DNA-based resolution of freshwater arthropod communities and interactions

Thesis submitted for the degree of Doctor of Philosophy

by

Elizabeth Anne Davidson, BSc. (Hons), MSc.

School of Biosciences

Cardiff University

July 2022

UK Centre for Ecology & Hydrology

CARDIFF UNIVERSITY

PRIFYSGOL CAERDYDD

Thesis Summary

Freshwater biodiversity is widely considered to be in crisis. The need for new methods to quickly assess and monitor biodiversity is urgent. DNA-based identification of biodiversity offers promising new methods. Compared to monitoring species presence and richness, DNA-based identification of ecological interactions through predator diets has gained much less attention. The study of these interactions has the potential to provide new insights into species ecology and new methods of identifying and monitoring changes in biodiversity.

This thesis aimed to explore how DNA-based resolution of predator diets can contribute to improving the assessment and monitoring of freshwater biodiversity and to identify the method development necessary to achieve this improvement. The studies in this thesis aimed to improve understanding of how DNA-based identification methods can be used to characterise freshwater arthropod communities and to resolve trophic interactions within those communities.

Gaps in the coverage of publicly stored reference sequences for UK freshwater arthropods coupled with low data quality will lead to poor taxonomic resolution and potential misidentifications. A systematic framework for prioritising current and future sequencing and curation needs was produced to address these issues. Thorough optimisation and validation of methods for metabarcoding of bulk zooplankton samples was found to be essential in order to produce meaningful community data. Optimisation of primers, bioinformatic processing and data analysis improved the detection of target taxa, reduced false positives and negatives, and improved relative abundance data. The combination of community metabarcoding of prey taxa and dietary screening of individual predators provided detailed and informative information on how predator-prey interactions change over time.

Overall, DNA-based resolution of predator diets has great potential for resolving freshwater food web interactions and providing new insights for improving assessment and monitoring. Curation of reference databases and method development for freshwater arthropods are urgent to realise this potential.

Acknowledgements

This research was funded by a NERC GW4+ Doctoral Training Partnership studentship from the Natural Environment Research Council [NE/L002434/1] and a CASE partnership with the National Trust.

I would like to say a huge thank you to my supervisory team, Steve Thackeray, Steve Ormerod, Ellie Mackay, Dan Read and Stewart Clarke for all their help and support throughout my PhD. I am especially grateful to Steve Thackeray and Ellie Mackay for their unwavering support and encouragement through all the challenges of the last almost five years. I have been very lucky to have such a great supervisory team.

I would also like to thank everyone at UKCEH, especially everyone in the Lakes Ecosystem Group. I feel very lucky to have been based within the lakes group and I am grateful for all the support from so many people throughout my PhD. A special thanks to Alex Elliott for being a great office friend, Glenn Rhodes for teaching and supporting me in the lab, and Ben James for all his support with night-time fieldwork. Thanks also to Freya Olsson and Naomi Lumsden for helping me with fieldwork and to Tim Goodall for supporting my sequencing work at Wallingford.

Finally, I would like to thank my family and friends who have been patient with how busy I have been, believed in me, and supported me throughout my PhD.

Table of Contents

List of Figures

List of Tables

1 General introduction

Fresh waters cover just 0.8% of the Earth's surface but are estimated to support at least 9.5% of described species, and as much as one third of all vertebrates (Dudgeon et al. 2006; Balian et al. 2008; Tickner et al. 2020). As such, freshwater ecosystems make a disproportionate contribution to global biodiversity which has both intrinsic value, based on the inherent worth of organisms, and utilitarian value, based on benefits provided to humans as 'ecosystem services', such as clean water and food. Freshwater biodiversity is widely considered to be in crisis (Reid et al. 2019; Tickner et al. 2020). Estimates suggest that approximately one third of freshwater species are threatened with extinction (Darwall et al. 2018) and the Living Planet Index reveals more rapid population decline in fresh waters than in either terrestrial or marine systems (Grooten and Almond 2018).

Most of the world's freshwater habitats have not been comprehensively assessed, so baseline data on what species are present in which habitats are not available (Abell 2002). Given this, we are likely to be losing species before their existence is even documented and are likely to be severely restricted in our understanding of the role of freshwater biodiversity in sustaining healthy ecosystems. Traditional methods of assessing and monitoring biodiversity are resource-intensive and dependent on expert taxonomist knowledge to identify species. It is impractical to attempt to fill in the gaps in our knowledge of species distributions with traditional methods alone (Cristescu and Hebert 2018). With declining expertise in freshwater taxonomy (Hopkins and Freckleton 2002), coupled with an increasing need for biodiversity assessments to address the baseline data gap, the need for new methods to quickly assess and monitor biodiversity is urgent.

DNA-based identification of biodiversity offers promising new methods that have the potential to revolutionise biodiversity assessment and monitoring (Lawson Handley 2015; Cristescu and Hebert 2018). Screening samples for particular target taxa using diagnostic PCR (using species- or group-specific primers to amplify only the taxa of interest) is a very sensitive and accurate

method for detecting species presence. Screening samples for great crested newts (*Triturus cristatus*) using diagnostic PCR is now the standard method for assessing the presence of the species because of its sensitivity and accuracy in comparison with traditional monitoring methods (Rees et al. 2014; Biggs et al. 2015). These assays require thorough development and validation but once they are developed, they provide high confidence for use in monitoring. However, they only provide data on the specific target taxa rather than whole communities.

Metabarcoding offers species detection at high-taxonomic resolution and has enabled quick, cost-effective analysis of many samples, facilitating much more rapid biodiversity assessments than is possible using traditional methods alone. Metabarcoding of community samples, especially eDNA, is widely believed to be a game changer for biodiversity assessment and monitoring (Lawson Handley 2015). Although metabarcoding of community samples is very effective at providing assessments of species present in a community, biases throughout the process may make it less suitable for monitoring change in communities over time (Cristescu and Hebert 2018).

Uncertainties in metabarcoding data are caused by biases at various points within the analytical process such as primer bias, reference database coverage, clustering and filtering (Alberdi et al. 2017). One of the main challenges with using metabarcoding data for monitoring purposes is that it is not currently possible to achieve accurate and unbiased measurements of abundance of the taxa detected (Elbrecht and Leese 2015; Pinol et al. 2015; Luo et al. 2022). This means that it is currently only possible to compare species richness and composition among samples rather than any changes in the abundance of species. Abundance data are critical for monitoring changes in populations, so it is currently necessary to continue use of traditional methods when abundance data are needed. Furthermore, false positives and/or negatives in metabarcoding data can cause artificial differences in the number of species detected so it can be challenging to use metabarcoding data even just to monitor changes in species richness (Cristescu and Hebert 2018). Before widespread adoption of metabarcoding approaches is possible, standardisation and validation with traditional methods are a key requirement.

Compared to monitoring species presence and richness, DNA-based identification of ecological interactions through predator diets has gained much less attention. The study of these interactions has the potential to provide new insights into species ecology and new methods of identifying and monitoring changes in biodiversity. Understanding specific interactions among species is essential for specific groups such as species of conservation concern and non-native species and can provide vital information on why populations are changing (Clare 2014). The interactions among taxa form the basis of food webs, which have been well-studied in freshwaters, revealing structural properties that underpin freshwater ecosystem functioning and stability (Stouffer 2010; Thompson et al. 2012b; Staudinger et al. 2021). In addition, predators are structurally important in ecosystems (Estes et al. 2011) and sensitive indicators of environmental change (Velarde et al. 2013). Therefore, by resolving the diet of freshwater predators, we can understand the resource needs and consumptive impacts of specific taxa, allowing us to estimate food web metrics that underpin ecological stability, and provide new early-warning indicators of environmental change (Bartley et al. 2019).

1.1 Aims and hypotheses

The overall aim of this thesis is to explore how DNA-based resolution of predator diets can provide benefits for the assessment and monitoring of freshwater biodiversity, and improve understanding of what method development is necessary to be able to gain these benefits.

The overall hypothesis is that a bias towards using DNA-based methods to replicate data from existing monitoring methods is limiting the full potential of what DNA-based approaches might offer.

The specific hypotheses tested are that:

1. Studies using DNA-based identification of freshwater predator diets are biased towards vertebrates, especially specific individual taxa of interest, rather than invertebrates.

2. Publicly available reference sequences for freshwater arthropods are not currently comprehensive and accurate enough to provide accurate taxonomic identification for most taxa.
3. Optimised and validated metabarcoding of bulk zooplankton samples can provide meaningful data on potential prey communities in lake ecosystems.
4. DNA-based screening of individual predator diets can provide a sensitive method for monitoring changes in communities.

1.2 Chapter structure

This thesis is organised into five subsequent chapters which address the specific hypotheses and synthesise the results to answer the overall thesis aim:

Chapter 2: DNA-based resolution of freshwater predator diets: benefits for biodiversity monitoring and conservation

This chapter provides broad-scale context for dietary studies later in the thesis, presenting a quick-scoping review of studies that have used DNA-based identification methods to resolve the diets of freshwater predators. The literature was systematically searched and then assessed to determine the different methods that have been used, the coverage and scale of trophic interactions investigated, and how these methods were used for the benefit of monitoring and conservation of freshwater biodiversity. Research needs identified in this chapter form the basis of the following chapters.

Chapter 3: DNA barcodes for UK freshwater arthropod species: coverage, curation and priorities for the future

This chapter assesses data availability for studies seeking to investigate freshwater communities and interactions, considering the overall coverage of UK arthropod species in the Barcode of Life Data System (BOLD), and the coverage of protected and non-native species. The quality of the stored sequences is analysed to assess the potential effect of quality on accurate identification of species using DNA-based methods. Geographic variation in

sequences is analysed to assess whether the origin of barcoded specimens is important in accurate taxonomic assignment.

Chapter 4: Using metabarcoding to identify freshwater zooplankton communities in lakes

This chapter presents work on the optimisation of methods and primers for metabarcoding bulk samples of target zooplankton taxa from the Lake District (United Kingdom) using public reference sequences and sequences of local individuals. Metabarcoding data are validated using morphological count data to understand the strengths and limitations of metabarcoding data.

Chapter 5: DNA-based analysis of the diet of phantom midge, *Chaoborus flavicans*, larvae in a lake ecosystem: combining community metabarcoding and dietary screening to analyse interaction strengths.

This chapter uses the metabarcoding methods optimised in Chapter 4 to characterise the potential diet of phantom midge, *Chaoborus flavicans*, larvae in Blelham Tarn (UK). Metabarcoding of samples from Blelham Tarn provided data on the potential prey community, the behaviour of *Chaoborus*, and sequences of the species in Blelham Tarn. These were used to optimise specific assays to detect prey taxa in dietary samples from *Chaoborus* individuals. Interaction strengths were analysed to showcase a new method for monitoring change in community interactions.

Chapter 6: General Discussion

The final chapter provides an overview and synthesises the findings of the previous chapters to address the thesis aim. Knowledge gaps are highlighted and areas for further research are suggested.

2 DNA-based resolution of freshwater predator diets: benefits for biodiversity monitoring and conservation

2.1 Summary

DNA-based identification can now enable trophic interactions between species to be resolved at unprecedented taxonomic, temporal and spatial resolution, providing new information to benefit biodiversity monitoring and conservation. DNA-based analysis of diet has, so far, been applied less in fresh water than in terrestrial and marine systems.

This study completed a quick-scoping review of studies that have used DNA-based identification methods to resolve the diets of freshwater predators to assess current progress and identify biases and knowledge gaps. This study tested the hypothesis that studies using DNA-based identification of freshwater predator diets would be biased towards vertebrates, especially specific individual taxa of interest, rather than invertebrates.

DNA-based identification of freshwater predator diets was found to already be providing benefits to freshwater biodiversity assessment and monitoring but was biased towards vertebrates, especially those that have piscivorous diets and species of conservation concern or non-native species. Many studies were focused on single predators and snapshots in time and space, but there was some scaling-up to multiple predators and larger spatial and temporal scales. Accurate and informative interaction data depend on thorough optimisation and validation of DNA-based methods.

Thorough method development for different taxonomic groups will enable high-taxonomic resolution of freshwater predator diets across time and space that provide a better understanding of freshwater ecosystems and potentially provide new metrics for monitoring changes in communities.

2.2 Introduction

2.2.1 Freshwater biodiversity

Fresh waters cover just 0.8% of the Earth's surface but are estimated to support at least 9.5% of described species and as much as one third of all vertebrates (Dudgeon et al. 2006; Balian et al. 2008; Tickner et al. 2020). The high biodiversity already known to exist in our freshwater ecosystems is likely a vast underestimation of the true amount for two main reasons. Firstly, our knowledge of biodiversity in global hotspots in tropical regions is very limited in comparison with that in temperate areas (Dudgeon et al. 2006). Secondly, biodiversity knowledge is mainly focused on vertebrates and therefore does not account for much of the vast diversity of other groups such as invertebrates, fungi and algae (Dudgeon et al. 2006; Reid et al. 2019). It is clear from current estimates, and predictions of what is missing from these estimates, that freshwater ecosystems make a disproportionate contribution to global biodiversity.

The biodiversity supported by fresh waters has both intrinsic value, based on the inherent worth of organisms, and utilitarian value, based on benefits provided to humans as 'ecosystem services'. These ecosystem services include the production of clean water, food, and livelihoods and are estimated to be worth over $4 trillion annually (Darwall et al. 2018). Whichever value system is used, it is clear that freshwater biodiversity is of great value to humans, and it is, therefore, essential to protect it now and into the future.

However, freshwater biodiversity is widely considered to be in crisis (Reid et al. 2019). Estimates suggest that approximately one third of freshwater species are threatened with extinction (Darwall et al. 2018) and indicators reveal more rapid population decline in fresh waters than in either terrestrial or marine systems (Grooten and Almond 2018). Pressures on freshwater biodiversity include climate change, non-native species invasions, overexploitation, destruction or degradation of habitat, water pollution, flow modification, harmful algal blooms, and emerging contaminants (Dudgeon et al. 2006; Reid et al. 2019). The pressures on freshwaters are predicted to increase in the future as human consumption continues to increase alongside rapid

environmental change (Darwall et al. 2018). Although freshwater ecosystem services are dependent on the biodiversity inhabiting them, policy and decision-making does not often take the impact on biodiversity into consideration, which leads to unsustainable policies.

It is vital that impacts on freshwater biodiversity are monitored effectively so that decisions and policies on resource-use can minimise the risk to the biodiversity that underpins these resources. A major factor that limits effective monitoring is gaps in our knowledge and understanding of freshwater biodiversity (Dudgeon et al. 2006). Most of the world's freshwater habitats have not been comprehensively assessed, so baseline data on what species are present in which habitats are not available (Abell 2002). Biases in conservation and research towards terrestrial ecosystems and vertebrate groups (Di Marco et al. 2017) results in a lack of data on freshwater biodiversity, particularly for invertebrates, which prevents necessary information being available for decision-making on resource-use.

It is impractical to attempt to fill in the gaps on our knowledge of species distributions with traditional methods alone (Cristescu and Hebert 2018). Traditional methods of assessing and monitoring biodiversity are resource-intensive and dependent on expert taxonomist knowledge to identify species. Environmental DNA (eDNA) has the potential to "revolutionise biodiversity science and conservation action by enabling the census of species on a global scale in near real time" (Cristecu and Hebert 2018, p.209). This is particularly true for aquatic habitats because organisms inhabiting the water are continually shedding DNA, meaning that biodiversity monitoring can be non-invasive and sensitive for aquatic species (Lawson Handley 2015). eDNA is particularly useful for monitoring purposes in freshwater ecosystems because DNA degrades rapidly in freshwater (days to weeks) (Dejean et al. 2011; Thomsen et al. 2012) which means that species detections are likely to show the recent presence of the organism rather than the presence of historic populations (Thomsen and Willerslev 2015). The use of eDNA for detecting the presence of species in freshwaters is developing rapidly and becoming standardised, enabling more comprehensive and widespread assessment of freshwater biodiversity than ever before.

The level of biodiversity assessment possible using eDNA is unprecedented and will provide the necessary baselines of which freshwater species are present in which habitats. However, effectively monitoring changes in, and impacts on, biodiversity requires more than the presence or absence of species.

2.2.2 Importance of interactions

The focus on species richness and the extinction of species is a necessary simplification of biodiversity in order to make the measuring and reporting of global biodiversity a tractable challenge (McMeans et al. 2016). However, biodiversity encompasses the variety of all life on Earth and its interactions, at all levels from genes to ecosystems. Understanding the loss of species, estimated to be at 100 to 1000 times above the natural background rate (Pimm et al. 1995; Ceballos et al. 2015; Ceballos et al. 2017), is of vital importance to understanding how ecosystems are affected by human impacts. However, understanding anthropogenic impact requires not only the knowledge of which species are present or absent, but also an understanding of the ecological processes that occur within ecosystems and how changes in these processes affect ecosystem functions and services (McCann 2007).

The loss of species is the tip of the iceberg in terms of the total loss of biodiversity that is occurring. Changes in species presence in a habitat take time to occur and have different causes. To monitor changes in and impacts on biodiversity, we need methods that are sensitive to the early changes in communities that precede changes in species presence. A very important, but rarely monitored, aspect of biodiversity that could enable more sensitive monitoring of changes in ecosystems is the ecological interactions between species (Tylianakis et al. 2008).

All species interact with other species in networks of interactions including predator-prey, pollinator-plant, and parasite-host interactions. The structure of these ecological networks has been shown to be more important in maintaining ecosystem function compared to species numbers (McCann 2007). Although conserving these ecological networks is dependent on conserving the species themselves, extinctions of interactions often precede species extinctions (Janzen 1974; Tylianakis et al. 2008; Stouffer 2010). Studies have

shown that species interactions are sensitive to environmental pressures and can change in frequency or be lost in response to environmental change (Tylianakis et al. 2008). Furthermore, the loss of an interaction might drive the extinction of one of the interacting species.

