDNA have recently been developed and deployed in the UK.

This study builds on this development by exploring the effect of reducing the number of samples taken within a lake and understanding what the effects of this are for biodiversity information.

A secondary goal was to explore if mammalian eDNA also identified from the lake could also contribute to biodiversity data for these species.

This report should be cited as: Sellers., G.S. and Hänfling, B. (2020) Meta-analysis of eDNA metabarcoding data from UK lakes: optimising species detection probability and sampling effort. *Natural England Commissioned Report Number 325.*

**Natural England Project Manager** – Dave Ottewell, Senior Hydrologist, Natural England: dave.ottewell@naturalengland.org.uk

**Contractor** – EvoHull, The University of Hull

**Keywords** – environmental DNA (eDNA), Lakes, metabarcoding, fish

**Further information**
This report can be downloaded from the Natural England Access to Evidence Catalogue: **http://publications.naturalengland.org.uk/**. For information on Natural England publications contact the Natural England Enquiry Service on 0300 060 3900 or e-mail **enquiries@naturalengland.org.uk**.



ISBN 978-1-78354-672-5

# Meta-analysis of eDNA metabarcoding data from UK lakes: optimising species detection probability and sampling effort

**Graham S Sellers[1], Bernd Hänfling[1]**

1. Evolutionary Biology Group, Biological and Marine Sciences, University of Hull

Corresponding author: Bernd Hänfling (b.haenfling@hull.ac.uk)

# Contents

# Acknowledgments

The Agencies that own the data are **Environment Agency, Natural Resources Wales** and **Scottish Environment Protection Agency**. In addition **Scottish Natural Heritage** have provided advice and **Natural England** have funded the project. A number of scientists from the **University of Hull**, the **University of Stirling** and the **Centre for Ecology & Hydrology** have contributed to the collection of the data. The individuals to be acknowledged are:

**Environment Agency:**
Dr. Graeme Peirson, Dr. Kerry Walsh, Keith Bell

**Natural Resources Wales:**
Dr. Tristan Hatton-Ellis

**Scottish Environment Protection Agency:**
Dr. Alistair Duguid, Dr. Willie Duncan, Sean Morrison

**Scottish Natural Heritage:**
Prof. Colin Bean

**Natural England:**
Dr. Ruth Hall, Chris Mainstone, Dr. Katie Clark, Debbie Leatherland, Dave Ottewell

**University of Hull:**
Dr Lori Lawson-Handley, Dr Jianlong Li, Dr Rosetta Blackman, Dr Harriet Johnson, Dr Rob Donnelly, Dr Lynsey Harper, Dr Marco Benucci, Dr Helen Kimbell, Cristina Di Muri, Nathan Griffith, Dr Christoph Hahn, Dr David Lunt, Dr Graham Sellers, Dr Bernd Hänfling

**University of Stirling**:
Prof Nigel Willby, Dr Alan Law, Dr Colin Bull

**Centre for Ecology & Hydrology**
Dr Ian Winfield, Dr Daniel Read, Dr Anna Oliver, Ben James, Janice Fletcher

Without exceptions land owners and fishery owners have supported the data collection through allowing access to the water bodies. A number of EA and SEPA officers helped during planning and evaluated the eDNA data against previous knowledge of fish distribution.

# Summary

Methods for environmental DNA (eDNA) based monitoring of lake fish communities have recently been developed and deployed in the UK. This approach combines eDNA with modern High-Throughput-Sequencing technology, so-called eDNA metabarcoding. The UK Environment Agencies have funded the collection of an eDNA meta-barcoding data set of vertebrates from 101 UK lakes covering a broad spectrum of environmental conditions. This dataset is based on analysing 20 water samples per lake and has successfully been used to develop a WFD compatible water quality assessment tool. Previous analyses on a subset of lakes have indicated that the number of samples is more than sufficient to reliably estimate species richness of lakes, but it is unclear how exactly reduced sampling effort affects other biodiversity estimates and inferences made about water quality. As the number of samples determines the cost of monitoring programmes it is essential that the sampling effort is optimised for a specific monitoring objective. Other biodiversity elements contained in the dataset such as mammals have so far not been explored. The objectives of this project were therefore twofold; first to explore the effect a reduced sampling effort would have on various biological inferences using algorithmic and statistical resampling techniques. A secondary goal was to extract information on the distribution of mammalian eDNA and to explore whether meaningful distribution patterns of key species could be inferred. The results showed that reliable estimation of lake species richness could be achieved already with a much lower number of samples. For example, almost 90% of lakes achieved a sample coverage of 95% with only 10 samples. However, rare species are more often missed with fewer samples, with implications for monitoring programs of invasive or endangered species. Estimates of community composition and the ecological quality ratio (EQR) responded slowly to decreasing sampling effort. For example subsets of 10 samples were in most cases much more closely related to each other than to sample sets from other lakes and showed very similar Ecological Quality Ratios. These results are able to inform the design of eDNA sampling strategies, so that these can be optimised to achieve specific monitoring goals. The analysis of mammalian eDNA revealed a total of 41 mammal species across the global dataset including a wide range of domestic animals, and terrestrial, semiaquatic and flying wild mammals. The detection probability was highest for domestic species such as cattle and sheep which were detected in the majority of lakes and samples within lakes. Most wild mammals were detected with a significantly lower probability within lakes and in a lower number of lakes. The most commonly detected among terrestrial wild mammals were various species of mice, voles and deer, and among semiaquatic animals otter and water shrew. An analysis of the spatial distribution of these eDNA records from some key species showed a good overlap with the known range of these species, but did not reflect the expected density of distribution within these ranges. This indicates that monitoring of aquatic eDNA could complement the monitoring of UK mammals but a high sampling effort would be required to reliably infer the absence of a species in a certain region.

# List of Tables

# List of Figures

# 1 Background and objectives

1.1 Environmental DNA (eDNA) metabarcoding of lake water has recently been used for the detection and monitoring of fish species and community structure. It is a non-invasive method proven to be more effective at detecting elusive species than established invasive surveying techniques such as electro fishing or fyke netting (Hänfling et al. 2016a, Lawson Handley et al. 2019, Li et al. 2019). Terrestrial and semi-aquatic species can also be detected from lake water eDNA metabarcoding (Harper et al. 2019, Sales et al. 2020).

1.2 In 2014 the UK Environment Agencies (UK-EAs) initiated a research programme to evaluate the suitability of eDNA metabarcoding approaches for monitoring lake fish communities largely with the objective to develop a tool which is compatible with requirements under the European Union Water Framework Directive (WFD), thereafter referred to as the "Project". The research output of the original pilot project was published in 2016 (Hänfling et al. 2016a), with subsequent development of the method published in Li et al. (2018), Sellers et al. (2018) and Lawson Handley et al. (2019). The findings of the pilot project demonstrated that 20 water samples were more than sufficient to detect the majority of fish species from England's largest lake, Windermere, and sufficient to provide meaningful semi-quantitative abundance estimates. The results further indicated that the efficiency of the approach could be optimised by collecting samples from the shoreline during the winter season. Using this approach additional data were collected between 2016 and 2019 (Li et al. 2019; Hänfling et al. 2020) resulting in a data set of 101 lakes. This data set was used to demonstrate that the eDNA metabarcoding data can be used to classify the ecological status of UK lakes (Willby et al. 2019).

1.3 DNA is not homogeneously distributed in lentic environments (Hänfling et al. 2016a) and hence the detection of species relies on the collection of an adequate number of samples from a water body to capture the eDNA signal. The precautious approach applied during data collection of the "Project" was based on 20 samples per lake. It is however likely that fewer samples yield sufficient data, especially if rare species are not the focus, but this issue has so far not been explored sufficiently. As the number of samples determines the cost of monitoring programmes it is essential that the sampling effort is optimised for a specific monitoring objective.