In fresh waters, trophic interactions are thought to be the most important interaction type (Woodward 2009). Trophic interactions between species depend on the interacting species co-occurring in space and time and this is affected by environmental change. For example, changes in lake temperature caused by climate change cause cold water adapted fish to shift their habitat use and foraging patterns within the lake, resulting in changes in their feeding interactions (Bartley et al. 2019). Shifts in timings caused by climate change can also cause changes in species interactions. Timing of diatom blooms in lakes have been shown to be advancing due to climate change. While one species of zooplankton that feeds on the diatoms has shifted its phenology to keep pace with its food, another species has failed to shift resulting in a phenological mismatch and uncoupling of the species interaction (Winder and Schindler 2004).

Although these changes in trophic interactions between species are mostly hidden from view, changes perpetuate through the population, community, and ecosystem levels where their effects become noticeable. Detection and monitoring of changes in trophic interactions by resolving the diets of consumer species can provide early warnings of problems before they result in the loss of species and ecosystem functions (Tylianakis et al. 2008; Bartley et al. 2019).

2.2.3 Resolving trophic interactions

However, accurately resolving predator-prey interactions can be challenging (Clare et al. 2009). Although it is sometimes possible to film larger predators under water (Oehm et al. 2017), predator-prey interactions cannot usually be observed directly in freshwater systems. Most interactions occur beneath the water surface and often involve small species/life-stages that are difficult to identify without closer inspection. Resolution of freshwater trophic interactions traditionally involves evidence from morphological analysis of stomach or faecal contents and/or experimental feeding trials (Woodward et al. 2010).

Morphological identification of partially digested prey individuals is very time consuming, relies on expert knowledge of prey taxonomy and morphological diversity, and can be biased towards prey species with hard body-parts that are more difficult to digest (Thompson et al. 2012b). In addition, morphological prey identification is not suitable for all predators, for example, prey eaten using extra-oral digestion (the breakdown of dietary components outside of the predator body) (Gamboa et al. 2012), and is difficult with very small predatory species (Jo et al. 2014) and small prey items such as eggs and larvae (Taguchi et al. 2014). Experimental feeding trials can enable observation of interactions and can provide important information such as, size preferences, handling limitations etc. but the interactions observed in experimental systems may differ from those in natural systems. Furthermore, once a trophic interaction has been resolved in a study once, it is often assumed that the species will interact when they co-occur but interactions are not static and these snapshots of interactions cannot resolve how interactions change in space and time and in response to environmental change. These methodological challenges have therefore limited the resolution of freshwater trophic interactions detectable in the past.

Recent developments in molecular methods have enabled the identification of prey taxa from the DNA in dietary samples (e.g. stomach contents, regurgitates or faecal samples) (Symondson 2002; King et al. 2008; Pompanon et al. 2012). These methods enable the identification of prey from highly digested samples, providing the opportunity to discover new interactions and providing higher taxonomic resolution than was possible with traditional methods. The use of DNA-based identification to study trophic interactions has been growing rapidly, providing new insights into predator-prey interactions (Clare 2014; Evans et al. 2016). Recent studies using DNA-based identification to construct food webs have shown that, even in environments with relatively low diversity, the trophic information provided has led to fundamentally different conclusions to studies using traditional methods (Smith et al. 2011; Wirta et al. 2014). However, DNA-based analysis of diet has, so far, been applied less in freshwater than in terrestrial and marine systems (Corse et al. 2010; Roslin and Majaneva 2016; Thalinger et al. 2016). DNA-based identification can now enable trophic

interactions between species to be resolved at unprecedented taxonomic, temporal and spatial resolution, providing new information to benefit freshwater biodiversity monitoring and conservation.

2.2.4 Aims and hypotheses

The overall aim of this chapter was to present a quick-scoping (evidence-based synthesis) review of studies that have used DNA-based identification methods to resolve the diets of freshwater predators. Specific aims were to assess the current progress in diet analysis of freshwater predators, identify biases and knowledge gaps, and to make recommendations for future research. The specific hypothesis tested in this chapter was that studies using DNA-based identification of freshwater predator diets would be biased towards vertebrates, especially specific individual taxa of interest, rather than invertebrates.

2.2.5 Review methodology

In brief, studies were included that used DNA in dietary samples to analyse the natural diets of animal predators with either the predator or prey taxa inhabiting fresh water (see Appendix S1 for full methods of searches, inclusion criteria, and categorisation) and included a final total of 67 studies. Using the results of this review, current progress in the field was assessed, biases and knowledge gaps in the amassed research base were identified, and recommendations for future research were made.

The results of the quick-scoping review are presented in three sections: the DNA-based methodology that has been used to resolve freshwater predator diets; the coverage and scale of the trophic interactions investigated; and the application of DNA-based identification of freshwater predator diets for monitoring and conservation of biodiversity. The review is concluded by setting out recommendations for how this area might be further developed in order to benefit the monitoring and conservation of freshwater biodiversity in a rapidly changing world.

2.3 Methodology used to resolve freshwater predator diets

In order to gain the benefits of DNA-based resolution of predator diets, the methods used must be carefully chosen and optimised to ensure that the data produced provide the information necessary to explore specific ecological questions in freshwater ecosystems. Decisions on methodology affect what kind of data can be gained at what taxonomic resolution so there is not a 'one-size fits all' approach that can be taken for all freshwater predator diet analysis. Instead, careful planning based on what data are needed must be done first and then decisions can be made about which method(s) will best provide that data and what is possible with the available resources. The basic workflow is: dietary samples are collected, DNA is extracted from the samples, primers are used to amplify the target DNA, the species whose DNA were present in the sample are identified (see Figure 2.1 for a basic overview of the molecular methodology used in the included studies). There are three main sections of this workflow that require decisions to be made. First, the type of dietary sample that will be analysed, second, which DNA-based method (barcoding, metabarcoding or screening) will be used to analyse the samples, and finally, the markers, primers and reference sequences that will be used within the chosen method. These three groups of analytical decisions are discussed in turn using evidence collected from the quick-scoping review.

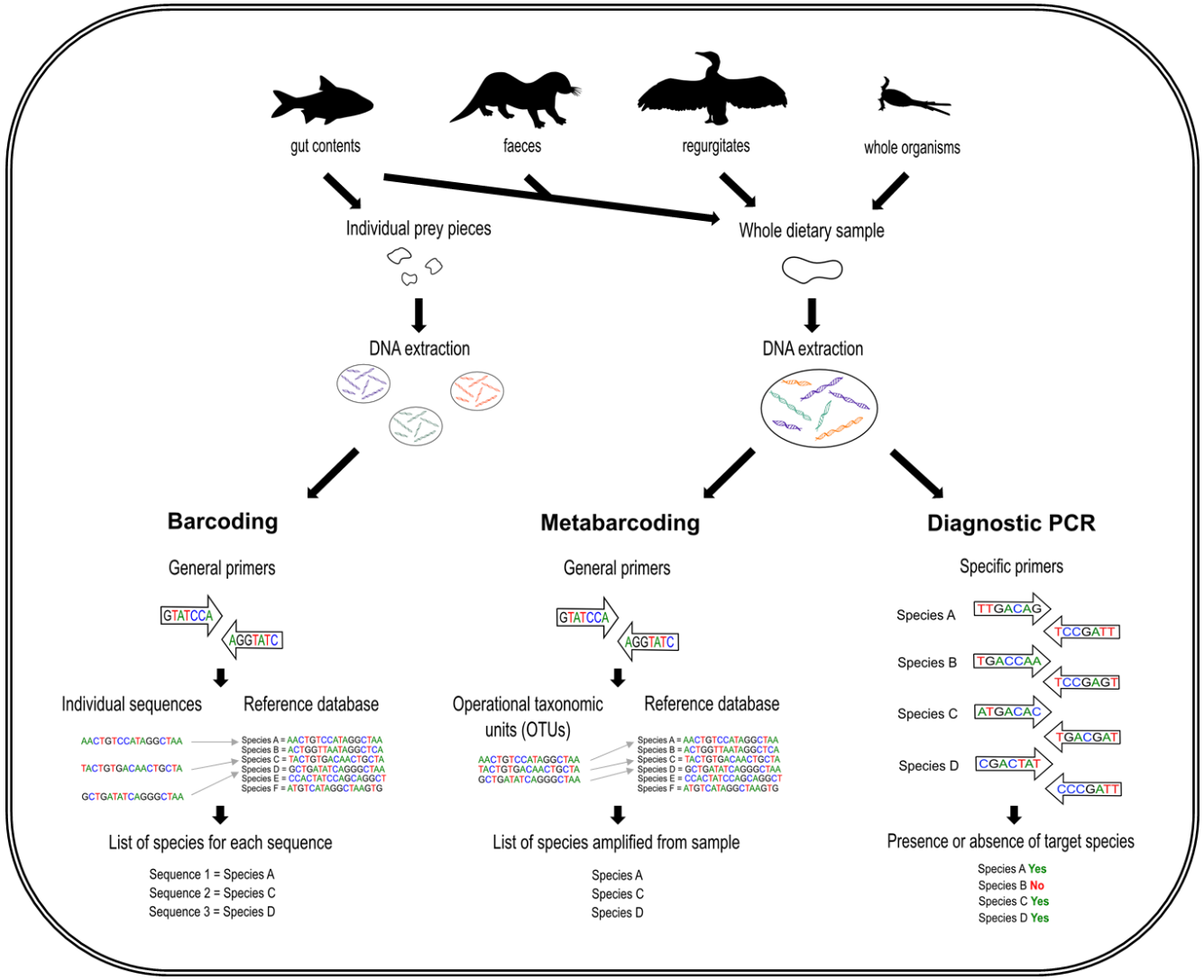

Figure 2.1 *Overview of the main methods used to analyse dietary samples of predators.*

2.3.1 Dietary sample types

The type of dietary sample used is dependent to a large extent on the taxonomic identity of the predator. Four types of samples have been used in the studies included in this review: faeces, regurgitates, gut contents, and whole organisms. The sample types might be invasive e.g. where whole organisms are used or the predator's guts are removed. These invasive sample types are sometimes opportunistically taken from animals that have died of other causes/for other purposes so are not always destructive. Samples might otherwise be non-invasive e.g. through the collection of faeces/regurgitates from predators and also through the use of gastric lavage (flushing the gut contents without the destruction of the predator). It is clearly beneficial to use non-destructive sample types where possible and where necessary due to legal protections of the predator taxa. However, analysis of faeces and regurgitates can be more challenging and limit the information on how many/which predator individuals are feeding on which prey as they are not specific to an individual.

The type of dietary sample analysed varied among taxa and studies (Figure 2.2a). Of the 67 studies included in this review, 52% used only non-invasive sample types, 45% used only invasive sample types and the remaining studies used a mixture of invasive and non-invasive sample types within the studies. The sample type varied depending on the predator taxa and the majority (90%) of studies in this review focused on vertebrate predators. Non-invasive analysis of faeces and/or regurgitates was used in all but one of the studies on bird and mammal predators. Where there is a choice of non-invasive dietary sample, it is important to know which sample will provide the most comprehensive information on diet. Oehm et al. (2017) compared faeces, pellets, and regurgitated fish samples from great cormorants and found that DNA-based analysis of pellets provided more dietary information, followed by regurgitated fish samples, then faeces. Fish predator diet studies often use invasive samples from culled individuals, but some studies used non-invasive methods (faeces (Corse et al. 2010; Guillerault et al. 2017) and gastric lavage (Kelling et al. 2016)) to obtain dietary samples from live individuals, and two studies used a combination of invasive and non-invasive sample types (Moran et al. 2016; Schmitt et al. 2017). For invertebrate predators, DNA can be extracted from

dissected gut contents (Gamboa et al. 2012; Northam et al. 2012; Pearson et al. 2018) or from the whole organism (Bradford et al. 2014). However, it is not always desirable to use invasive methods with invertebrates and one study used the faeces of damselflies for DNA-based analysis (Cheng and Lin 2016). Non-invasive dietary samples are preferable (or essential) in many cases, but these samples can be more difficult to extract DNA from and have a higher risk of contamination. Where animals are culled for other purposes, gut contents can provide samples that are less likely to be contaminated and that contain less degraded DNA than faecal or regurgitate samples. In addition, it is advantageous to use samples that can be directly linked to the predator individual so that data on number/size/age/condition of individuals can also be recorded. When faeces/regurgitates are collected from sites the predator has previously occupied, the number of predator individuals (and associated information about those individuals) is not available for analyses unless observational or genotyping data are available.

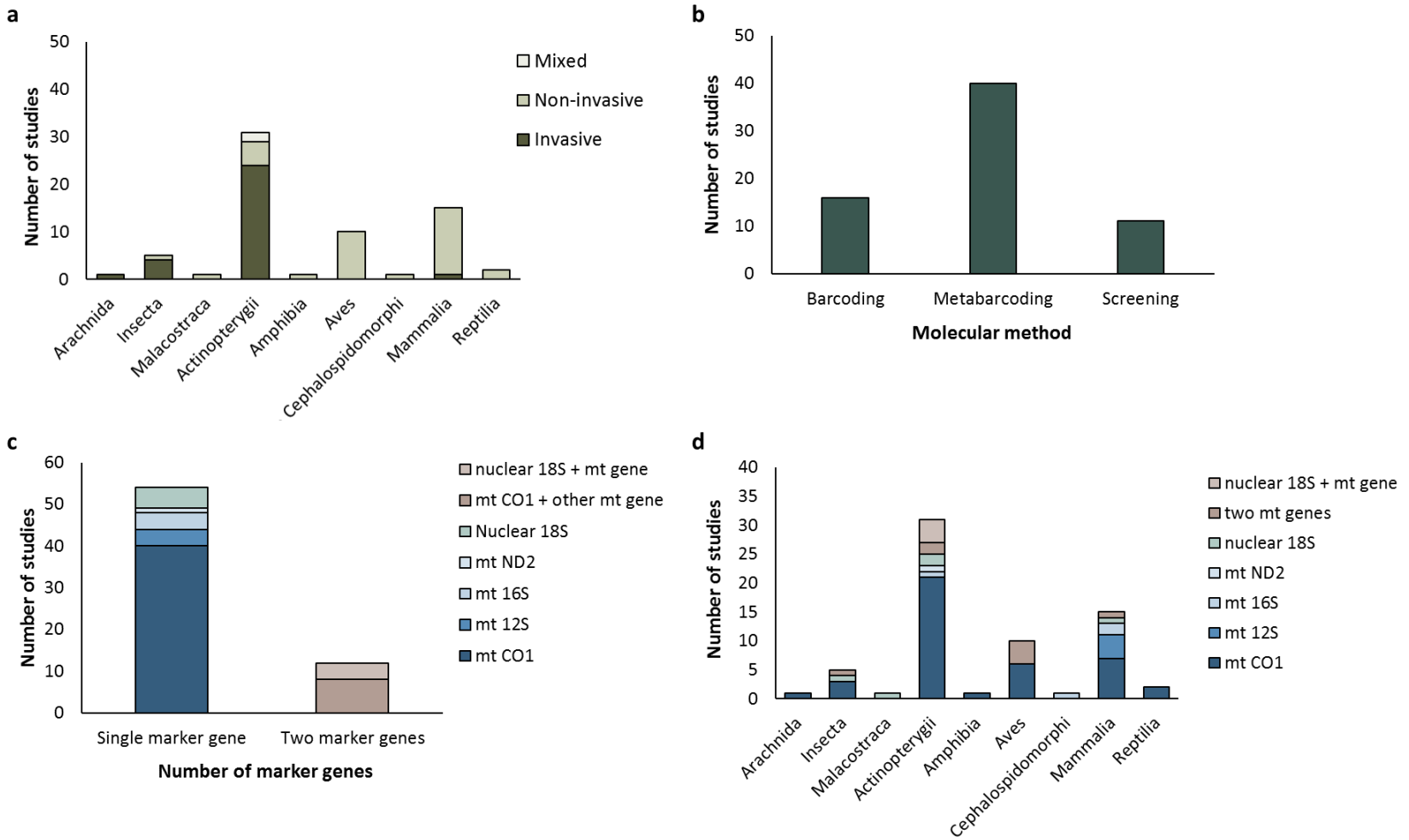

Figure 2.2 *Number of studies included in this quick-scoping review by (a) dietary sample type grouped by predator class (invasive (gut contents or whole organism), non-invasive (faeces, regurgitates) or a mixture of invasive and non-invasive methods), (b) molecular methods used to analyse dietary samples, (c) the number of marker genes used to analyse dietary samples and the specific marker genes used, (d) the marker genes used grouped by predator taxa studied.*

2.3.2 Barcoding, metabarcoding and screening

Dietary samples can be analysed using different molecular methods. For the purposes of this review the term DNA barcoding is used to describe molecular methods that involve the production of DNA sequences from single individuals to identify species. DNA metabarcoding is used to describe the production of DNA sequences from a sample that includes DNA from multiple individuals through high throughput sequencing (HTS). The term screening is used in this review to describe methods that target specific prey taxa and do not involve the generation of DNA sequences, and includes diagnostic polymerase chain reaction (diagnostic PCR) and restriction fragment length polymorphism (RFLP).

The majority of studies in this review (60%) used DNA metabarcoding methods (Figure 2.2b). Metabarcoding produces many DNA sequences from bulk samples containing a mixture of different taxa and can be used to analyse the DNA in highly digested dietary samples. A target gene is chosen and an alignment of sequences from target and non-target taxa is used to identify conserved sites and design general primers that will amplify the target prey taxa. The general primers are used in PCR to amplify a specific region of the target gene in the prey taxa of interest in the study/assessment. This results in DNA sequences from different individuals in the sample which are then often clustered together into species-like groups called operational taxonomic units (OTUs). These OTUs are then matched to sequences in a reference library in order to identify the detected taxa. Metabarcoding is increasingly used in diet analyses as it has the potential to detect the DNA of 'all' the taxa in a sample, making it possible to analyse even the broad diets of generalist species (Clare 2014). However, whether or not a species whose DNA is present in a sample is actually extracted, amplified, and assigned to the correct species is dependent on many methodological/analytical decisions.