1.4 Recent studies have demonstrated that the detection and monitoring of mammal species is also possible from aquatic eDNA samples, yet detection probability is low depending on population densities (Harper et al. 2019, Sales et al. 2020). However, the reliability and comprehensiveness of this method needs to be extensively tested in comparison to existing records.

1.5 The objective of this study was to carry out a meta-analysis of the 101 lakes data used in (Willby et al. 2019) to

a) further explore the effect of sample number on estimation of lake biodiversity metrics such as species richness, community composition and eDNA base EQR estimation using random and non-random data resampling techniques.

b) determine detection probability and distribution of mammals across the UK based on the presence/absence of eDNA signals from lake water. Specifically focussing on four key species: Eurasian beaver (*Castor fiber*), Eurasian otter (*Lutra lutra*), Eurasian red squirrel (*Sciurus vulgaris*) and the European pine marten (*Martes martes*), we compared the reliability of their positive eDNA signal distribution in relation to existing records.

# 2 Methods

## The data sets

### Study lakes, sample collection and DNA sequencing

2.1 We utilised eDNA metabarcoding data from 101 lakes which were generated during various project phases of the Project (**Figure 1**). The water samples were collected between January 2015, and March 2019 largely during the winter season (November - March). A consistent approach was used for sample collection and filtration across different project phases as described in Hänfling et al. (2016b; 2016c). Each individual sample contained 2l of surface water collected from five shoreline points within a radius of 10m. Where possible 20 samples were collected at roughly equidistant points around the perimeter of each lake and filtered within 24 hours. Due to logistic constraints and varying objectives during early project phases, the actual number of samples collected across all lakes ranged from 10 to 21. Samples were further processed and sequenced following metabarcoding protocols established at the University of Hull using a vertebrate specific 12S marker (Riaz et al. 2011). Some modifications to the molecular protocols were made during the course of the project as described in Hänfling et al. (2016a), Li et al. (2019) and Willby et al. (2019).

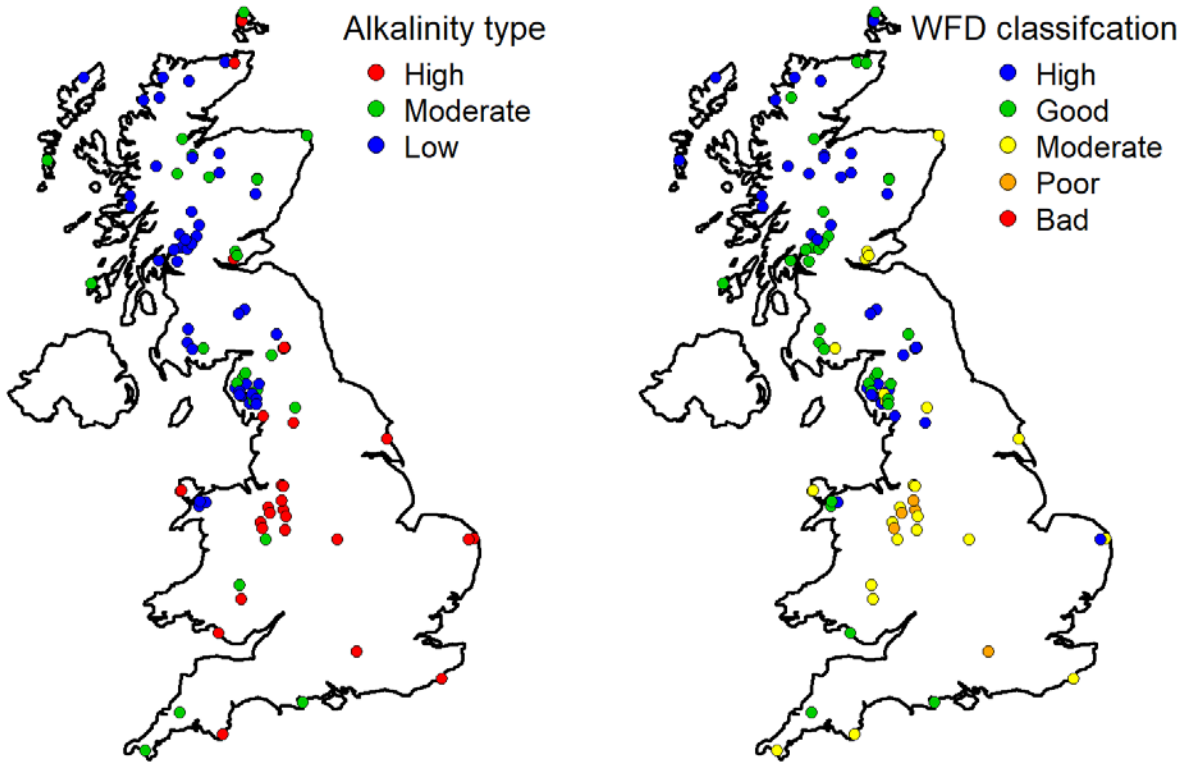


**Figure 1** Distribution and characteristics of 101 UK lakes with associated eDNA data. Shown are alkalinity type (left) and existing WFD classification (right) for each lake. Redrawn based on data from Willby et al. (2019).

### Bioinformatics and data analysis

2.2 Raw sequence data collected during various phases of the Project were re-analysed using the same version of the bioinformatics pipeline as described in (Willby et al. 2019). Sequencing reads from all lakes underwent taxonomic assignment against a curated UK fish

species reference database. Additionally, sequencing reads from 81 of the lakes had taxonomic assignment against a curated UK mammal species reference database. This resulted in two separate datasets that were then treated identically for sequence read clean up. Following taxonomic assignment a noise threshold of 0.1% of total reads per sample was applied to remove low frequency reads (Hänfling et al. 2016a). Most reads were assigned to the species level, but as the molecular marker used here cannot distinguish certain species reliably, the reads belonging to these species were assigned to the next possible higher taxonomic level. Fish belonging to the genera *Coregonus; Lampetra* and *Salvelinus* and mammals belong to the genera *Apodemus*, *Microtus* and *Myotis* were assigned to the genus level, and two members of the fish family Percidae (*Perca fluviatilis, Sander lucioperca*) were assigned to the family level. All remaining assignments to taxonomic levels higher than species were excluded from the analysis. Reads assigned to positive controls, reads which could not be assigned to any taxon and samples with no taxonomically assignable reads were also removed from the data set.

### Description of the final data sets

2.3 The final cleaned fish dataset for all 101 lakes consisted of 1819 samples. Individual lakes ranged from having 9 to 21 successfully sequenced samples with the majority (n = 71) having ≥ 20 samples. A total of 40 fish taxa were recorded across all lakes. Fish species richness of individual lakes ranged from 2 to 18 species (mean 7.79 ± 3.37 SD). A total of 41 mammal taxa were recorded across the 81 lakes with mammal data. Mammal species richness of individual lakes ranged from 3 to 23 species (mean 12.43 ± 4.91 SD).

## Lake fish biodiversity metrics

2.4 The following biodiversity metrics were calculated based on all samples of each lake and for each reduced sample number replicate (see section 2.3).

### Total species richness (ɣ – diversity)

2.5 Read counts data (number of raw reads assigned to fish species) for each lake were converted into species presence/absence. Total species richness was calculated as the total number of fish species detected across all samples.

### Relative abundance

2.6 The proportion of positive samples for a species per lake (site occupancy) was used as a surrogate for relative abundance (Hänfling et al. 2016a). Mean site occupancy, the mean of all fish species occupancy, was used as a proxy for overall fish detection probability in a lake.

### Community composition

2.7 Total read counts per species across all samples from a lake were converted to relative reads (proportion total reads) to create a standardised community composition estimate.

# Effect of sample number on lake fish biodiversity metrics species

2.8 Two principal approaches were used to evaluate the effect of reduced sampling on fish detection and community composition estimation from lake water eDNA metabarcoding, statistical estimation of sampling threshold and data re-sampling techniques.