Twenty four percent of studies included here used barcoding rather than metabarcoding (Figure 2.2b). The sequences produced have to be matched to known sequences in a reference library in order to identify which taxon each sample is from. In the barcoding-based studies included here, individual prey items were manually separated from the rest of the dietary sample to enable

separate sequencing of each item (for examples see (Boileau et al. 2015; Nelson et al. 2017)). This method is similar to how gut content analysis was done traditionally but the identification is DNA-based rather than morphologically-based providing higher accuracy and the identification of prey that might be unidentifiable morphologically due to digestion. Barcoding is low cost, per sample, compared to metabarcoding but this method will only be able to identify taxa where there are complete undigested parts remaining which might bias results against more soft-bodied/easily digested prey taxa.

The identification of the taxa in samples, analysed using both barcoding and metabarcoding, is dependent on the ability to match the detected sequence/OTUs to sequences in reference databases. The taxonomic resolution of diet analysis using these methods is therefore dependent on having a high-quality, comprehensive reference sequence database that includes the potential prey taxa of interest. Gaps in reference databases can severely limit the successful application of these methods, with some metabarcoding studies unable to assign up to 40% of the obtained sequences (Yang et al. 2017a). Several studies state comprehensiveness of reference databases as a limitation to identification of prey taxa (e.g. Coissac et al. 2012; Leray and Knowlton 2015; Elbrecht et al. 2016; Yang et al. 2017). However, the limitation imposed by incomplete reference databases will decrease in the future as more species barcodes are produced. Any OTUs from current studies that have not been able to be assigned to species will be able to be revisited in the future in order to identify them using more complete reference databases. In the meantime, when species diversity is relatively low, the construction of indigenous species barcode databases for particular studies can improve the identification of species in metabarcoding (Yang et al. 2017a). One freshwater predator diet study included in this review focuses on the construction of a species barcode database for use in fish conservation (Hardy et al. 2011). This kind of reference database construction for freshwater taxa is key to enable DNA-based analysis of predator diets to be used to its full potential.

An alternative method of DNA-based diet analysis is screening. Eleven studies included in this review (16%) used screening methods rather than barcoding or metabarcoding methods to analyse predator diets (Figure 2.2b). All but one of

these studies used diagnostic PCR to screen samples for specific prey taxa. In this method, primers are designed and optimised to enable the detection of specific taxa rather than many different taxa. Amplification of the specific DNA sequences indicates the presence of those taxa in the dietary sample. Therefore, amplified sequences do not need to be matched in reference databases. Diagnostic PCR results in presence or absence data for each targeted species. The majority of studies that used diagnostic PCR used standard/conventional PCR rather than quantitative PCR (qPCR). Quantifying the biomass or number of prey individuals consumed using PCR is complex (due to many differences e.g. in the size/stage of prey, time since consumption, DNA copy number between species, and primer efficiencies) but qPCR can provide improved sensitivity and/or specificity in addition to quantification (King et al. 2008) and therefore can offer benefits for diet analysis. Only one study included in this review used qPCR, which they used to study the impact of predators on juvenile salmon (Michel et al. 2018). One study (Nelson et al. 2017) in this review used RFLP to screen for a particular prey species rather than diagnostic PCR. This method uses a restriction enzyme that will cleave a particular point in the sequence of the target species but not in sequences of other species likely to be in the sample. The presence of the shorter, cleaved, fragments indicates the presence of the target species in the sample. Screening methods provide a cost-effective method to screen dietary samples for targeted prey taxa.

Although metabarcoding has the potential to amplify all the OTUs in a dietary sample of mixed taxa whereas screening methods can only identify targeted species, the latter might sometimes be the preferred choice for analysing dietary samples. Dietary samples contain different numbers of prey taxa, which can affect the consistency of detection using metabarcoding. In addition, dietary samples contain large amounts of the consumer's own DNA, which can swamp the sequence reads when analysed using metabarcoding if the chosen primers also amplify DNA from the predator taxon. Although the problem of predator DNA can be resolved by using blocking primers, they require careful design and testing and they are not always a suitable option if the predator and prey species are closely related (Piñol et al. 2014). Screening methods do not

involve amplification of the predator's DNA because the primers are specific to the targeted prey taxa and detection is less dependent on the amount of other DNA present. It should be noted that the neither of these approaches would enable detection of cannibalism in predator diets which is common in freshwater food webs (Boukal 2014).

A recent study comparing diagnostic PCR and metabarcoding for dietary analysis found that diagnostic PCR gave more consistent results than metabarcoding (Rennstam Rubbmark et al. 2019). As diagnostic PCR and RFLP tests screen samples for the presence of targeted prey sequences, they can only be used when the potential diet is known *a priori* or when only specific prey species are of interest. However, Nielsen et al. (2018) recommend that the consumer's potential diet should be known *a priori* in all studies as it is usually not possible to get a complete diet assessment, or choose the best method or reference library, without some knowledge of the consumer's feeding behaviour and the available resources in the habitat. If the prey species of interest are known *a priori*, if the potential prey diversity is relatively low, or if only specific prey taxa are of interest then screening methods might offer a more consistent and cost-effective method of diet analysis (Nelson et al. 2017; Rennstam Rubbmark et al. 2019).

2.3.3 Markers, primers and reference sequences

For all of these methods it is important that genetic markers are chosen with regards to the ecological question (King et al. 2008; Pompanon et al. 2012). As the studies included in this review focus on animal prey species, the marker genes used are eukaryotic nuclear and mitochondrial genes. Higher variation in mitochondrial genes provides better taxonomic resolution but lower taxonomic coverage than nuclear genes (Deagle et al. 2014). Mitochondrial genes are often chosen for diet analysis due to the high copy number per cell allowing greater sensitivity and the availability of a large number of published primer sets for these genes (King et al 2008). The majority of studies in this review used a single marker gene and most of these studies (61%) used only the mitochondrial gene cytochrome c oxidase I (COI, Figure 2.2c). This gene is a popular choice in animal studies as it was suggested to be able to form the core of a global

bioidentification system for animals (Hebert et al. 2003b) and taxon barcodes for this marker have been continually added to a curated reference database (Barcode of Life Data Systems (BOLD) (Ratnasingham and Hebert 2007)). An additional benefit of using a protein-coding gene, such as COI, is that it enables better identification and removal of errors/non-target sequences (e.g. pseudogenes, insertions and deletions) (Yu et al. 2012). However, a lack of conserved primer binding sites can limit amplification success and the 16S rRNA gene has been proposed as an alternative to provide species-level resolution and broad amplification (Clarke et al. 2014, Deagle et al. 2014). Only 6% of studies reviewed here used only the 16S rRNA gene. Eighteen percent of studies used a combination of two different marker genes. The use of multiple markers has been recommended in order to maximise amplification success across taxa as well as maximising taxonomic resolution (Pompanon et al. 2012; Taberlet et al. 2012) but causes an increase in cost and time (Zhang et al. 2018) and might not be necessary if the diversity of the target group is low. Most of the studies using multiple markers used COI with another mitochondrial gene (8 studies) but four studies on fish used a combination of nuclear 18S with either mitochondrial COI or 12S.

For most classes, there is little variation in marker choice but classes with higher numbers of dietary studies show more variation in the choice of single marker and include studies that use multiple markers (Figure 2.2d). Studies on fish diet show the greatest variation in marker choice with the majority of studies using COI on its own but four studies using different single markers and six studies using multiple markers. Studies on mammal diets show a similar pattern with the majority using COI on its own, seven studies using other single markers and two studies using multiple markers.

The best choice of marker is dependent on the potential prey of the target predator. For piscivorous fish, birds, mammals the marker(s) chosen need(s) to be suitable for amplifying from one taxonomic group. However, the prey of some generalist predators might include multiple different taxonomic groups. Knowledge of the potential prey is vital in choosing which and how many markers are necessary in order to obtain a true representation of diet and known limitations in the chosen markers to amplify other possible prey taxa

should be clearly stated to avoid a lack of amplification being assumed to be a lack of interaction.

Knowledge of potential prey enables assessment of which markers and primers will provide the best coverage and resolution of the target prey taxa. These decisions need to take into account which marker might offer the desired level of taxonomic resolution, how comprehensive and accurate reference databases for the marker are, and whether primers have been designed and well-validated for the target taxa. For example, although there has been a dedicated campaign to construct the Fish Barcode of Life (FISH-BOL) for COI (Ward et al. 2009), previous work on fish taxa has meant that some reference databases for other marker genes might be more complete for particular fish species than for COI at present (Hardy et al. 2011). Previous work to validate primers and construct reference databases enables future studies to use the same markers and primers with confidence. However, optimisation and validation of new primers might enable broader taxonomic amplification or better taxonomic resolution of the target prey and so studies focusing on this are essential for the application of DNA-based resolution of freshwater predator diets.

2.3.4 Optimisation of molecular methods

It is vital that DNA-based methods are carefully optimised for the system and question of study (Elbrecht and Leese 2017). If this is not the case, it could lead to biased and/or incomplete results. Within the studies included in this review, eight specifically developed DNA-based methods for particular taxa or systems. Four of these studies developed methods for detecting fish prey in piscivorous fish, mammals or birds (Carreon-Martinez et al. 2011; Moran et al. 2016; Thalinger et al. 2016; Oehm et al. 2017). The focused attention given to detecting fish species in dietary samples provides the necessary tools to enable ecological studies on piscivorous predators. Where optimisation work has not been done previously for the target taxa, initial method development is necessary to enable meaningful ecological inferences to be made. Three of the studies in this group each developed DNA-based methods for more challenging prey groups: invertebrate prey (Corse et al. 2010); a range of vertebrate prey using a blocking primer for the vertebrate predator (Kumari et al. 2019); and a

wide range of vertebrate and invertebrate prey to assess the diet of a reptilian predator (Ducotterd et al. 2021).

Finally, method development to allow dietary analysis of a semi-aquatic mammal, the Pyrenean desman, from faeces along streams (Gillet et al. 2015) provides a good example of how initial optimisation facilitates future studies: resolving the diet, comparing the trophic overlap between the Pyrenean desman and the Eurasian water shrew, and developing a bioinformatics pipeline (Biffi et al. 2017a; Biffi et al. 2017b; Hawlitschek et al. 2018).

2.4 Coverage and scale of trophic interactions

2.4.1 Taxonomic coverage

This review found that, to date, DNA-based methods have been used to study the diets of freshwater predators from nine animal classes (Figure 2.3a). Currently published studies are biased towards vertebrate predators (90%) over invertebrate predators and there have been a particularly large number of studies focused on fish predators compared to any other class of animals (46%) (Figure 2.3a). While the number of studies published per year on freshwater fish diet using DNA-based methods has remained fairly steady since 2014, numbers of studies on other vertebrate predators have been increasing. Indeed, the first DNA-based studies analysing the diets of amphibians and reptiles have only been published in the last two years suggesting that DNA-based methods are only just beginning to be used to study diets in other groups in fresh waters.

Very few studies have so far used DNA-based methods to analyse the diets of freshwater invertebrates. The invertebrate studies reviewed here include five insect predator studies, one arachnid study, and one malacostracan study. In contrast to the vertebrate studies, where there are often several studies on the same fish, bird or mammal species, each of the invertebrate studies represents the only example of the use of DNA-based methods to analyse the diet of the study species. In fact, each of the studies represents the only example for each of these insect orders and the only example for the classes of Malacostraca and Arachnida. As described above, the first studies on new freshwater taxa often include considerable method development to find the best methods for the particular sample type and diet of the study species. As methods are developed for more freshwater taxa, it becomes easier to use the methods to study ecological questions about those taxa, enabling the benefits of DNA-based methods for freshwater trophic interactions to be realised.

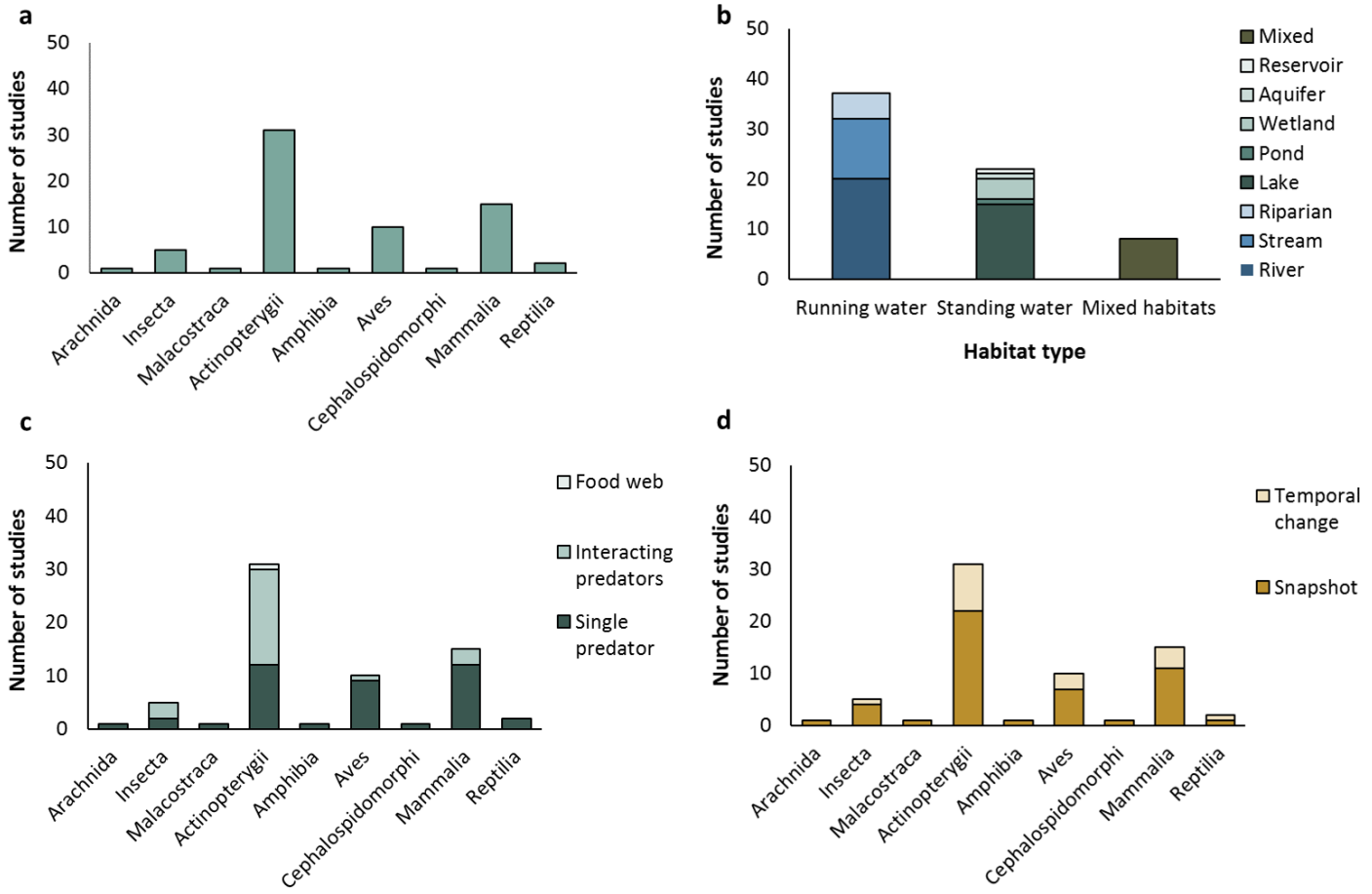

Figure 2.3 *Number of studies included in this quick-scoping review by (a) different classes of animal predators studied. (b) different freshwater habitats studied. (c) the trophic scale of the study (single predator taxon, interacting predator taxa, food web structure) grouped by predator class, (d) temporal resolution of the study (snapshot (single time point or composite snapshot, combining dietary data from different time points) or change in diet over time) grouped by predator class.*

2.4.2 Habitat coverage

The majority of DNA-based freshwater predator diet studies have so far focused on species living in and around running water habitats (55%), with the majority of those studies on predators inhabiting rivers (54% of running water studies) (Figure 2.3b). Five studies focus on the riparian habitat with predators feeding on freshwater taxa for all or part of their diet. Studies of predators in standing waters (33% of studies overall) are mostly focused on lake habitats (68% of standing water studies), studies on wetlands (18%) are showing a recent increase, and three other standing water habitats (ponds, aquifers, and reservoirs) are each represented by only a single study. Eight studies (12% of all studies) have looked at predators that forage across mixed habitat types. These studies include predators that forage between different types of freshwater habitats as well as predators that forage at the boundary between freshwater and coastal environments (i.e. species of birds and bats). At present, the habitat type studied appears to largely be a consequence of the habitat used by particular taxa of interest rather than the targeting of specific habitats. However, as DNA-based methods become more widely used to study freshwater predator diets, it will provide the opportunity to explore trophic interactions across different freshwater systems.

2.4.3 Trophic scale

In contrast to the findings of a review focused on parasite interactions and large vertebrate interactions (Clare 2014), the trophic scale of studies in freshwater systems is largely restricted to analyses of the diet of either two or more predator taxa within the same system, or single predator taxa (Figure 2.3c). Where multiple predator taxa are studied, the number of taxa ranges from two to twenty-seven, showing a scaling up of the number of predator diets resolved within a study system. However, only one study (Bartley et al. 2015) focused on multiple predators with the specific aim of resolving network structure as opposed to selected predator-prey interactions. The studies of higher numbers of predator taxa (more than 3 taxa) all focus on fish predators. The larger number of fish studies in total, and the higher numbers of predator taxa targeted within these studies, might partly reflect the relative ease of obtaining dietary samples of many fish species compared to the challenges in obtaining

samples from other taxa due to protection of some species or challenging sample collection. In addition, the optimisation of DNA-based methods for detecting fish means that analysis of piscivorous fish diets can be done relatively easily without needing to first optimise all the methods. While studies on the diets of single predator taxa can be very informative when a particular species is of interest or concern, studies focusing on multiple predator taxa within the same system can provide much more information. Studies on multiple interacting predator taxa are beneficial in that they not only provide a broader understanding of the factors affecting individual species, but also enable the study of aspects of ecological network structure that underpin ecosystem state and stability.