## Sampling threshold

2.9 Presence/absence data were used to determine the “sample coverage”, an algorithmic coverage-based rarefaction and extrapolation method to measure samples by completeness of species richness (Chao et al. 2014). Sampling threshold is defined as the minimum number of samples required to get 95% sample coverage of the lake, so determining the sample size sufficient to capture ≥ 95% of complete species richness for a given lake. This provides a measure of sampling efficiency which is independent of species richness and therefore comparable across different lakes.

## Random resampling of lake fish metabarcoding data

2.10 Random data resampling techniques were used to generate replicate data sets with reduced sample numbers for each lake. In order to improve comparability across the dataset, only lakes with ≥ 15 samples (83 lakes) were used for resampling. For each lake set consisting of *n* samples (*n* ranging from 15-21) all possible unique sample combinations at different sample sizes were generated with sample size ranging from 2 to a maximum of n-2. The number of possible sample combinations vary depending on total *n* and range from 105 (*n*=15, 13 samples drawn) to 352,716 (*n*=21, 10 samples drawn). For each lake, subsets of 100 unique combinations per sample size were randomly drawn and used as resampling replicates. Unlike a bootstrap resampling approach, there was no possibility of replicate duplication.

2.11 The effect of sample number on species detection and community composition estimates was investigated in the following way. First, the number of undetected species compared to the full data set was calculated for all combinations at each sample size. Second, the average deviation of a combination’s community composition from the full lake sample was quantified for each sample size using pairwise dissimilarity measures (Bray-Curtis dissimilarity index). In order to quantify the effect across all lakes the proportion of lakes which fall above a certain threshold value at each sample size was calculated. Threshold values of 1, 2, and 3 were used as “minimum undetected species”. An arbitrary value of 0.1 was used for the dissimilarity index threshold. The sample size at which 95% of the lakes achieved less than these thresholds was considered.

## Non-random reduced sampling of lake metabarcoding data

2.12 Random resampling provides the opportunity to explore a wide range of sample numbers but ignores the spatial context in which the samples are collected. Hence, under the assumption the eDNA is not randomly distributed it might not represent a realistic sampling strategy. For example for the data set analysed here samples were collected at equidistant points around a lake perimeter. To address this we created replicate data sets, which better reflected the original sampling design by splitting the samples from each lake into two interleaved subsets, i.e. two sets of 10 equidistantly distributed samples. Practically this was achieved by grouping samples into odd and even sample numbers since samples were continuously numbered along the shoreline transect. Only lakes with exactly 20 (n = 70) were

used for this comparison. Undetected species and dissimilarity indices were calculated for each lake subset as above and tested against the maximum threshold values decided for each, 1 and 0.1 for undetected species and dissimilarity indices respectively. Additional non-metric multidimensional scaling (NMDS) ordination was used to visualise differences in lake community estimates between all lakes and sample subsets based on proportion reads. Finally, we applied the overall fish EQR of Willby et al. (2019), a eutrophication relevant metric, to determine the effect of reduced sampling on lake water quality assessment from fish metabarcoding data.

## Mammal distribution from lake water eDNA metabarcoding

2.13 Detected mammal taxa from the 81 lakes were divided into four groups loosely representative of eDNA sources; domesticated, terrestrial, semiaquatic and flying mammals. All species records were converted to presence/absence per lake. For all species, we calculated the number of lakes where a species was recorded and the mean site occupancy within the lake.

2.14 From the presence/absence data we focused on four key species; two semiaquatic mammals, Eurasian beaver (*C. fiber*) and Eurasian otter (*L. lutra*), and two terrestrial mammals Eurasian red squirrel (*S. vulgaris*) and the European pine marten (*M. martes*). We assessed the ability of water sample eDNA metabarcoding to predict the distribution of these species of interest across England and Scotland.

# 3 Results

## Effect of sample number on lake fish species biodiversity metrics

### Sampling threshold

3.1 Regardless of actual sample size, 96% of the 101 lakes achieved sample coverage ≥ 95% for fish species detection at 20 samples (**Figure 2)**. 89% lakes achieved ≥ 95% sample coverage at a sample size of 10. 95% of all the lakes achieved ≥ 95% sample coverage at a sample size of 15 (**Figure 3**). Sampling threshold for lakes ranged from 1 to 25 samples with the mean sample threshold at 5.67 (SD 4.73). Sampling threshold correlated with total species richness ($r_s$ = 0.42, $p < 0.05$) and mean occupancy ($r_s$ = -0.83, $p < 0.05$). There was no correlation between sampling threshold and lake area ($r_s$ = -0.09, $p = 0.39$).

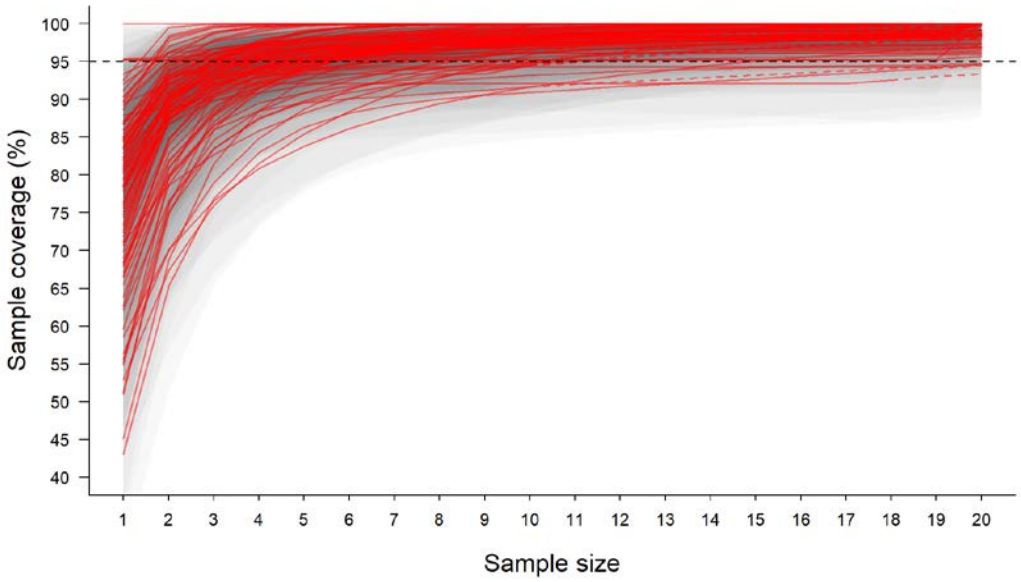


**Figure 2** Sample coverage for all 101 UK lakes used in this study. Sample size cut off at 20 for uniformity. Solid red lines are the interpolated sample coverage. Grey area shows range of upper and lower confidence intervals. Horizontal dashed line indicates 95% sample coverage.

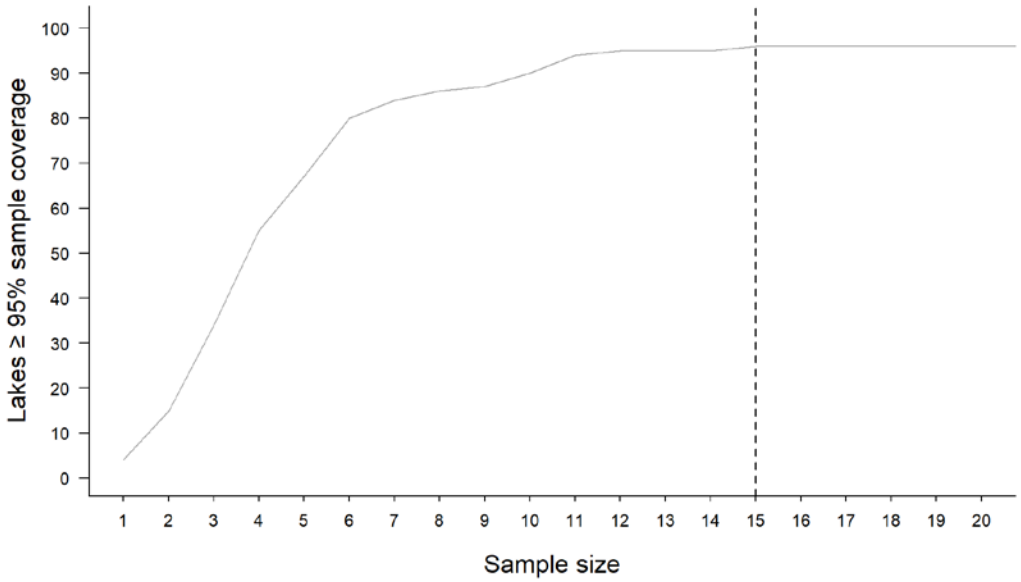


**Figure 3** Proportion of lakes used in this study with ≥ 95% sample coverage per sample size. Sample size cut off at 20 for uniformity. Vertical dashed line indicates sample size at which 95% of lakes attain 95% sample coverage.