2.4.4 Spatio-temporal scale

Studies on trophic interactions often provide a snapshot of diet at either one point in time or by combining dietary data from different time points to give a composite snapshot of diet for a particular time period. A large proportion (73%) of DNA-based studies of freshwater predator diet included here provide this kind of dietary snapshot (Figure 2.3d). However, trophic interactions are not static, but show temporal and spatial variability (Berlow et al. 2004; Boukal 2014). It is this variability that could provide sensitive indications of the impacts of environmental pressures and "early warnings" of impending biodiversity change.

Snapshots from a single time point are useful for method development studies and when comparing to other species at the same time point, but might not be representative of the diet at other times and places and should therefore be treated with some caution. Composite snapshots provide a more comprehensive analysis of the overall diet of a species but, unless the temporal and spatial variation in diet is explicitly reported, this pooling of dietary information can mask ecologically important changes and over- and/or under-emphasise the importance of different prey resources. DNA-based methods enable high temporal and spatial resolution of diet due to the ability to quickly process large numbers of samples from different times and places compared to morphological analysis. This reduction in the time necessary to process samples

is one of the major benefits of DNA-based analysis of diet and enables changes in diet to be assessed over time and across systems. The eighteen studies reviewed here that assess diet at different time points are on a variety of taxa (Figure 2.3d) and thirteen of these studies use the information for applied purposes (see next section for details of applied purposes). The use of DNA-based methods to analyse predator diets over time and across systems will be particularly effective in studies assessing the effects of environmental change on freshwater species.

2.5 Application for biodiversity monitoring and conservation

The studies included in this review have a wide range of ecological motivations. It is beyond the scope of this review to discuss all of these areas in detail, and many studies focus on a combination of areas. However, this section summarises key areas that are beginning to benefit from DNA-based analysis of predator diets. In total, 66% of the studies included in this review used DNA-based methods to resolve freshwater predator diets for applied uses. The majority of these applications are to facilitate conservation and the management of non-native species but small numbers of studies show how these methods can benefit other applied areas as well (Figure 2.4).

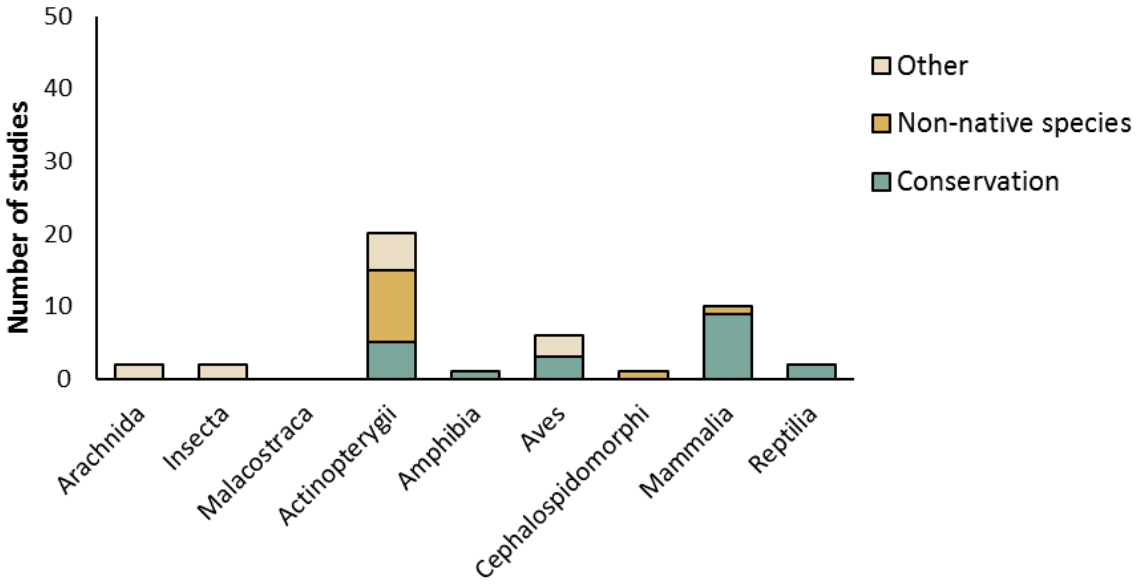

Figure 2.4 *Number of studies included in this quick-scoping review that have an applied focus (e.g. conservation, assessment of non-native species) grouped by predator class.*

2.5.1 Increased taxonomic resolution

Clear motivation for using DNA-based methods to study the diet of any freshwater predator of interest is the increased taxonomic resolution in comparison with methods used previously such as the ability to obtain species-level dietary information (e.g. Vesterinen et al. 2013). When DNA- and morphologically-based analyses of predator diets are compared, DNA-based analyses detected more species of prey than morphological analyses (Carreon-Martinez et al. 2011; Bartley et al. 2015; Moran et al. 2016; Thalinger et al. 2016; Oehm et al. 2017). This higher resolution dietary information can provide new insights into many different areas of ecology including: foraging behaviour (Arroyave and Stiassny 2014; Boileau et al. 2015; Lu et al. 2016; Kovac et al. 2019); spatial and temporal changes in diet (Clare et al. 2011); ontogenetic dietary shift (Jo et al. 2014); sex-specific prey choice (Thalinger et al. 2018) (Thalinger et al. 2018); trophic niche partitioning (Cheng and Lin 2016); speciation (Bradford et al. 2014); and food web structure (Bartley et al. 2015). These new insights can provide benefits to the monitoring and conservation of many species.

2.5.2 Conservation

Analysing diet to understand the ecology of species of conservation concern is the focus of twenty of the studies in this review. Research on birds (three studies) and mammals (nine studies) used DNA-based methods to analyse the diet of the species of concern. In these studies, DNA-based methods yielded new information about species' diets, including the use of prey from different systems, (Gerwing et al. 2016; Trevelline et al. 2016), and the comparison of diets between species, enabling trophic partitioning and overlap to be examined (Biffi et al. 2017a; Trevelline et al. 2018a). Investigations motivated by the conservation of fish species (five studies) use DNA-based methods to detect predation of specific prey of conservation concern in the diets of many predator taxa (Waraniak et al. 2018a; Waraniak et al. 2018b; Waraniak et al. 2019; Bunch et al. 2021). These studies show the potential for DNA-based analysis of predator diets to benefit conservation through the ability to assess the diets of species of concern and predation pressures on these species.

2.5.3 Non-native species detection/monitoring

The use of DNA-based methods in dietary analyses can provide new understanding of the effects of non-native species on communities. Twelve studies in this review use DNA-based methods to explore interactions involving non-native species. All but one of these studies focuses on non-native fish taxa. DNA-based analysis of non-native predatory fish diets enabled a much larger number of prey species to be identified and also revealed selective predation that showed which non-native predator was likely to have a greater impact on prey species of concern (Moran et al. 2016; Schmitt et al. 2017; Schmitt et al. 2019). DNA-based methods have also been used to explore the effects of non-native prey species on native predators important to fisheries, finding evidence that predation on the non-native prey provided benefits to the native fish species studied (Nelson et al. 2017). DNA-based analysis of the diets of non-native species can provide important information that can help to predict impacts on native communities and monitor the new interactions that occur as species establish in new habitats.

2.5.4 Other applications

Predators can be sensitive indicators of environmental change (Velarde et al. 2013) and are structurally important in ecosystems (Estes et al. 2011). DNA-based analysis of predator diets provides new opportunities to understand and monitor changes. A small number of very recent studies have used DNA-based analysis of predator diet to explore the effects of environmental/anthropogenic changes in freshwater habitats (Pearson et al. 2018; Trevelline et al. 2018b; Jo et al. 2019). A study using DNA-based methods to measure changes before and after dam construction in rivers in South Korea found that the variation of largemouth bass (*Micropterus salmoides*) diet decreased after dam construction, suggesting a change in foraging behaviour due to the change in habitat (Jo et al. 2019). DNA-based analysis of the diet of the Louisiana waterthrush (*Parkesia motacilla*), showed that the reduction in preferred aquatic prey taxa, due to stream acidification, caused an increase in dietary richness and diet breadth through the inclusion of more terrestrial prey (Trevelline et al. 2018b).

The analysis of freshwater predator diets can also be a useful additional tool in assessing biodiversity. While eDNA from the water and sediment can provide high resolution biodiversity assessments, some taxa can be missed. Predator diets might provide useful additional sampling tools for assessing biodiversity as they forage in different micro-habitats and might consume prey that shed less DNA into the water making them more difficult to detect. As such, predator diets may uncover "hidden biodiversity" in fresh waters (Jo et al. 2016). Studies using diet analysis to assess terrestrial mammal biodiversity have shown increased diversity (Schnell et al. 2012; Meyer et al. 2020) and this is potentially an overlooked benefit of DNA-based analysis of diet to freshwater biodiversity assessment.

A small number of studies included in this review used DNA-based identification of predator diets for other purposes. Three studies have used these methods to determine factors affecting species important in fisheries management, such as predation, competition, and the effect of turbidity on predation (Carreon-Martinez et al. 2014; Kelling et al. 2016; Nelson et al. 2017). DNA-based identification of predator diet also enabled the importance of freshwater bodies for house-farm swiftlets (*Aerodramus* sp.) in urban farms to be revealed (Chan et al. 2019); the resolution of the diet of the creeping water bug (*Naucoris* sp.) which use extra-oral digestion and are potentially involved in disease transmission (Gamboa et al. 2012); and the testing of whether bioaccumulation of mercury in songbirds' nests might be via predation on emerging mayflies by wolf spiders (Lycosidae) (Northam et al. 2012).

The variety of applications for DNA-based analysis of predator diets shown here suggests that these methods are just beginning to be used for applied purposes in freshwaters and a wide variety of areas might benefit from these methods in the future.

2.6 Conclusions and recommendations

2.6.1 Conclusions

DNA-based identification of freshwater predator diets offers the potential for trophic interactions to be resolved at unprecedented taxonomic, temporal and spatial resolution, providing new data for biodiversity monitoring and conservation. Although it has been used more in terrestrial and marine systems previously (Corse et al. 2010; Roslin and Majaneva 2016; Thalinger et al. 2016), it is increasingly being applied in freshwater systems. This review aimed to assess the use of DNA-based diet analysis in freshwater ecosystems, focusing on the DNA-based methodology used; the coverage and scale of the trophic interactions investigated; and the application for monitoring and conservation. The following conclusions, biases and areas for further work were drawn from the assessed studies.

DNA-based identification of freshwater predator diets is already beginning to benefit a wide range of ecological areas that could enable better monitoring and conservation of freshwater biodiversity. These methods are already frequently used to benefit the monitoring of non-native species and conservation and are beginning to be applied to a broader taxonomic range. There is a current bias towards vertebrates (especially fish or piscivorous birds and mammals) and towards running water systems. However, more recent studies are showing the methods are beginning to be used to analyse invertebrate diets and taxa in a wider variety of habitat types.

Studies on single predator taxa and snapshots of diet in time and space are useful (and essential) to answer specific questions and optimise methods for particular taxa and environments. However, the real power of DNA-based predator diet analysis is in being able to benefit from the ability to process large numbers of samples from multiple individuals, multiple taxa, and multiple points in time and space (due to the low cost and efficiency of DNA-based methods in comparison with traditional methods). Individual-level data enable data collection on the strength of interactions between taxa as well as the presence of an interaction. Changes in the strength of an interaction (due to

changes in abundance/habitat use etc.) will occur before the interaction disappears altogether and so provide an early-warning signal of change in a community. Resolving the diet of multiple taxa enables data on interactions between predator and prey taxa within a community to be assessed and monitored as a food-web module. These modules provide a clearer picture of change within a community than the monitoring of any single taxon.

In order to monitor change in biodiversity it is essential that data are collected over multiple time points so we can see how diets shift in response to environmental pressures and not simply a composite picture of the complete diet of a taxon. When predator diet data are collected over these larger scales, it enables the quality and breadth of ecological analysis to be much higher including, at larger trophic scales, metrics of ecosystem stability.

Being able to gain high-quality dietary data over time depends on thorough optimisation and validation of the methods. Choosing markers and primers that provide the desired breadth and resolution is essential. Although it is unlikely that any single choice will be perfect for a potential prey group, there is value in the DNA-based identification community finding/choosing a core marker for particular taxonomic groups as this then enables further optimisation and reference database construction that will help provide higher-quality data in the future. The high reference database coverage for COI is a strong motivation for choosing to use this marker in metabarcoding studies and enables the design of more primers for target taxa. Increased effort in improving reference database coverage for other mitochondrial genes (i.e. 12S and 16S rRNA genes) would enable better comparisons between markers and allow taxonomic resolution and amplification success to be assessed for different potential prey groups. Additional markers could then be used if needed to increase the taxonomic breadth/resolution.

To monitor change over time it is essential that the same markers/primers are used and that any biases are known. Changes in reference databases over time will lead to higher taxonomic resolution in barcoding/metabarcoding data but these changes can be accounted for by repeating the taxonomic assignment of sequences from all time points with the most up-to-date reference databases.

Comprehensive, high-quality reference databases are key to all the DNA-based methods discussed in this review. Barcoding and metabarcoding depend on matching amplified sequences to identify what taxa are present in each sample. Specific primers for diagnostic PCR can only be designed if reference sequences for the potential prey are present in reference databases. The building and curation of reference databases is key to enabling the identification of prey from different taxonomic groups and the highest level of taxonomic resolution possible.

2.6.2 Recommendations

Continued building and curation of comprehensive, high-quality reference databases for freshwater biodiversity is vital for DNA-based resolution of predator diets both for the accuracy of taxonomic assignment and for the development of new primers to target specific species or groups of potential prey. This is especially true for freshwater invertebrates which are key components of freshwater food webs but currently underrepresented in studies using DNA-based identification of predator diets.

Optimisation and validation of DNA-based methodology for different taxonomic groups of prey in freshwater would enable more widespread application. The development and validation that has been done for detecting fish needs to be replicated for other taxonomic groups such as crustaceans, molluscs and insects. Prior knowledge of potential prey communities in different freshwater habitat types is essential for the development of methods for DNA-based dietary analyses.

Scaling up from single predators to food web modules and to increased spatial and temporal scales will enable DNA-based interaction data to build on the extensive work on freshwater food webs and generate new high-resolution, individual-level data that can provide a better understanding of freshwater ecosystems and potentially provide new metrics for monitoring changes in communities.

3 DNA barcodes for UK freshwater arthropods: coverage, curation and priorities for the future

3.1 Summary

Accurate DNA-based identification depends on high-quality reference sequences for the species of interest. Gaps in reference databases can limit the identification of specimens. This study assessed the coverage of UK freshwater arthropods in the Barcode of Life Data System (BOLD) including protected and non-native species. The quality of the stored DNA sequences was then analysed to assess any ramifications for accurate species identification. Geographic variation in sequences was also analysed to assess whether the origin of barcoded specimens is important in accurate taxonomic assignment.

This study shows a total of 60% of UK freshwater arthropod species are represented by publicly available reference sequences in BOLD. Representation is biased toward the classes: Malacostraca, Branchiopoda and Insecta, and protected and non-native species are covered more fully than other taxa. Stored sequences include misidentifications and errors causing a lack of barcoding gap for some species which can prevent accurate identification. Species within this study showed high intraspecific geographic variation suggesting that the origin of barcoded specimens is important for accurate taxonomic assignment. Only 5% of UK freshwater arthropod species are represented by sequences from UK specimens.

The accurate identification of UK freshwater arthropods using DNA-based methods requires that gaps in reference sequences are filled particularly for species that are not represented by any sequences. In addition, it is important that more specimens, from multiple locations, are barcoded so that intraspecific variation is more fully represented. In species where multiple sequences are stored, analysis and curation of sequences can improve the accuracy and resolution of taxonomic identification while preventing misidentification or restricting analysis to higher taxonomic levels.

3.2 Introduction

3.2.1 DNA-based identification of freshwater biodiversity

Declines in biodiversity are thought to be more rapid in freshwater ecosystems than in terrestrial and marine systems (Grooten and Almond 2018; Tickner et al. 2020). Assessment and monitoring of freshwater biodiversity is, therefore, vital in detecting the responses of freshwater species to threats such as climate change, invasive species and pollutants, and in prompting management and conservation (Lawson Handley 2015).

Invertebrate communities form the basis of many biomonitoring programmes for assessing ecological quality. However, current gaps and biases in our knowledge of the distribution and status of freshwater biodiversity are hindering this effort (Darwall et al. 2011; Di Marco et al. 2017). For effective biodiversity management, we need to be able to characterise the richness and composition of whole communities and also monitor particular species more closely, for example those that are of conservation concern or are invasive. This requirement represents a key challenge for traditional approaches to assessment and monitoring.

DNA-based identification can provide a more efficient approach to species identification in freshwater habitats compared to morphological biodiversity assessment. Morphological assessment is time-consuming and dependent on taxonomic expertise (Reid et al. 2019) and a decrease in the number of qualified taxonomists is likely to reduce capacity further (Hopkins and Freckleton 2002; Thomsen et al. 2012). The main DNA-based methods used for biodiversity assessment and monitoring are DNA metabarcoding and screening methods such as qPCR (quantitative polymerase chain reaction). DNA metabarcoding uses general primers (short oligonucleotides that bind to particular genetic sequences) designed to amplify DNA from a wide range of target taxa (e.g. macroinvertebrates) using high throughput sequencing (HTS). Diagnostic PCR uses specific primers that are designed to only amplify the DNA of the single target species or genus.

These DNA-based identification methods have high detection capability and reduced costs compared to morphological methods and can also enable the

detection of species that might otherwise be missed due to identification challenges or constrains on sampling effort – notably rare species, cryptic species or larval stages (Cristescu and Hebert 2018). With careful methodological optimisation, DNA-based methods can deliver important data on both whole communities and focal species for the assessment of biodiversity and ecological state.