## Random resampling of lake metabarcoding data

3.2 Both the number of undetected fish species and dissimilarity index of community composition decreased continuously with increasing sample size (**Figures 4 and 6**). The point at which 95% of the lakes fall below the thresholds of 1, 2 or 3 species undetected were at sample sizes of 15, 10 and 6 respectively (**Figure 5**). Undetected species at a sample size of 10 correlated with total species richness ($r_s = 0.72$, $p < 0.05$) and mean occupancy ($r_s = -0.68$, $p < 0.05$). There was no correlation between undetected species at sample size 10 and lake area ($r_s = 0.04$, $p = 0.71$). 95% of the lakes fell below the dissimilarity index threshold of 0.1 at a sample size of 16 (**Figure 6**).

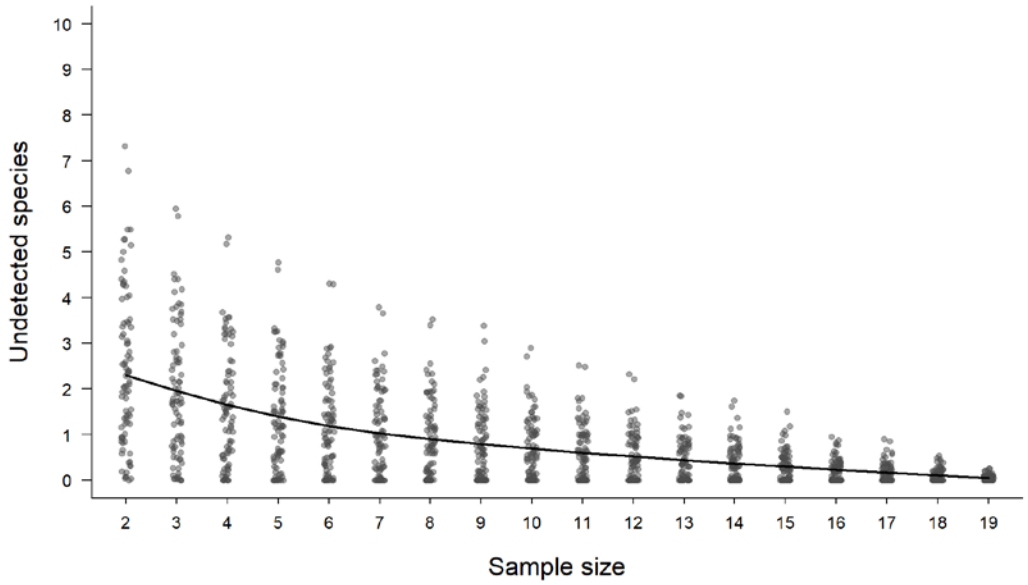


**Figure 4** Number of undetected fish species from resampling of 83 lakes used in this study. Each lake analysed had a successfully sequenced sample size of ≥ 15 (maximum 21). Points represent the mean count of undetected species for combination replicates of a lake at a given sample size. Solid line shows the mean of all points at a sample size.

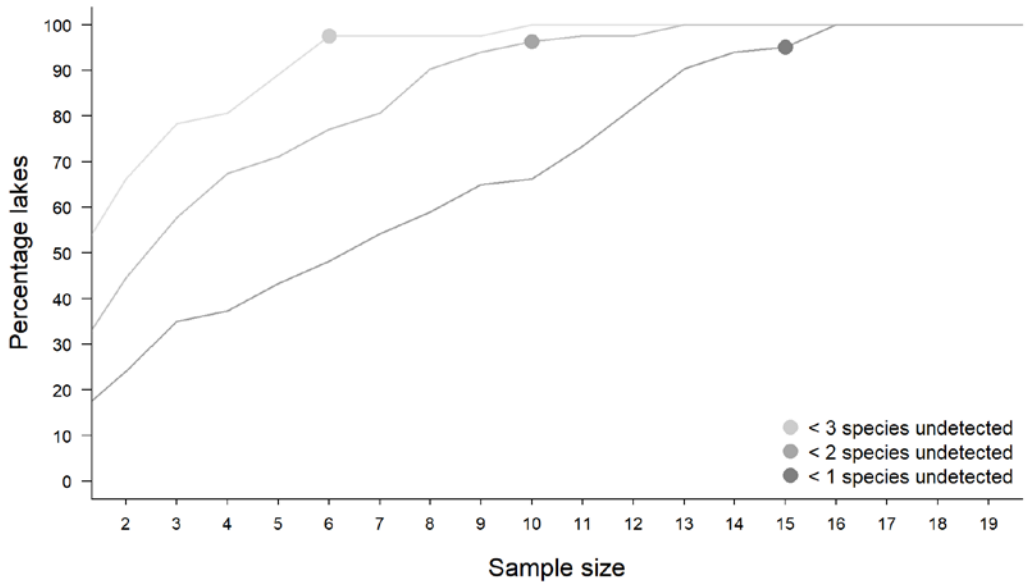


**Figure 5** Proportion of the 83 lakes used in this study where less than 1, 2 and 3 species were undetected respectively at different sample sizes. Each lake analysed had a successfully sequenced sample size of ≥ 15 (maximum 21). Points indicate at which sample size 95% or higher of lakes achieved less than 1, 2 or 3 undetected species.

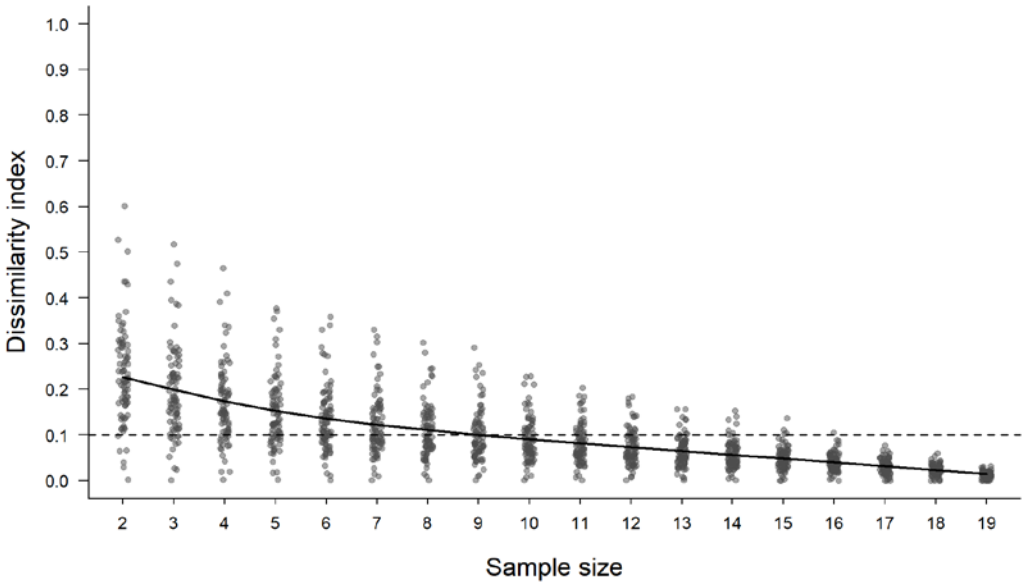


**Figure 6** Bray-Curtis dissimilarity index of fish communities from resampling of 83 lakes used in this study. Each lake analysed had a successfully sequenced sample size of ≥ 15 (maximum 21). Points represent the mean dissimilarity index for species composition of combination replicates in comparison to that of the whole lake at a given sample size. Solid line shows the mean of all points at a sample size. Horizontal dashed line indicates the decided dissimilarity index threshold (0.1).