Development and optimisation of methods for vertebrate taxa has led to successful application of DNA-based identification methods for biodiversity assessment and monitoring of some vertebrate groups. For example, screening for the presence of great crested newts (*Triturus cristatus*) using diagnostic PCR now provides more effective monitoring than standard methods (Rees et al. 2014; Biggs et al. 2015), and metabarcoding of environmental DNA (eDNA) from water samples currently provides an efficient, non-invasive and cost-effective method of surveying freshwater fish species composition in standing and flowing waters (Lawson Handley et al. 2019; Di Muri et al. 2020).

3.2.2 Reference databases

The success of DNA-based identification is dependent on the relevant DNA sequences of target species being present in reference databases. With metabarcoding, identification of the taxa in the sample depends on matching the amplified DNA sequences to the sequences in barcode reference databases. In order to design specific primers for diagnostic PCR, the sequences of the target species and the sequences of related non-target species are needed to ensure that the primers will amplify only the DNA of target species and not DNA from other taxa in the sample. Gaps in reference databases can limit our ability to design and validate primers and can cause severe limitations to the identification of specimens in biodiversity assessments (Leray and Knowlton 2015). However, to date, reference databases have not been constructed in a systematic way, resulting in biased coverage, which has yet to be formally quantified for many taxonomic groups (Weigand et al. 2019).

3.2.3 DNA-based identification of freshwater arthropods

The freshwater taxa targeted so far in studies using DNA-based identification methods do not reflect the relative occurrence or abundance of animals in fresh

waters (Belle et al. 2019a). Although freshwater arthropods are extremely diverse, comprising 74% of UK freshwater animals (UKCEH UK Checklist of Freshwater Animals: Gunn et al. 2018); functionally important in freshwater ecosystems (Dudgeon et al. 2006); important in water quality assessment schemes (e.g. the Water Framework Directive (WFD); and as citizen science tools (e.g. The Riverfly Partnership), they are under-represented in eDNA studies (Belle et al. 2019a; Blackman et al. 2019).

Optimising DNA-based identification methods for freshwater arthropods is more challenging than for vertebrates due to their very high diversity. In addition, arthropods might shed less DNA into the environment than vertebrates or other invertebrates such as molluscs (Tréguier et al. 2014; Harper et al. 2020) making their detection in eDNA samples more difficult. Many recent studies have focused on improving the methodology for metabarcoding bulk samples of macroinvertebrates for ecological quality assessments (Elbrecht et al. 2017; Elbrecht and Leese 2017; Pereira-da-Conceicoa et al. 2021). In addition, newly-designed primers, that minimize amplification of non-target DNA (e.g. fungi, algae, bacteria), (Leese et al. 2021) will enable more efficient detection of macroinvertebrates in eDNA samples. However, to date, less attention has been given to the development of reference databases for this phylum. As a result, incomplete reference databases will mean that many of the sequences that are derived from metabarcoding of eDNA or bulk samples will not be assigned to species, reducing the utility of these surveys for biodiversity and ecological quality assessments. In order to use DNA-based identification methods to assess and monitor UK freshwater arthropods for biodiversity and ecological quality assessments, it is vital that comprehensive reference databases are available.

3.2.4 United Kingdom freshwater arthropod reference sequences

The United Kingdom (UK) provides an excellent case study for exploring the coverage of freshwater arthropod species in reference databases. There is a long history of interest in freshwater groups and biomonitoring in the UK and there is a large community of amateur and professional recorders providing data using traditional methods. There are also clear ambitions to improve the environment which will require better species data (e.g. A Green Future: 25 Year Environment Plan to Improve the Environment, HM Government (2018)).

The coverage of UK freshwater arthropod species in reference databases has yet to be formally quantified. Understanding the existing coverage will enable gaps to be filled through prioritisation of these species in sequencing projects. Several projects aim to tackle this issue. The FreshBase project (https://freshbase.myspecies.info/) aims to create a modern collection of expertly identified freshwater invertebrates preserved for genomic analysis. UKBOL (https://ukbol.org) and BIOSCAN Europe (https://www.bioscaneurope.org/) are both part of the International Barcode of Life Consortium (iBOL) (https://ibol.org/) and aim to assemble and curate barcodes for UK and European species. The Darwin Tree of Life project (https://www.darwintreeoflife.org/) aims to sequence the genomes of all the eukaryotic species in Britain and Ireland.

While several genes are currently used for DNA-based identification, a region of the cytochrome *c* oxidase I (COI) mitochondrial gene is the main marker used for animal barcoding and one of the main reference databases for animal barcodes is the Barcode of Life Data System (BOLD: http://www.boldsystems.org/). BOLD is a collaborative public resource currently (at time of writing) containing barcodes for 227,000 animal species. Prioritising the sequencing of UK freshwater arthropod species that are not represented in BOLD would enable more complete identification of UK freshwater biodiversity using metabarcoding and would enable the future development of primers to target specific arthropod groups or species.

Accurate DNA-based identification depends not only on sequences of the target taxa being present in reference databases but also on the quality of the stored sequences. Low quality or misidentified sequences stored in reference databases can lead to misidentification of taxa when DNA-based identification methods are used (Weigand et al. 2019). In addition, DNA-based identification to species is only reliable if a 'local barcoding gap' can be detected i.e. the intraspecific variation in the marker sequence is less than the interspecific divergence (Hebert et al. 2003a; Hebert et al. 2004). Originally, a standard threshold for the 'barcoding gap' was proposed (Hebert et al. 2004) but several studies have shown that the size or presence of a 'barcoding gap' varies among species and can be an artefact of insufficient sampling (Wiemers and Fiedler 2007; Robinson et al. 2009; Virgilio et al. 2010; Čandek and Kuntner 2015). What

is important for specimen identification is that the maximum variation within a species is lower than the minimum variation to its nearest neighbour (Figure 3.1) so that taxonomic assignment is not ambiguous (Meier et al. 2008; Collins and Cruickshank 2013). It is therefore essential that where local barcoding gaps do not exist among UK species this is identified to prevent acceptance of incorrect assignments of specimens. Increased intraspecific variation over large geographical scales has been suggested as a factor that might reduce the barcoding gap (Bergsten et al. 2012; Čandek and Kuntner 2015; Koroiva and Kvist 2018) and more accurate identification is possible when reference databases are focused on smaller geographical scales (Bergsten et al. 2012). It is therefore important that studies, such as this PhD project, that are targeted at DNA barcoding at a national level, consider the origin of sequenced specimens when assessing the quality of reference databases.

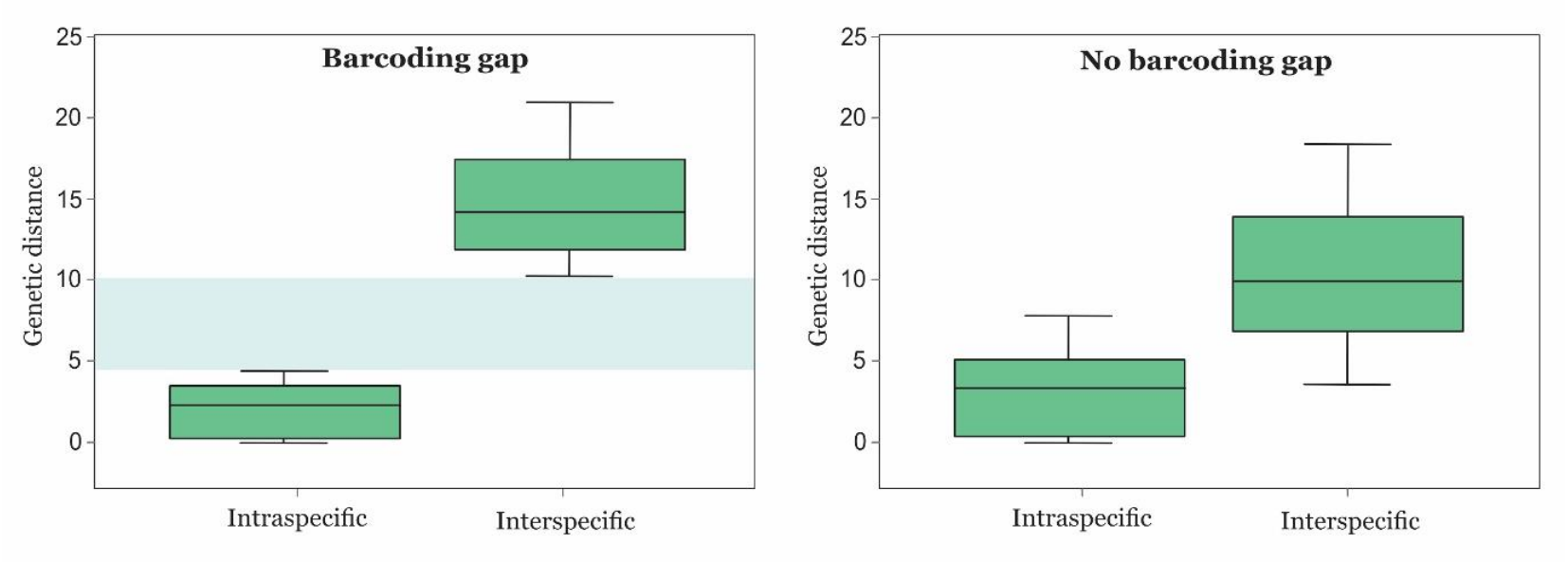

Figure 3.1 *Schematic diagram to provide an example of the presence (a) and absence (b) of a barcoding gap. The barcoding gap between the maximum intraspecific genetic distance and the minimum interspecific distance is shown by the light green band a. There is no barcoding gap in b due to the overlap of the intraspecific and interspecific genetic distances.*

3.2.5 Aims and hypotheses

Using freshwater arthropod sequences from the UK, this study addresses several potential biases in DNA-based species identification, and thus biodiversity and/or ecological quality assessment. The study tests the specific hypothesis that publicly available reference sequences for freshwater arthropods would not currently be comprehensive and accurate enough to provide accurate taxonomic identification for most taxa. Specifically, the following key questions are addressed:

- How complete is the current coverage of UK freshwater arthropod species in BOLD?
- How complete is the coverage of protected and non-native species?
- Do the sequences stored in BOLD show high intra- and/or low inter-specific variation that could cause difficulties in accurately identifying UK specimens?
- Could geographic variation in sequences be an important factor in accurate identification of UK specimens?

Recommendations are made for database assessment, curation and prioritisation of future UK species barcoding.

3.3 Methods

3.3.1 Coverage

3.3.1.1 *Coverage of UK freshwater arthropod species in BOLD*

To inform subsequent searches in BOLD, a single, comprehensive list of UK freshwater arthropods was created. Herein, freshwater arthropod species are taken to be those that, at any stage in their lifecycle, inhabit and/or feed in running or standing waters, and their associated riparian habitats. A list of freshwater arthropods was taken from the FreshBase (https://freshbase.myspecies.info/) freshwater invertebrate species list and was split into orders. The FreshBase order lists were then cross-checked against the UKCEH UK Checklist of Freshwater Species (Gunn et al. 2018), the NHM UK Species Inventory (https://www.nhm.ac.uk/our-science/data/uk-species.html) and the Global Biodiversity Information Facility (GBIF)

(https://www.gbif.org/) in order to gather the most commonly-used synonyms for each species name in the list. Taxonomy was matched to the NHM UK Species Inventory as disagreements in taxonomy were found among the lists. The NHM UK Species Inventory taxonomy was chosen so that this work was aligned with the taxonomy used by the UK Biological Records Centre. Species were listed as the name shown in the UK Species Inventory. Where the species names differed among the three lists, the names from the UK Checklist of Freshwater Species and the accepted name in GBIF were recorded as synonyms. Checklists for each order were created for use in reference sequence searches.

Searches in BOLD v4.0 (www.boldsystems.org) were conducted between the 1/9/20 and 3/9/20 using the BOLD Systems Workbench. Order checklists were uploaded to the BOLD Systems Workbench and a Record Search was completed for each order. Public COI reference sequences over 500 bp (the length accepted as a formal barcode standard in BOLD (Ratnasingham and Hebert 2007)) matching the checklist species names were downloaded. Records that BOLD had flagged as contaminated, sequences containing stop codons, and sequences that had been flagged for errors or misidentifications were removed from all analyses. The records were then summarised to obtain the total numbers of sequences for each species and the country that each specimen had been collected in. The BOLD public searches included all sequences recorded under each synonym as separate records. The total numbers of sequences listed under each synonym and accepted species name were combined to provide the final numbers.

Records in BOLD can be stored privately, and although the sequences and metadata cannot be accessed for private records, the total numbers of public and private sequences can be obtained through BOLD's checklist progress reports. Progress reports for each order level checklist were downloaded on 15/9/20. Progress reports did not include sequences recorded under the uploaded synonyms, so synonyms were searched for using separate checklists and results were then combined so that the total numbers included all records stored for each species.

In order to quantify the representation of UK specimens, the numbers of barcodes from UK specimens for each species were collated from the BOLD

public records. The proportions of species with sequences from UK specimens (thresholds of 'at least one sequence' and 'at least five sequences' (as in Weigand et al. 2019)) were calculated for each order.

Representation with more than one sequence for each species is important, both for verifying that individual sequences are stored under the correct taxonomic name, and for quantifying intraspecific variation in sequences. A threshold of three or five sequences has been used in studies as a suggested minimum number of sequences (Oliveira et al. 2016; Weigand et al. 2019; Fontes et al. 2021). In this study, a threshold of five public sequences was chosen to assess the number of species with a minimal level of within species representation in BOLD.

3.3.1.2 *Coverage of protected and non-native freshwater arthropod species*

A list of UK species with conservation designations was downloaded from JNCC (downloaded August 2020 https://hub.jncc.gov.uk/assets/478f7160-967b-4366-acdf-8941fd33850b) and cross-checked with the list of UK freshwater arthropods for this study. Both the recommended taxon name and the designated name from the JNCC list were cross-checked with all synonyms in the freshwater arthropod list used in this study. The list was filtered to remove the designations 'least concern', 'data deficient', 'not evaluated', 'insufficiently known' and 'indeterminate' prior to cross-checking. A list of non-native species was downloaded from the NBN Atlas (downloaded October 2020). This list was cross-checked with the freshwater arthropod list used in this study.

3.3.2 Sequence Variation

3.3.2.1 *Intra- and interspecific variation in stored sequences*

The maximum intraspecific genetic distance and the minimum distance to the nearest interspecific neighbour were calculated in BOLD Systems using the Barcode Gap Analysis tool (BOLD v4.0 (www.boldsystems.org)). Within these analyses, the BOLD aligner and the Kimura 2-Parameter (K2P) distance model (Kimura 1980) were used to analyse the sequences that were at least 500 bp in length stored for each species in each freshwater arthropod order. Sequences that were flagged in BOLD as contaminated, containing stop codons, misidentifications or errors were excluded from analyses. Gaps or ambiguous

bases in the sequences were handled with pairwise deletion. In addition, the BIN Discordance tool in BOLD was used to analyse which species in each order shared a BIN (barcode index number (Ratnasingham and Hebert 2013)).

The maximum intraspecific distance was plotted against the minimum distance to the nearest neighbour for each species within each order to show the presence or absence of a 'barcoding gap' (a gap between the intra- and interspecific distances (Robinson et al. 2009; Collins and Cruickshank 2013)). The proportion of species in each order that do not show a 'barcoding gap' was calculated. The proportion of species in each order that showed less than 2% divergence from their nearest neighbour was also calculated as this is a very low divergence for distinct species (Hebert et al. 2003a; Hebert et al. 2003b) and indicates a need for verification. If closely related species are unrepresented by a public barcode in BOLD, the barcoding gap might appear to be larger than it is, as the missing species might be the true nearest neighbour. It is therefore important to recognise that the barcoding gap can only be accurately assessed if all UK species within the family are represented by barcodes. In this study, barcoding gaps were still calculated where all species within the family were not represented by barcodes but these are highlighted (coloured orange) as they might be less accurate due to missing species. Amphipoda and Plecoptera were chosen as example orders in this study because they have relatively high proportions of species represented by barcodes, include species that lack a barcoding gap, and include species with barcodes from UK specimens.

Boxplots comparing intraspecific variation and interspecific distance to other species within the same family for both uncurated and curated data were plotted using R Statistical Software v4.0.2 (Core Team, 2020). Species from the order Plecoptera were chosen because most families within this order had full representation of all the UK species within the family (no families within Amphipoda had full representation). Three Plecoptera species that did not show a barcoding gap in the previous analysis were chosen (*Isoperla grammatica, Leuctra fusca* and *Amphinemura sulcicollis*). *Isoperla grammatica* and *Leuctra fusca* did not have a barcoding gap but did show over 2% divergence from their nearest neighbours. The families of both of these species have complete barcode representation in BOLD (Perlodidae and Leuctridae). Although

Nemouridae had one species that was not represented by a barcode in BOLD, *Amphinemura sulcicollis* was chosen in order to analyse a species which showed less than 2% divergence from its nearest neighbour. For each of the three chosen species, the pairwise genetic distances (K2P distances) between all of the publicly stored sequences in BOLD that were at least 500 bp in length were downloaded using the Distance Summary tool in BOLD Systems. The resulting data were curated and sequences that were likely to include errors or be misidentified were removed.

3.3.2.2 *Geographic intraspecific variation*

In order to investigate whether increased intraspecific variation over large geographical scales might reduce the barcoding gap and lead to ambiguous taxonomic assignment, data for the three species from the above analysis were processed in order to match the sequence process identification numbers to the specimens' country of origin. Sequences with >1% ambiguous bases were excluded from analyses. To visualise the similarity between sequences from different countries, principal coordinates analyses (PCoA) were performed in R (VEGAN package v2.5-7 Oksanen et al. 2020) on matrices of the K2P distances for each species.

3.4 Results

3.4.1 Coverage

3.4.1.1 *Private and public barcode coverage by class*

The checklists used in these searches gave a total of 3251 species of freshwater arthropod that are found in the UK. Seventy-three percent of these species are represented by at least one private or public barcode in BOLD. The percentage of UK freshwater arthropod species that are represented by public or private records in BOLD is very varied among classes (Figure 3.2), ranging from 46% of Arachnida species to 100% of Chilopoda species. However, as there is only one UK Chilopoda species associated with freshwaters this class is not comparable with the others. After Chilopoda, Malacostraca has the highest percentage (94%) of species represented by at least one privately or publicly stored barcode.

Only 60% of UK freshwater arthropod species are represented by publicly stored barcodes. While privately stored barcodes are used in BOLD when assigning an identity to a new sequence, the private sequence and metadata are not available for further analyses and verification. While, for some UK freshwater arthropod classes, all the barcodes stored in BOLD are publicly available (i.e. Chilopoda and Branchiopoda), other classes have large percentages of species that are currently only represented by barcodes that are not publicly available. The percentage of species in each class represented only by privately stored barcodes in BOLD ranges from 0% to 33%, with particularly high percentages (18-33%) in the classes Maxillopoda, Ostracoda and Arachnida.

3.4.1.2 *Public barcode coverage by class*

Three classes show high coverage with public barcodes in BOLD, with over 75% of species in these classes represented by publicly available sequences and metadata (Chilopoda, Branchipoda and Malacostraca). In addition, although the percentage of freshwater Insecta species represented by at least one publicly stored barcode is lower (67%), the total number of species in this class (2526) is much larger than any other class so very high numbers of insect species are represented. Representation of UK freshwater species in Maxillopoda, Ostracoda and Arachnida with public barcodes is currently very low (21-33%) in comparison with the other classes.

3.4.1.3 *Private and public barcode coverage by order*

The barcode coverage of UK freshwater arthropod species in BOLD shows high variation among orders, ranging from 18-100% of species represented by at least one private or public barcode (Figure 3.2). In eight of the thirty-two orders, every species is represented by at least one privately or publicly stored barcode. However, in some orders, only 18-50% of species are represented (i.e. Hymenoptera, Bathynellacea, Arguloida, Harpacticoida, Poecilostomatoida, Siphonostomatoida, Acarina, Oribatida and Prostigmata). In twenty-four of the orders, over three-quarters of the stored sequences are stored publicly. However, between 30 and 100% of the stored sequences for Hymenoptera, Calanoida, Harpacticoida, Siphonostomatoida, Podocopida, Acarina, Oribatida and Prostigmata are not publicly available.

3.4.1.4 *Public barcode coverage by order*

Six orders have at least one publicly stored sequence for every species (Megaloptera, Geophilomorpha, Anostraca, Notostraca, Mysida and Araneae). These orders all have very low numbers of UK freshwater species (ranging from one to five species). There are, however, other orders with similarly low numbers of species that do not have every species represented by at least one barcode (Neuroptera, Bathynellacea, Arguloida and Oribatida). Several orders with much higher numbers of UK freshwater species also have a very high percentage of species represented by at least one publicly stored barcode. At least three-quarters of UK freshwater species in the following orders have a publicly stored barcode in BOLD: Coleoptera, Ephemeroptera, Hemiptera, Odonata, Plecoptera, Trichoptera, Amphipoda, Decapoda and Isopoda.

In comparison, eleven UK freshwater arthropod orders have much lower percentages (50% or lower) of species represented by a publicly stored barcode in BOLD (Hymenoptera, Collembola, Bathynellacea, Arguloida, Calanoida, Cyclopoida, Harpacticoida, Poecilostomatoida, Siphonostomatoida, Podocopida, Prostigmata). Two Arachnid orders (Acarina and Oribatida) have no species represented by publicly stored sequences.

Although some orders have comparatively lower percentages of species with barcodes, there are large differences in the number of species in each order so these proportions represent very large numbers of species in some cases. Only just over half of Diptera species are represented by publicly stored barcodes, but this represents a total of 928 species sequenced. Similarly, the 80% of Coleoptera species that are represented by at least one public barcode equates to 347 species. Despite Insecta having a much larger number of species than any other class, insect orders (with the exception of Hymenoptera) have very high proportions (mean = 84% (SD = 14%)) and/or very high numbers of species represented by barcodes.

3.4.1.5 *Representation from UK specimens*

In total, only 5% of UK freshwater arthropod species have public sequences stored in BOLD that are from specimens collected in the UK. The proportion of species represented by sequences from UK specimens is very low in most classes

and very variable between orders, ranging from 0 to 100% of species (Figure 3.2). Only Bathynellacea (two species), Notostraca (one species) and Lepidoptera (nine species) have at least half the species represented by UK sequences. Half of the orders have no species represented by publicly stored UK sequences.

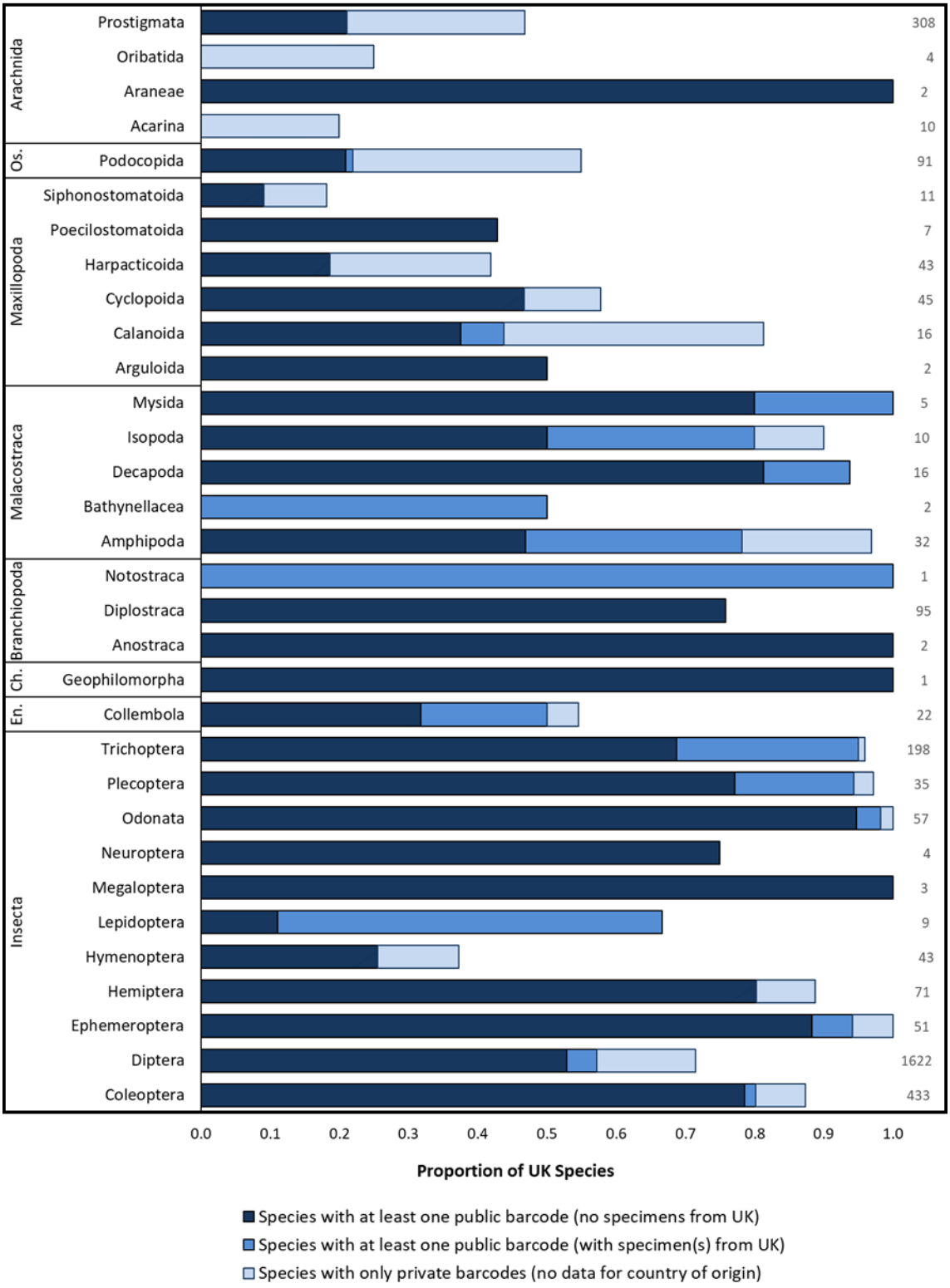

Figure 3.2 *Proportion of UK freshwater arthropod species in each order represented by COI barcodes stored in BOLD (searches conducted between the 1/9/20 and 3/9/20). Total number of species in each order shown to the right of each bar. (Abbreviated class names: Entognatha (En.); Chilopoda (Ch); Ostracoda (Os.)).*

3.4.1.6 *Coverage of protected and non-native UK freshwater arthropod species in BOLD*

A total of 660 UK freshwater arthropod species in the checklists used in this study are designated as protected species in the UK. Eighty percent of these species have at least one private/public barcode stored in BOLD. Only fifteen of the thirty-two orders include protected species and the barcode coverage of protected species in BOLD varies among those orders, ranging from 50-100% (Figure 3.3). Eight of the orders only include one or two protected species and all of these orders show 100% coverage except for Neuroptera (which has the lowest percentage coverage of all the orders with 50% (one of the two species)). In the remaining seven orders that include higher numbers of protected species (all of which are in the class Insecta), coverage varies from 67-100%. Most of the protected species have publicly available barcodes (68%). However, Diptera, Ephemeroptera and Plecoptera have at least 10% of the barcodes for protected species currently stored privately. Representation of protected freshwater arthropod species with barcodes from UK specimens is very low (2% of protected species).

Thirty-seven non-native species, from ten orders, are included in the checklists used in this study. Seventy-three percent of these species are represented by at least one private or public barcode in BOLD (Figure 3.3). Seven of the orders have 100% of the non-native species represented by private or public barcodes. There are no barcodes stored for the three non-native species in the order Siphonostomatoida. The remaining two orders have 20-25% of non-native species represented by barcodes in BOLD. In most orders (eight of the nine orders that have barcodes stored for non-native species) the barcodes stored are all available publicly. Thirty-three percent of the non-native species in Amphipoda are currently only stored privately. Representation of non-native freshwater arthropod species with barcodes from UK specimens is very low (8% of non-native species).

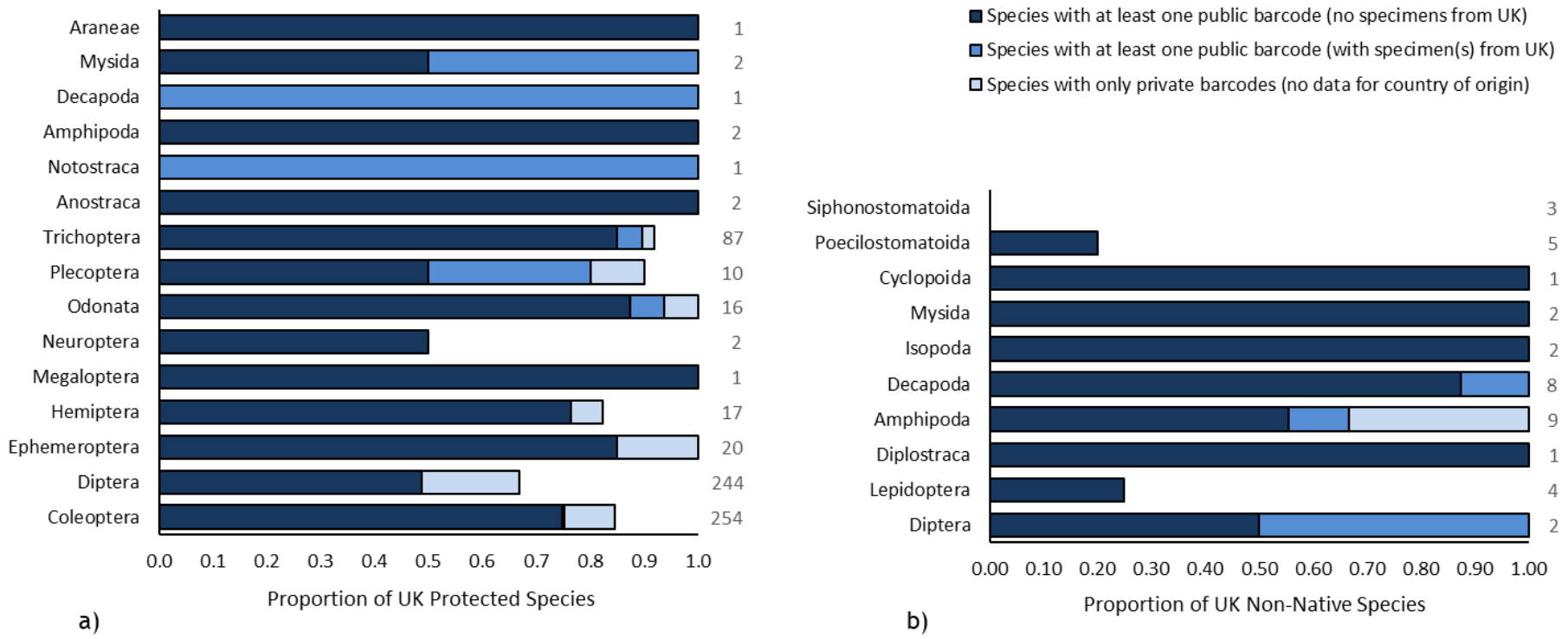

Figure 3.3 *Proportion of UK freshwater arthropod species in each order represented by COI barcodes stored in BOLD for protected species (a), and non-native species (b) (searches conducted between the 1/9/20 and 3/9/20). Total number of species in each order shown to the right of each bar*

3.4.1.7 *Coverage of UK freshwater arthropod species with multiple sequences*

Only 37% of UK freshwater arthropod species are represented in BOLD by at least five public barcodes (Figure 3.4). Most orders show much lower percentages of species represented by at least five public barcodes. However, seven orders (Lepidoptera, Anostraca, Notostraca, Decapoda, Mysida, Siphonostomatoida and Araneae) have at least five public barcodes for all the species that are represented by public barcodes. In four orders (Geophilomorpha, Bathynellacea, Arguloida and Poecilostomatoida), none of species that are represented by public barcodes have at least five sequences stored.

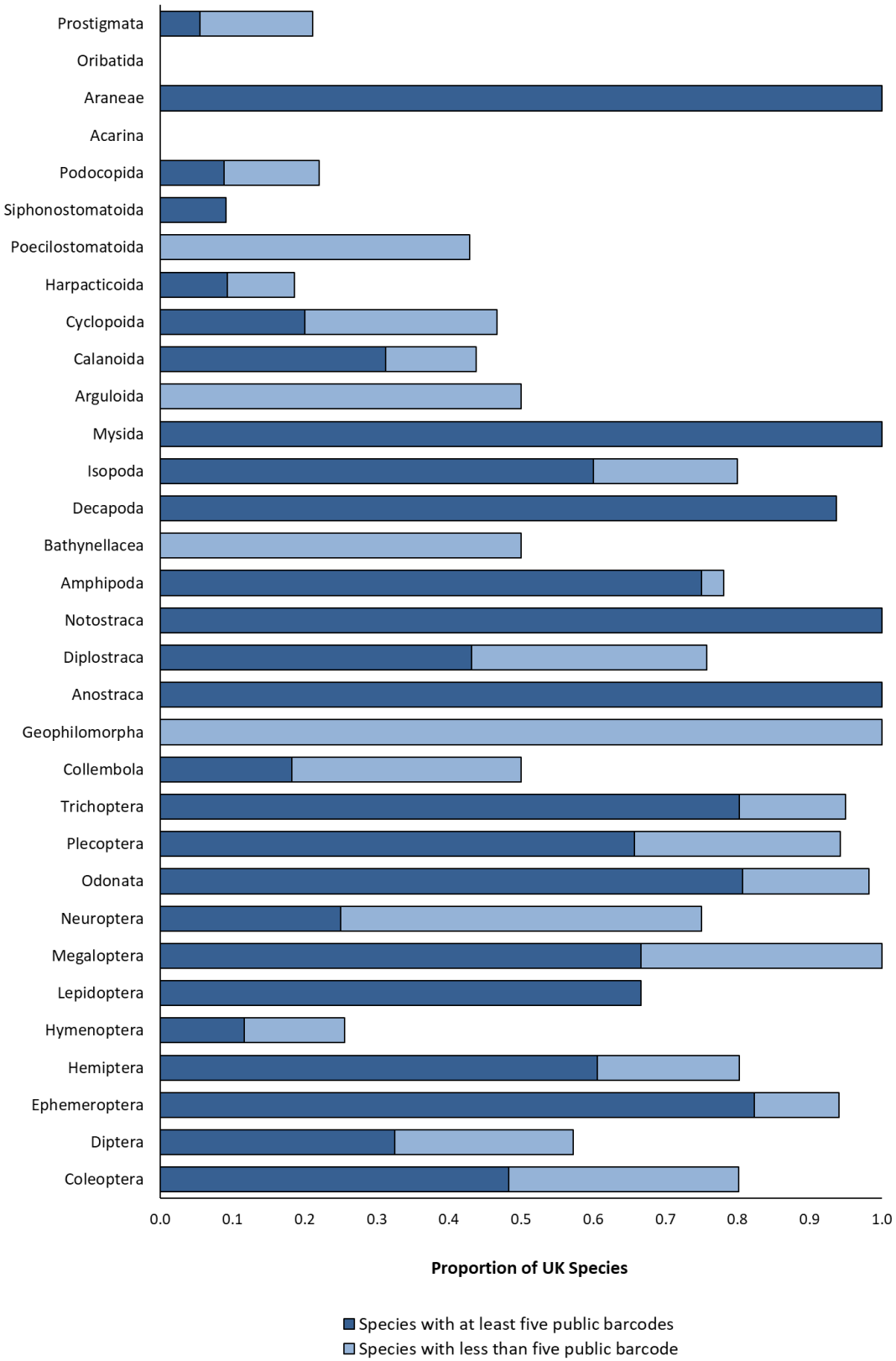

Figure 3.4 *Proportion of UK freshwater arthropod species in each order represented by at least five public COI barcodes stored in BOLD in comparison with the total proportion of species represented by at least one public barcode (searches conducted between the 1/9/20 and 3/9/20).*

3.4.2 Sequence Variation

3.4.2.1 *Intra- and interspecific variation in stored sequences*

The intra- and interspecific genetic distances of all species in each of the thirty-two orders were plotted in order to show whether each species has a clear barcoding gap from other closely related UK species (based on the sequences publicly available in BOLD). Two orders (Amphipoda and Plecoptera) were chosen to use as examples in this study because they have relatively high proportions of species represented by barcodes, include species that lack a barcoding gap, and include species with barcodes from UK specimens. (Figure 3.5). Eight UK Plecoptera (23%) and six UK Amphipoda (19%) species lack a barcoding gap (Figure 3.5). In addition, six Plecoptera species have less than two percent divergence from their nearest neighbour. Interspecific distances might be overestimated where all confamilial species do not have publicly stored barcodes present in BOLD due to the possibility that a missing species is the most closely-related species. Species from families that are fully represented by publicly stored barcodes in BOLD are shown as blue points in Figure 3.5, whereas species from families with unrepresented species are shown as orange points. No families in the order Amphipoda have representation in BOLD for all species. In the order Plecoptera, all families except one (Nemouridae) have representation for all species within the family.