## Non-random reduced sampling of lake fish species metabarcoding data

3.3 In most cases, the number of undetected species was equal between subset or differed by a single species (**Figure 7**). However, in a few cases the number of undetected fish species differed greatly in subsets. Only 27 lakes detected all species present in both subsets, all other lakes showed differences in the number of species detected. The size of differences in species detection between odd and even subsets correlated with total species richness ($r_s$ = 0.40, $p < 0.05$) and mean occupancy ($r_s$ = -0.53, $p < 0.05$). There was no correlation with lake area ($r_s$ = 0.00, p = 0.97).

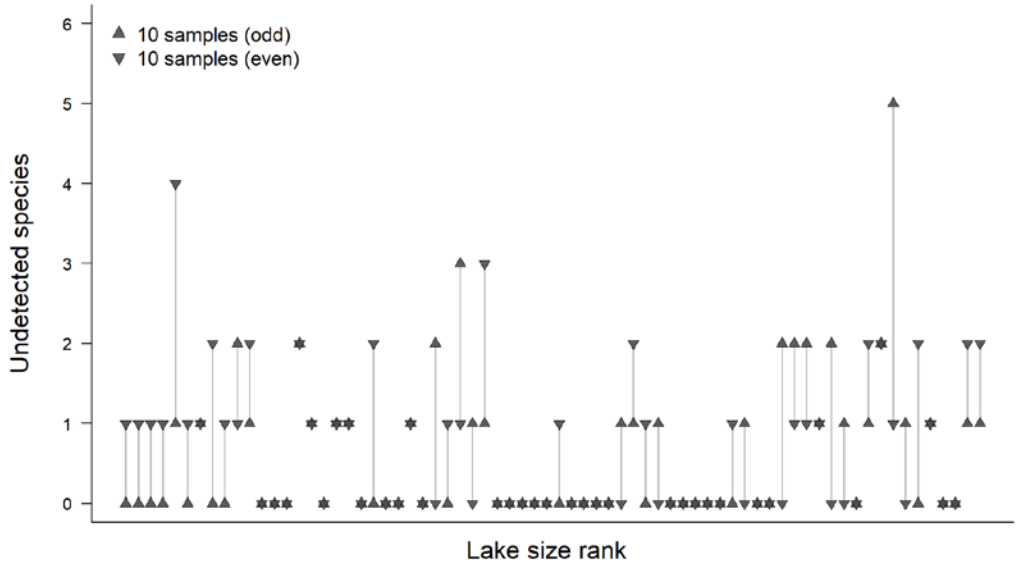


**Figure 7** Undetected fish species counts for sample subsets of 70 lakes. All lakes had 20 samples divided into odd (triangles) and even (inverted triangles) 10-sample subsets. Undetected fish species counts were calculated from comparison of each subset to the whole lake. Lakes are ranked by area; smallest (left) to largest (right). Where present, vertical lines are visual links for corresponding odd and even subsets.

3.4. Bray-Curtis dissimilarity indices of the fish communities represented in odd and even subsets per lake were not greatly dissimilar and remained closely associated (**Figure 8**). All but 4 of the lakes had dissimilarity indices for both subsets below the 0.1 threshold.

3.5. Non-metric multidimensional scaling of whole lake fish species community estimates (species proportion reads) and those of their odd and even 10-sample subsets demonstrated there was little overall difference (**Figure 9**). With the exception of 13 of the selected 70 lakes (those with extended ellipses), all whole lake ordinations were tightly grouped with those of their respective odd and even subsets.

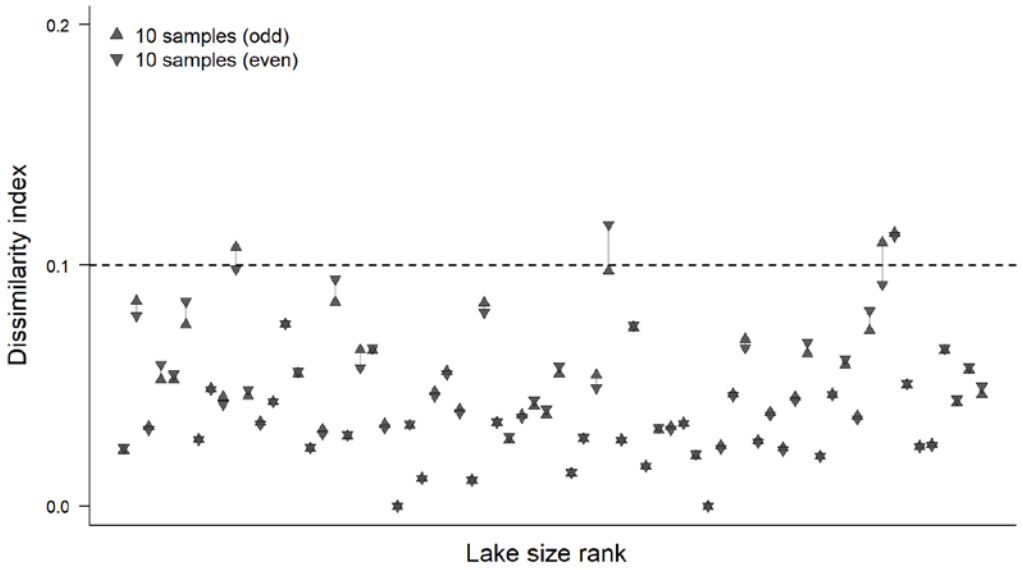


**Figure 8** Bray-Curtis dissimilarity indices for fish communities in sample subsets of 70 lakes. All lakes had 20 samples divided into odd (triangles) and even (inverted triangles) 10-sample subsets. Dissimilarity indices were calculated from comparison of each subset community composition (proportion reads) to the whole lake. Lakes are ranked by area; smallest (left) to largest (right). Where present, vertical lines are visual links for corresponding odd and even subsets.

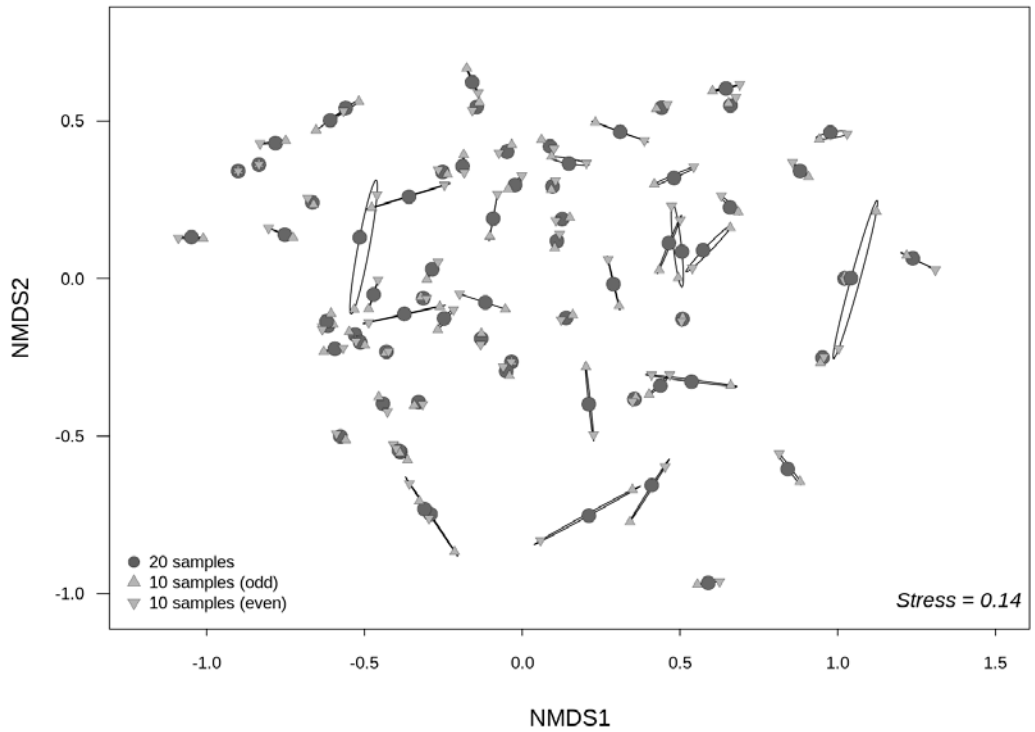


**Figure 9** Non-metric multidimensional scaling (NMDS) ordination for fish communities of 70 lakes based on two subsets of 10 samples. NMDS generated from species composition (proportion reads) estimates using Bray-Curtis dissimilarity method in 3 dimensions (stress = 0.14). All lakes had 20 samples divided into odd (triangles) and even (inverted triangles) 10-sample subsets. Whole lake ordinations (circles) are shown in relation to their odd and even subsets. Ellipses denote the overall spread between subset composition estimates to that of the lake as a whole.

3.6 Application of the eutrophication relevant metric of Willby et al (2019) showed little deviation between odd and even 10-sample subsets to those of the whole lake (**Figure 10**). There were no distinct increases across the established class boundaries of water quality. All points from odd and even subsets to those of the whole lake remained tightly grouped.

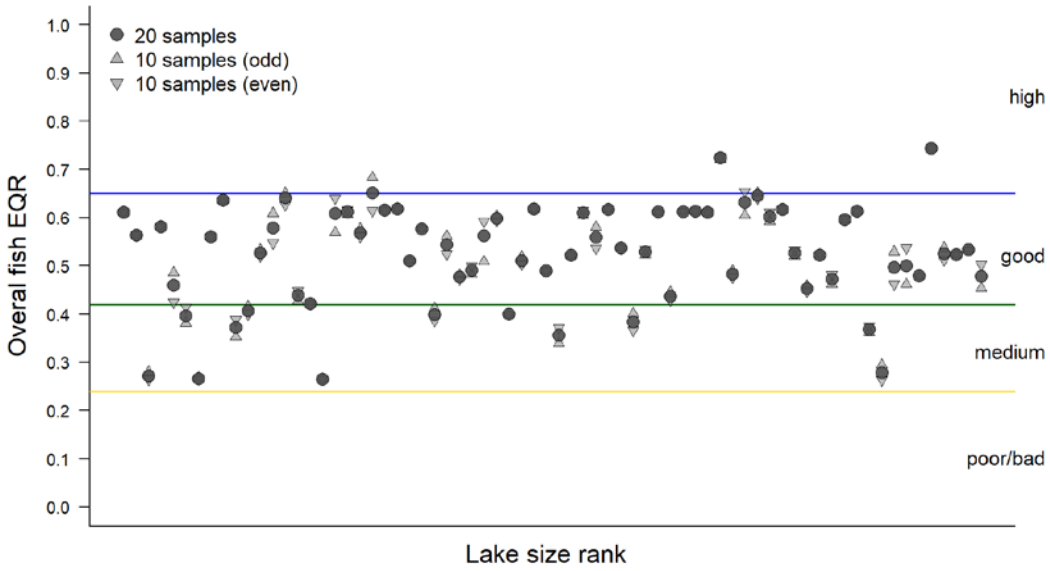


**Figure 10** Water quality for 70 lakes as based on overall fish EQR based on two subsets of 10 samples. All lakes had 20 samples (circles) divided into odd (triangles) and even (inverted triangles) 10-sample subsets. Overall fish EQR, a eutrophication relevant metric, was calculated as per Willby et al. (2019). Lakes are ranked by area; smallest (left) to largest (right). Horizontal lines indicate class boundaries of water quality as per Willby et al. (2019).

## Mammal distribution from lake water eDNA metabarcoding

3.7 Mammal eDNA across the 81 lakes was predominantly from domesticated mammals (**Table 1**). These species were found in over 75% of the lakes and had some of the highest detection probabilities of all mammal taxa. Detected wild mammal taxa were largely terrestrial mammals (26 taxa), whereas semiaquatic and flying mammals were appreciably lower (6, 5 and 4 taxa respectively). Bank voles (*Microtus spp*.) were the most numerous of terrestrial mammals, detected in over 70% of the lakes and having the highest detection probability of non-domesticated mammals. Despite a wide distribution, most wild mammal species had low detection probability across lakes in which they were detected (**Table 1**).

3.8 The four focal key species; Eurasian beaver (*C. fiber*), Eurasian otter (*L. lutra*), Eurasian red squirrel (*S. vulgaris*) and the European pine marten (*M. martes*) had low detection probability (**Table 1**). From the presence/absence data of the 81 lakes (Figure 10), species distribution closely matched existing records of the species in the UK (Campbell-Palmer et al. 2018, IUCN Red List, NBN Atlas).

**Table 1** Distribution summary of 41 identified mammalian taxa across 81 lakes. Shown are species' common names, latin binomials and the number of lakes in which it was detected with positive eDNA signal. Additionally, mean occupancy indicates detection probability of the species across all the lakes in which it was detected. Taxa are divided into four representative groups; domestic, terrestrial, semiaquatic and flying mammals.

| Common name | Latin binomial | Number of lakes with positive eDNA signal detection | Mean occupancy across lakes with positive detection |
|---|---|---|---|
| **Domestic mammals** | | | |
| Domestic cattle | Bos taurus | 73 | 0.68 |
| Domestic sheep | Ovis aries | 72 | 0.67 |
| Domestic dog | Canis lupus | 67 | 0.31 |
| Domestic pig | Sus scrofa | 64 | 0.28 |
| Domestic horse | Equus caballus | 13 | 0.13 |
| Domestic goat | Capra hircus | 8 | 0.12 |
| **Terrestrial mammals** | | | |
| Field voles | Microtus spp. | 69 | 0.45 |
| Field and wood mice | Apodemus spp. | 57 | 0.24 |
| Bank vole | Myodes glareolus | 44 | 0.14 |
| Red deer | Cervus elaphus | 43 | 0.6 |
| European roe deer | Capreolus capreolus | 42 | 0.3 |
| Red fox | Vulpes vulpes | 41 | 0.15 |
| Common shrew | Sorex araneus | 40 | 0.14 |
| European rabbit | Oryctolagus cuniculus | 36 | 0.27 |
| Brown rat | Rattus norvegicus | 35 | 0.27 |
| Eurasian pygmy shrew | Sorex minutus | 30 | 0.09 |
| European badger | Meles meles | 24 | 0.12 |
| European hare | Lepus europaeus | 20 | 0.19 |
| European mole | Talpa europaea | 18 | 0.13 |
| Eastern grey squirrel | Sciurus carolinensis | 15 | 0.21 |
| Mountain hare | Lepus timidus | 9 | 0.34 |
| Fallow deer | Dama dama | 6 | 0.12 |
| Eurasian harvest mouse | Micromys minutus | 6 | 0.15 |
| Eurasian red squirrel | Sciurus vulgaris | 6 | 0.08 |
| Sika deer | Cervus nippon | 4 | 0.05 |
| European pine marten | Martes martes | 4 | 0.15 |
| House mouse | Mus musculus | 4 | 0.06 |
| Chinese muntjac | Muntiacus reevesi | 3 | 0.1 |
| Stoat | Mustela erminea | 3 | 0.07 |
| Common weasel | Mustela nivalis | 3 | 0.17 |
| European polecat | Mustela putorius | 1 | 0.05 |
| Reindeer | Rangifer tarandus | 1 | 0.2 |
| **Semiaquatic mammals** | | | |
| Eurasian otter | Lutra lutra | 45 | 0.14 |
| Eurasian water shrew | Neomys fodiens | 41 | 0.16 |
| European water vole | Arvicola amphibius | 28 | 0.36 |
| American mink | Neovison vison | 4 | 0.08 |
| Eurasian beaver | Castor fiber | 3 | 0.18 |
| **Flying mammals** | | | |
| Common pipistrelle | Pipistrellus pipistrellus | 5 | 0.1 |
| Mouse-eared bats | Myotis spp. | 4 | 0.09 |
| Brown long-eared bat | Plecotus auritus | 1 | 0.05 |
| Grey long-eared bat | Plecotus austriacus | 1 | 0.05 |