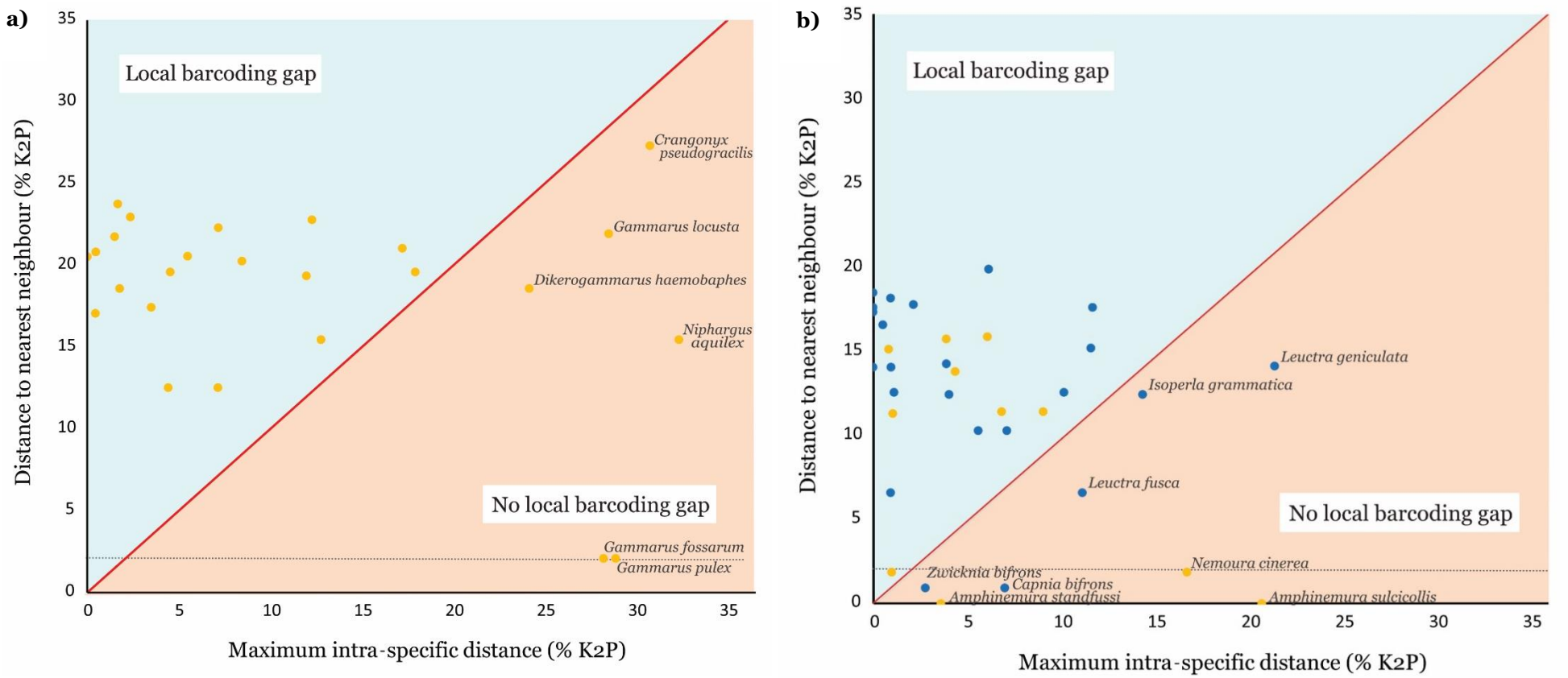

Figure 3.5 *Barcoding gap plots comparing the maximum intra-specific distance of a species with the minimum distance to the nearest neighbour (Kimura 2-Parameter) for the orders: Amphipoda (a) and Plecoptera (b). Species above the line show a local barcoding gap and those below the line lack a local barcoding gap (based on the publicly available sequences in BOLD (1/9/20 and 3/9/20)). Species without a barcoding gap are labelled. Two percent divergence from the nearest neighbour is marked with a dotted line. Blue points show species where all confamilial UK species have stored public barcodes. Orange points show species from families that do not have complete coverage of UK species within the family.*

The high intraspecific and low interspecific variation present in the publicly stored sequences indicates potential problems for accurate species identification using DNA-based methods. Analysis of all 32 orders showed that 19 orders include species that do not have a barcoding gap based on current publicly stored sequences and 13 orders include species that have two percent or less divergence from their nearest neighbour species. Of the total 1949 UK freshwater arthropod species that have publicly stored barcodes, 29% have no local barcoding gap and/or less than two percent divergence from their nearest neighbour (17% of all UK freshwater arthropod species (Figure 3.6)).

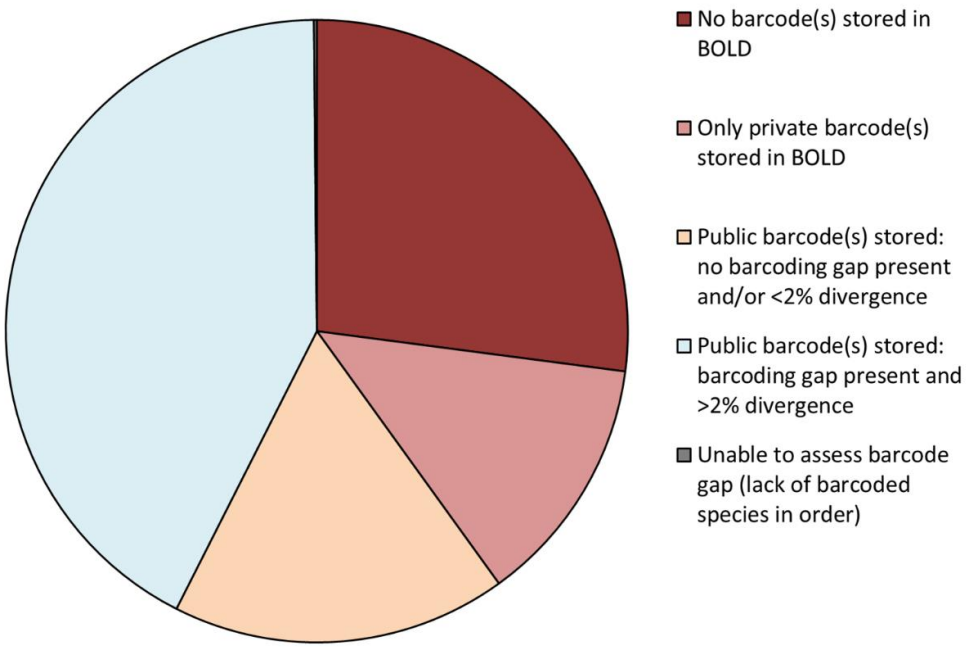

Figure 3.6 *Proportions of UK freshwater arthropod species with: no barcodes stored in BOLD; only private barcodes stored in BOLD; publicly stored barcodes in BOLD but without a barcoding gap and/or <2% divergence (potentially problematic for accurate species-level identification); publicly stored barcodes in BOLD with a local barcoding gap present and >2% divergence from their nearest neighbour (therefore should be able to be accurately identified to species-level); too few barcoded species to assess the barcode gap. Genetic distances calculated using the Kimura 2-Parameter based on the COI barcodes stored in BOLD (searches conducted between the 1/9/20 and 3/9/20).*

3.4.2.2 *Intraspecific Variation*

Out of the two orders chosen as examples in the intra- and interspecific variation section above (Amphipoda and Plecoptera), only Plecoptera had families with complete species barcode representation. Plecoptera species were therefore chosen to further explore intraspecific variation in barcodes. Although in the above analysis (Figure 3.5b) there were two species (*Zwicknia bifrons* and *Capnia bifrons*) with less than 2% divergence from their nearest neighbours that also had complete species barcode representation within their families, *Capnia bifrons* is a synonym of the accepted name *Zwicknia bifrons*. There are sequence records for both names in BOLD and records were combined in these analyses, but the BOLD tools treat them as separate species still so they appear as two species with less than 2% divergence from each other.

Intraspecific variation analyses for individual species included lower numbers of sequences than the previous analyses. This made it feasible to download and process the data to remove sequences with >1% ambiguous bases, enabling more accurate measures of maximum intraspecific distances to be calculated for these species. *Isoperla grammatica* showed a maximum intraspecific distance of 13.98% using the BOLD Barcode Gap Analysis tool and a minimum distance to its nearest neighbour of 12.41% so did not have a barcoding gap. A box plot of the pairwise intraspecific and interspecific distances (from *I. grammatica* to other species within the same family) (Figure 3.7a) showed that most intraspecific pairwise comparisons were much more similar and the high maximum distance was caused by a small number of outliers. The removal of two sequences that included >1% ambiguous bases reduced the maximum intraspecific distance to 7.21% which then shows a clear barcoding gap between *I. grammatica* and the other UK species within the same family (Figure 3.7b).

Leuctra fusca showed a maximum intraspecific distance of 10.85% using the BOLD Barcode Gap Analysis tool and a minimum distance to its nearest neighbour of 6.58%, and so did not have a barcoding gap. A box plot of the pairwise intraspecific and interspecific distances (Figure 3.7c) showed a group of outliers that had very high pairwise distances to the majority of sequences for the species. One sequence had >1% ambiguous bases but was not the cause of the outlier points and so removal of this sequence did not produce a

barcoding gap. The outlier points were all caused by one sequence of 1261 bp in length. Removal of this sequence would reduce the maximum intraspecific distance to 4.29% which then shows a barcoding gap between *L. fusca* and other UK species within the same family (Figure 3.7d). This flags this sequence as a potential misidentification or error. Removal of this long sequence would also reduce the number of outliers on the interspecific distance box plot but would not remove all of them. Identification of the sequences causing the interspecific outliers determined that this group of more similar sequences were all from one species *Leuctra moselyi*, which is the nearest neighbour to *L. fusca*.

Amphinemura sulcicollis showed a maximum intraspecific distance of 20.15% using the BOLD Barcode Gap Analysis tool and a minimum distance to its nearest neighbour of 0% so did not have a barcoding gap and was chosen because it showed <2% divergence from its nearest neighbour. A box plot of the pairwise intraspecific and interspecific distances (Figure 3.7e) showed outliers that were as distant from most of the *A. sulcicollis* sequences as the sequences for other species within the family. In addition, outliers for the interspecific box showed that it includes sequences that were as similar to *A. sulcicollis* sequences as those within the species. This suggested some sequences stored in BOLD as *A. sulcicollis* were potentially misidentified/mislabelled and belonged to a different species within the family. The BOLD Taxon ID Tree tool was used to create a nearest-neighbour joining tree for the genus which confirmed that four *A. sulcicollis* sequences group together with *Amphinemura standfussi* (the nearest UK neighbour) rather than *A. sulcicollis*. These species are morphologically similar so are likely to be misidentified. Removal of these four sequences reduced the maximum intraspecific distance to 5.41% and increased the minimum interspecific distance to 17.06% which then shows a clear barcoding gap between *A. sulcicollis* and the other species within the family (Figure 3.7f).

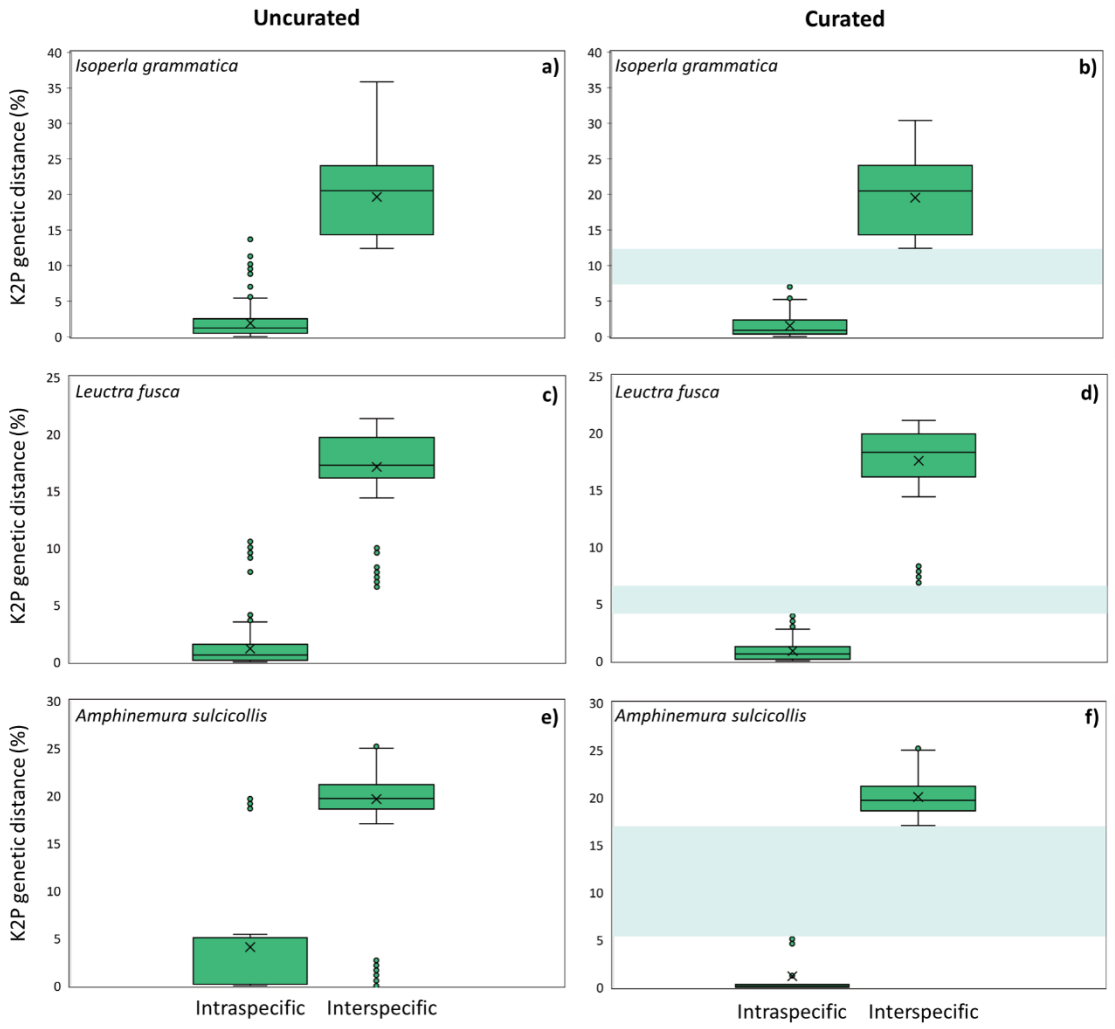

Figure 3.7 *Boxplots for three Plecoptera species comparing intraspecific variation and interspecific distance to other species within the same family: Isoperla grammatica (a and b), Leuctra fusca (c and d), and Amphinemura sulcicollis (e and f) (K2P genetic distance). Uncurated sequences include errors/misidentifications that can inflate the intraspecific distance and reduce the barcoding gap (a, c, and e). Curation of sequences that cause outlier points and removal of sequences that are likely to be errors/misidentifications can improve assessment of the barcoding gap (b, d, and f).*

3.4.2.3 *Geographic Variation*

The lack of barcoding gaps for *I. grammatica* and *A. sulcicollis* are clearly caused by the inclusion of sequences with errors and/or misidentifications in the reference database. Although the sequence causing the lack of a barcoding gap for *L. fusca* is not as clear an error or misidentification, it is a very large distance from the other *L. fusca* sequences and the fact that it is the only sequence to be so dissimilar provides low confidence in the accuracy of the sequence.

The recalculated maximum intraspecific distance for *I. grammatica* in particular is still relatively high for within species variation. This species has a wide distribution and a high number of previous subspecies and synonyms listed in GBIF (https://www.gbif.org/species/2004039) so the high variation could be due to cryptic species. In this case, the high intraspecific distance is not a major concern for identification of specimens using DNA-based methods due to the presence of the barcoding gap. However, in some species, high intraspecific variation that cannot be attributed to errors/misidentifications could affect the barcoding gap and limit accurate and confident identification of specimens to species level. In addition, it is important that the sequences used to design specific or group primers to detect specific species are as accurate as possible and relevant to the target geographic location. Where intraspecific variation is naturally high and/or potential errors cannot be easily identified, further analysis can elucidate the sequences that are most likely to be accurate and those that group closely with sequences from the target location.