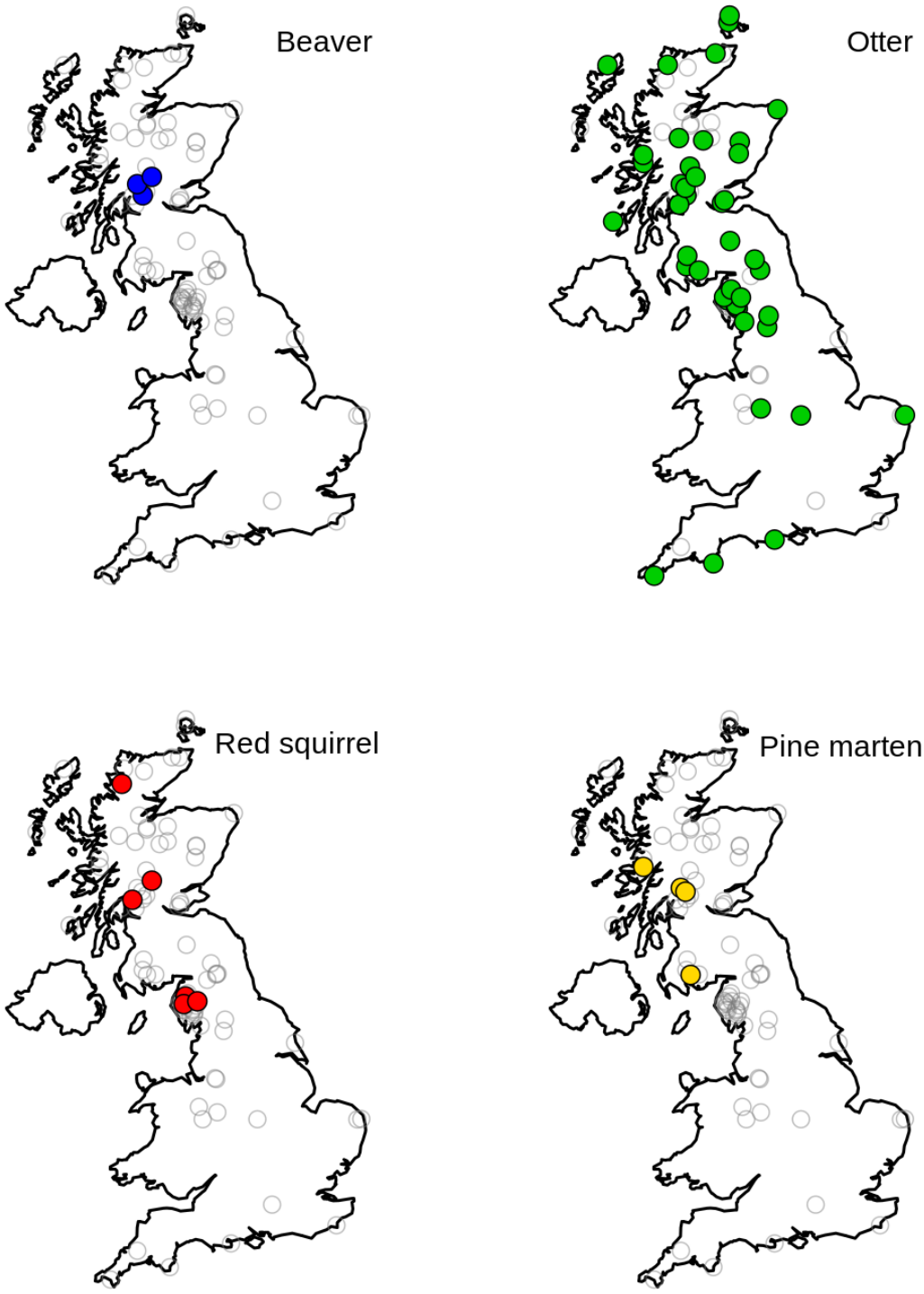


**Figure 11** Distribution of the eDNA signal from key mammal species in England and Scotland from 81 lakes used in this study. Shown are beaver (*Castor fiber*) (blue), otter (*Lutra lutra*) (green), red squirrel (*Sciurus vulgaris*) (red) and pine marten (*Martes martes*) (yellow). Hollow grey points indicate sampled lakes with no positive eDNA record of the species.

# 4 Discussion

## Reduced sampling on fish species detection and community composition estimation

4.1 The results of the sample coverage analysis confirmed that the sampling design used to create the original data set i.e. 20 samples from equidistant locations around the lake provided a very reliable estimation of the true species richness with only 4% of lakes having an estimated sample coverage below 95% at this sample size (Hänfling et al. 2016a; Willby et al. 2019). However, for most lakes the sample coverage curves are starting to reach a plateau at much lower sample numbers indicating that the loss of signal is relatively small even with a substantially lower sampling effort. This was confirmed by the resampling analysis which indicated that in the majority of lakes on average less than 2 species remain undetected with a sample size of 10 randomly distributed samples and even lower rate of undetected species when samples are non-randomly distributed as it would be applied in most real world scenarios. Interestingly lake size does not directly determine the required sampling effort. However, as the required sample size increases with species richness, the prior knowledge of expected species richness can be used to design efficient sampling strategies.

4.2 We demonstrated that a minimum of 10 samples taken from a lake was sufficient to represent fish species composition and does not greatly affect ecological community analysis. The condition and quality of a lake can still reliably be inferred through the presence of easily detected, commonly occurring species with this lower sample size. A reduction of sampling effort does have drawbacks as detection of rare or low abundance species is reduced. Therefore, sampling strategies aiming to provide accurate distribution records of endangered or invasive species should be based around the higher sample numbers as used in the original data collection. The reduced sampling approach is best suited to the low diversity lakes of the UK and is ideal for use with established fish-based water quality assessment metrics (ie Willby et al. 2019). Increased diversity, as is found in mainland European lakes, will possibly demand an increase in sample size.

## Lotic eDNA in lentic systems

4.3 In the data set analysed here we detected fish species more typically associated with river systems (rheophilic fish) in lake water samples, such as bullhead, grayling, lamprey and salmon. Rivers have been shown to transport eDNA over great distances (Deiner et al. 2016), although eDNA quantity decreases rapidly during this process (Pont et al. 2018). Hence some detections, especially rare ones, could reflect contamination from river water (Deiner et al. 2017). However, rheophilic fish also occur in lake estuaries, stray into the lake and might even breed there. From sequencing data alone it is impossible to disentangle if a detection is true occupancy or transit of eDNA from rivers to the lake. In any case the fish fauna of a lake and its tributaries are closely connected and it is more appropriate to regard the eDNA sampling in lakes as a catchment approach.

## Mammal distribution from lake water eDNA metabarcoding

4.4 A positive eDNA signal for many wild UK mammal species was detected across all the lakes in this study but all had relative low detection probabilities compared to fish (**Table 1**).

The source of this eDNA could be from the animals interacting directly with lake shorelines or from river inflow close to sample collection points. Of the four key mammal species focussed upon, the Eurasian beaver had positive eDNA signal from three lakes and Eurasian otters were widespread across many. Eurasian red squirrel and European pine marten, both being arboreal in nature, were also successfully detected in a small number of lakes. This demonstrates that both arboreal and semiaquatic mammal presence can be reliably detected from lake water eDNA, providing sample size is sufficient to accommodate for low detection probabilities.

## Marker limitations: false negatives and positives

4.5 In this study we used the 12S marker of Riaz et al. (2011) which can distinguish between many vertebrate taxa. The marker region is short (106bp for most vertebrate species) and therefore has limited resolution to identify very closely related species. We have already alluded to the limitations of detection in fish taxa; *Percidae*, *Coregonus spp*., *Salvelinus spp*. and *Lampetra spp*., and also in mammal taxa; *Apodemus spp.*, *Microtus spp.* and *Myotis spp.* Coupled with this, in some cases a single sequencing error could cause a change of species assignment although this would result in a very low number of sequences which in most cases would fall well below the filtering threshold. Additionally the reference database is incomplete and a detailed in-silico analysis hasn't been carried out for this group. Therefore the taxonomic assignments include a high uncertainty. A specific case is the European wild cat (*Felis silvestris*) where the reference data sequences differ only by one bp but it is likely that they share sequence haplotypes due to widespread hybridisation. As this will result in a large number of false positives and negatives, these species were excluded from the data set. Sequences assigned to the wild cat were found in 13 lakes across Scotland but the distribution did not match that of known records. The reference database for mammals is also still incomplete for some subgroups such as bats and is restricted to a single sequence for a number of taxa, which increases the risk of false negatives and miss-assignments respectively. Birds can also be detected with the 12S marker used here but the resolution is low for a number of key groups such as ducks and gulls which in many cases can't be assigned to species level.

## Conclusion and further recommendations

5.1 Sample size is an important factor to consider and depends greatly on the question to be answered. The results of this study provide an important overview of how sample effort affects various metrics of fish species richness which will provide guidance on designing the best sampling strategy for individual projects. The best approach would be to define clear objectives in terms of acceptable error and then use the figures presented in this report to decide the specific sample number required for each project. As a general rule to achieve an overview of species composition in low diversity lakes, as typical for many UK regions, a sample size of 10 will suffice, regardless of lake size. This would make for simpler logistics, less intensive sampling surveys and allow for sampling of more lakes. Fewer samples also removes pressure during downstream sample processing, e.g. filtration and DNA extraction.

5.2 However, sample size will need to be increased if rare species detection is required, or when sampling high diversity lakes. Sample sizes of at least 20 per lake will be required if rare species should be detected with a high certainty. It is important to note that our results depend on the specific sequencing library preparation method used here. If methods are used which have lower or higher detection probabilities (eg different number of PCR replicates) within individual samples the results would look different.

5.3 The development of eDNA metabarcoding approaches has progressed rapidly over the last few years and these approaches have already contributed to monitoring biodiversity. However, there are still a number of aspects which require further research and resources should be directed towards tackling them in order to improve the effectiveness and cost-efficiency of these methods. This include the following issues:

1) The effect of increasing the number of PCR reactions on detection probabilities within samples. This could be carried out using existing eDNA samples from the lake fish study and should include a cost benefit analysis. The results would show whether fewer samples are needed if more PCR replications are carried out and whether this is a more cost-efficient approach.
2) Seasonal variation in eDNA signal. The approach used to generate the lake fish data set was based on collecting samples during winter as data from a small number of lakes indicated that this is the most appropriate sampling season to generate an unbiased picture of a lake fish community. However, other objectives might be better achieved by sampling during different seasons. For example, detection probabilities of rare fish might be higher during high activity periods such as spawning or migration. There is currently a lack of data to support such potentially highly effective sampling strategies.
3) Interpreting spatial variation of eDNA in lotic habitats. There is an increasing number of studies demonstrating that eDNA can be used efficiently in lotic habitats to characterise fish communities on a catchment scale. However, hydraulic modelling approaches are required in order to relate the eDNA signal to a precise species distribution. This will require a larger scale research project such as a PhD studentship.
4) Reference databases are currently still incomplete and poorly covered for UK mammals and birds. This would require generating de novo sequences for a number of species and carry out a detailed in-silico analysis to determine which species can be reliably detected.
5) Further exploration of the distribution of mammal eDNA in relation to existing species records is required to fully understand how much eDNA can contribute towards monitoring of these taxa.

# 5 References


Campbell-Palmer R, Puttock A, Graham H, Wilson K, Schwab G, Gaywood MJ, Brazier RE (2018) Survey of the Tayside area Beaver population 2017-2018. *Scottish Natural Heritage Commissioned Report* No. 1013.

Chao A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, Ellison AM (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* 84:45–67.

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology* 26:5872–5895.

Deiner K, Fronhofer EA, Mächler E, Walser J-C, Altermatt F (2016) Environmental DNA reveals that rivers are conveyer belts of biodiversity information. *Nature Communications* 7:12544.

Hänfling B, Lawson Handley L, Harper LR, Benucci M, Sellers GS, Di Muri C, Griffiths N, Jaques R, James J (2020) Development of an eDNA-based lake fish classification tool – Collection of data from English Lakes. *Report to the Environment Agency, UK* (Unpublished).

Hänfling B, Lawson Handley L, Read DS, Hahn C, Li J, Nichols P, Blackman RC, Oliver A, Winfield IJ (2016a) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology* 25:3101–3119.

Hänfling B, Lawson Handley L, Read DS, Winfield IJ (2016b) The development of an eDNA-based approach for fish sampling in lochs for WFD - Phase 2. *Report to the Scottish Environment Protection Agency, UK*.

Hänfling B, Lawson Handley L, Read DS, Winfield IJ (2016c) eDNA-based metabarcoding as a monitoring tool for fish in large lakes. Report – SC140018/R. *Report to the Environment Agency, UK*.

Harper LR, Lawson Handley L, Carpenter AI, Ghazali M, Di Muri C, Macgregor CJ, Logan TW, Law A, Breithaupt T, Read DS, McDevitt AD, Hänfling B (2019) Environmental DNA (eDNA) metabarcoding of pond water as a tool to survey conservation and management priority mammals. *Biological Conservation* 238:108225.

IUCN Red List [online]. URL: https://www.iucnredlist.org [Accessed January 2020].

Lawson Handley L, Read DS, Winfield IJ, Kimbell H, Johnson H, Li J, Hahn C, Blackman R, Wilcox R, Donnelly R, Others (2019) Temporal and spatial variation in distribution of fish environmental DNA in England's largest lake. *Environmental DNA* 1:26–39.

Li J, Hatton-Ellis TW, Lawson Handley L, Kimbell HS, Benucci M, Peirson G, Hänfling B (2019) Ground-truthing of a fish-based environmental DNA metabarcoding method for assessing the quality of lakes. *The Journal of Applied Ecology* 56:1232–1244.

Li J, Lawson Handley L, Read DS, Hänfling B (2018) The effect of filtration method on the efficiency of environmental DNA capture and quantification via metabarcoding. Molecular Ecology Resources 18:1102–1114

NBN Atlas [online]. URL: https://nbnatlas.org [Accessed January 2020].

Pont D, Rocle M, Valentini A, Civade R, Jean P, Maire A, Roset N, Schabuss M, Zornig H, Dejean T (2018) Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Scientific Reports* 8:10361.

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.

Sales NG, McKenzie MB, Drake J, Harper LR, Browett SS, Coscia I, Wangensteen OS, Baillie C, Bryce E, Dawson DA, Others (2020) Fishing for mammals: Landscape-level monitoring of terrestrial and semi-aquatic communities using eDNA from riverine systems. *The Journal of Applied Ecology* 57:707–716.

Sellers GS, Di Muri C, Gómez A, Hänfling B (2018) Mu-DNA: a modular universal DNA extraction method adaptable for a wide range of sample types. *Metabarcoding and Metagenomics* 2:e24556.

Willby N, Law A, Bull C, Hänfling B, Lawson Handley L, Winfield IJ (2019) A tool for classifying the ecological status of lake fish in Britain based on eDNA metabarcoding. *Report to the Scottish Environment Protection Agency, UK*.