The recalculated maximum intraspecific distance in the three species analysed above ranges from 4.29% to 7.21% (K2P). Two of the species chosen for intraspecific variation analysis are represented by barcodes from specimens from the UK enabling comparison of the similarity of UK sequences to those from other countries. Analyses of pairwise distances of the barcodes stored in BOLD for each species show very different patterns of similarity within and among the countries of specimen origin (Figure 3.8). While the majority of sequences for a species often cluster together in similarity, there are often many sequences that are much less similar to the main cluster which sometimes form separate clusters of similar sequences.

Isoperla grammatica (Figure 3.8a) shows one large cluster of more similar sequences from specimens from the UK, Norway, France, and Switzerland but also a smaller cluster of similar sequences from specimens from Portugal. A sequence from Germany and a sequence from Austria form a third cluster and one individual sequence from the UK is not similar to any of the other sequences currently stored. *Leuctra fusca* (Figure 3.8b) shows two main clusters and two single sequences that are less similar. The least similar sequence here (from a specimen from Italy) is the single long sequence that caused the outlier points in the intraspecific analysis but has been retained here as it was not a clear error or misidentification. The largest cluster consists of sequences from Germany, Switzerland, and Austria. The smaller cluster consists of sequences from France and the remaining single sequence is from a specimen from China. *Amphinemura sulcicollis* (Figure 3.8c) shows two clusters: one consisting of sequences from the UK, Norway, and Germany and the other cluster from Switzerland, France, Portugal, and Germany. For this species, sequences from specimens in Germany are located in both clusters.

The similarity of sequences from UK specimens to sequences from specimens from other countries is varied. In the two species with UK representation, UK sequences are clustered with sequences from other countries but with distinct dissimilarity to the sequences from some countries. The minimum distance between clusters of UK sequences and those from countries that cluster separately ranges from 1.98 to 3.31% for *I. grammatica,* and from 4.61 to 4.93% for *A. sulcicollis*. Where there are no sequences from UK specimens, like for *L. fusca*, it is unknown whether UK sequences would cluster together with the other countries, and what the distance would be to other sequences currently stored for the species.

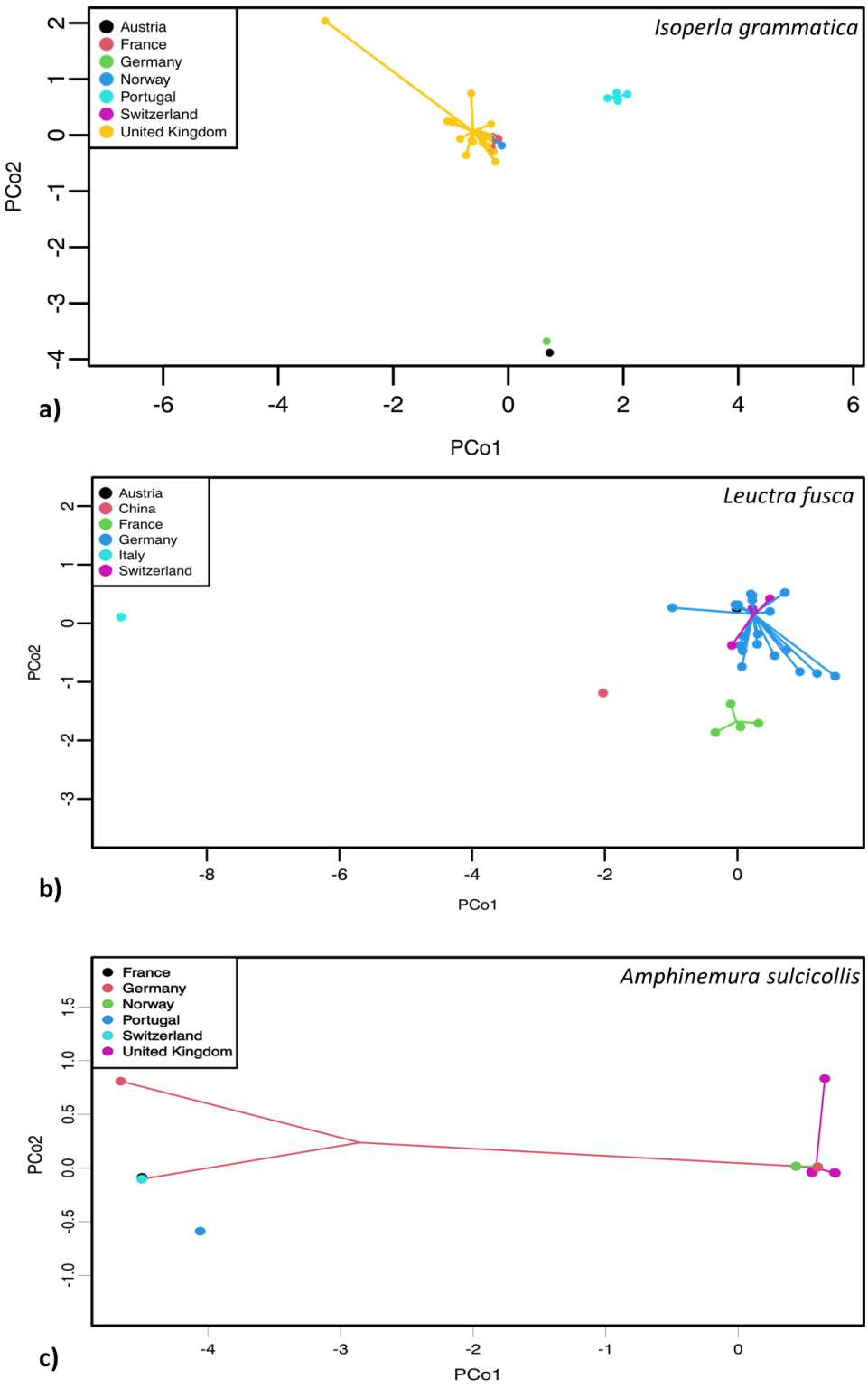

Figure 3.8 *Principal Coordinate Analyses (PCoA) of the similarity of COI barcodes (Kimura 2-Parameter) stored publicly in BOLD (searches conducted between the 1/9/20 and 3/9/20) for three Plecoptera species: Isoperla grammatica, Leuctra fusca, and Amphinemura sulcicollis. Colours show the countries where the sequenced specimens originated. Maximum intraspecific distance (sequences with >1% ambiguous bases removed) are Isoperla grammatica: 7.21%; Leuctra fusca: 4.29% (not including the single sequence from Italy identified earlier as an outlier (10.85% including the sequence from Italy); Amphinemura sulcicollis: 5.41%).*

3.5 Discussion

3.5.1 Biodiversity assessment for UK freshwater arthropods

The priority in being able to identify UK freshwater arthropod species using DNA-based methods is to have barcodes from expertly identified voucher specimens of every species stored in publicly accessible reference databases. Biodiversity assessments are usually done using metabarcoding where the DNA of multiple species is amplified and then matched to barcodes in reference databases to identify which species are present. Overall coverage of UK freshwater arthropod species in BOLD is good at 73% represented altogether, but only 60% of species are represented by publicly stored barcodes. Privately stored barcodes are used by BOLD in identification of species, but the actual sequences and metadata cannot be accessed for quality control. The release of these privately stored barcodes to the public will provide sequences and metadata for another 422 UK freshwater arthropod species, and will make a particularly large difference to the classes Maxillopoda, Ostracoda and Arachnida but will also provide data for many more species in Insecta and Malacostraca.

Representation of UK freshwater arthropod species with publicly available barcodes in BOLD is currently biased, with three classes much less well-represented than the others (Maxillapoda, Ostracoda, and Arachnida). The bias towards some classes might be partly explained by the proportion of protected or non-native species a class contains. The three classes with the highest proportions of species represented by barcodes (not including Chilopoda which is just one species) are the only classes to contain protected species. Only small proportions of Branchipoda and Malacostraca are protected (3% and 8% respectively) but 26% of species in Insecta are protected. Most classes do not contain any non-native species but 32% of species in Malacostraca are non-native and low proportions of Branchipoda and Maxillopoda are non-native.

Orders with very high species representation are often small in terms of species number i.e. Megaloptera, Geophilomorpha, Anostraca, Notostraca, Mysida, and Araneae all have 90-100% coverage but only 1-5 species in the order. The remaining orders that have 90-100% coverage but higher numbers of species (ranging from 35-198) are all in the class Insecta (Ephemeroptera, Odonata,

Plecoptera, and Trichoptera). There is a long history of ecological quality monitoring based on the orders Ephemeroptera, Plecoptera, and Trichoptera and they remain important orders for monitoring Ecological Quality Status of aquatic ecosystems under the Water Framework Directive (WFD, Directive 2000/60/EC) (European Commission, 2000). Therefore, species-level barcodes are needed for these orders especially (Weigand et al. 2019). The proportion of protected species appears to have a strong influence on which orders have high coverage, with all orders that contain protected species having at least 70% coverage. It is possible that having protected species in an order drives not just the barcoding of those species, but also of other related species too, due to the need to test specific assays (see Monitoring protected and non-native species section below).

Interest in having barcodes for protected and non-native species, and for species that are important for monitoring of water quality might have driven the biases seen in the coverage of UK freshwater arthropod species in BOLD. These biases are undesirable in the long term as they cause severe limitation in the taxonomic groups that can be identified using DNA-based methods. One of the benefits of these methods is the potential to detect and monitor species that are not already monitored using morphological methods due to their small body size and/or challenging identification. A focus on barcoding taxa that are already monitored, may provide faster and more comprehensive assessment of these groups than was possible before but it limits the potential application of DNA-based methods to other questions. However, the high coverage of some groups provides a very good foundation on which to build. Assessment of which groups have high coverage and where the gaps are now will now enable a more focused approach to barcoding species that are not already represented. This will greatly increase the proportion of UK freshwater arthropods that can be identified and monitored using DNA-based methods in the future.

3.5.2 Monitoring protected and non-native UK freshwater arthropod species

Monitoring of target species, both protected and non-native, is a particularly important goal for the use of DNA-based methods. Biodiversity assessments, aiming to detect and identify multiple species in a community using

metabarcoding, can detect the DNA of target species present in the community and can provide an effective initial screening in cases where there are multiple target taxa. However, screening methods that target single species or a group of closely-related species are usually more sensitive than metabarcoding and so are often preferable when monitoring species that are threatened or potentially invasive as false positives or negatives could have serious implications for management. Specific assays need thorough development and validation before they can be used to screen samples for target species.

In order to design specific assays for screening for target species, it is obviously essential that the barcodes of the target species are available in reference databases so that assays can be designed that target the DNA sequence for that species. However, it is also essential that the barcodes of closely-related species are also available in reference databases so that the assays can be designed to amplify the target species but not any other species found in the habitat. Which other species are important depends on which related species are potentially present in the target habitat so although many protected/non-native species are represented in reference databases, specific assays for them cannot be designed without barcodes for other potentially relevant species. This assessment provides an overview of the protected and non-native species that are represented in BOLD as a first step in determining whether protected and non-native species have the necessary representation in reference databases to enable them to be monitored using DNA-based methods. Our ability to monitor them in practice depends on related species found in the habitat also having barcode representation, and assays being designed, thoroughly tested, and validated to ensure specificity.

The coverage of both protected species and non-native species is higher than the overall coverage of species (68% and 65% respectively compared to 60% overall). There is very high coverage (90-100%) of protected species in nine orders (total of 15). However, most of these orders (seven) have very low numbers of protected species (i.e. one or two protected species). The orders with higher numbers of protected species are all in the class Insecta in which coverage ranges from 67-100%. Non-native species have very high coverage (90-100%) in six orders (total of ten). All orders have low numbers of non-native species (maximum of nine). The higher coverage of protected and non-native

species shown in this study and the tendency for orders containing protected/non-native species to have higher coverage overall suggests that the desire to monitor target species has so far driven a lot of the barcoding effort. Filling the gaps in barcode reference databases will enable the development of specific assays to target current species of interest and also enable the rapid development of these assays as needed as species distributions shift under climate change.

3.5.3 Database curation

Having a barcode stored publicly for each species is the first step in enabling DNA-based identification of UK freshwater arthropods but is not sufficient to ensure that identification is accurate. Assessment of the barcodes stored in BOLD shows that it is essential that species are represented by multiple barcodes and that barcodes are curated to remove any records that contain errors or that are misidentified. Errors and misidentifications in reference databases can lead to misidentifications of sequences generated via metabarcoding, causing false positives and negatives in results. They can also lead to specific assays failing to detect target species and/or detecting non-target species in error. These consequences could be extremely serious and undermine the accuracy and reliability of DNA-based identification methods.

Having multiple sequences stored per species is important for two reasons. Firstly, such replication provides confidence that sequences are stored under the correct taxonomic name. The accuracy with which a sequence can be matched with the correct taxonomic name depends upon taxonomic expertise, and minimisation of potential errors in the laboratory and during data management. If there is only one sequence stored for a species, a misidentification or error might remain as the sequence for that species. Multiple morphologically identified sequences per species (especially when the specimens and sequences originate from different groups/locations) enables individual sequences that contain errors or that have been misidentified to stand out as different from the rest of the sequences and so be checked and, if necessary, removed from the database. Many sequences currently stored in public reference databases have been identified using reverse taxonomy (where a new sequence is matched to stored reference sequences and assigned their

taxonomic name rather than specimens being expertly identified prior to sequencing) which causes any initial errors/misidentifications to perpetuate through databases and makes curation of sequences more challenging. Although the method of identification is a database field for submitting records to BOLD, it is often left blank so it is difficult to curate sequences based on the identification method.

Secondly, multiple sequences per species enables intraspecific genetic distances to be calculated which is important in determining whether species can be identified to species level from the genetic marker in use. The maximum intraspecific variation is needed to calculate whether the species has a local 'barcoding gap' which is necessary for accurate identification to species. Ideally, to gain a good measure of intraspecific distance, the specimens need to originate from different geographic locations, in this case, from the different countries and regions of the UK. If very few specimens of the species have been sequenced or if all the specimens originate from the same area, intraspecific variation for the species is likely to be underestimated which could lead to the assumption of a local barcoding gap that does not necessarily exist in reality.

This study found that only 37% of UK freshwater arthropods are represented by at least five barcodes in BOLD. This means that the majority of UK freshwater arthropod species do not have sufficient sequences to measure intraspecific variation and so it is not known if these species have a local barcoding gap and therefore if they can be accurately identified using the COI marker. In addition, the geographic representation of the sequences is often limited so the intraspecific variation is likely to be underestimated in these cases even if multiple sequences for the species are stored. The number of barcodes stored per species is highly variable ranging from one to 9737. Most species have relatively low numbers of barcodes stored but some species have hundreds or even thousands of sequences stored publicly. Only Lepidoptera, Anostraca, Notostraca, Decapoda, Mysida, Siphonostomatoida and Araneae have at least five sequences for all the species that have publicly stored barcodes.

Although the low numbers of sequences stored for the majority of UK freshwater arthropod species would suggest intraspecific variation is likely to be underestimated, 29% of the species that have public barcodes stored in BOLD

currently lack a barcoding gap and/or have less than two percent divergence from their nearest neighbour. This is likely to result in inaccurate or uncertain taxonomic identification for these species. Where species have very low numbers of sequences stored (or sequences only from one area), the addition of new sequences might cause an increase in intraspecific variation which would reduce the current barcoding gap. In addition, when species are added to BOLD for closely-related species that currently do not have sequences stored, the interspecific distance to the nearest-neighbour might decrease for some species, again, potentially reducing the barcoding gap. The barcoding gap assessments in this study are based on the sequences available at the time of the searches. Where all species in the family are not represented by publicly available barcodes in BOLD, the interspecific distance might be overestimated and so should be reassessed once barcodes for all confamilial species are available.

A lack of barcoding gap can be caused by natural variation in sequences. Rapid or recent radiation events and/or the presence of cryptic species can cause the sequences of different closely-related species to be very similar (Hajibabaei et al. 2006; Ward et al. 2009). In these cases, the species cannot be discriminated with the marker and sequences generated from metabarcoding will only be identifiable to a higher taxonomic level. Knowledge of whether or not a barcoding gap exists for species of interest, prior to using metabarcoding, can help to inform whether the marker can provide the desired level of data for particular biodiversity assessments or whether a different marker would be more suitable.

However, a lack of barcoding gap can also be caused by errors and misidentifications in the reference databases. Where sequences with errors are included in databases, the intraspecific variation is likely to be artificially high due to the greater differences between the accurate sequences and the ones containing errors. The inclusion of misidentified sequences can cause a greatly inflated intraspecific distance that is as large as or larger than the distance to the nearest neighbour because the sequence that has been misidentified is likely to belong to a related species (especially if the misidentification was at the stage of morphological identification of the specimen rather than lab or data management errors). Misidentifications can therefore cause a lack of

barcoding gap and a less than two percent distance to the nearest neighbour. It is essential to know if a lack of barcoding gap is caused by natural variation or by errors or misidentifications as curation of databases can remove problematic sequences and enable accurate identification of the species.

The case study of Plecoptera herein shows that many of the species which lack a barcoding gap and/or have less than two percent divergence from the nearest neighbour are caused by errors and misidentifications in the reference sequences rather than natural variation. More detailed analysis of the sequences revealed that the inclusion of sequences that have over one percent ambiguous bases in the sequence tools available in BOLD, can cause intraspecific distances to be exaggerated and removal of these sequences can reduce the intraspecific variation to the point that a barcoding gap exists (e.g. *I. grammatica*).

The genetic distances reported using the Barcode Gap Analysis tool in BOLD might represent an overestimate of intraspecific distances in some cases, due to the inclusion of sequences that have high percentages of ambiguous bases. BOLD systems analyses provide options to exclude sequences that are below a minimum length, recorded as contaminated, contain stop codons, or were flagged as misidentifications or errors, which was done here, but sequences with high percentages of ambiguous bases cannot be removed prior to the use of the BOLD tools. Where sequences with high percentages of ambiguous bases are included in the Barcode Gap Analyses, intraspecific distances are likely to be higher.

Sequences with ambiguous bases are not always the cause though. Analysis of sequences stored for *L. fusca* showed that single sequences that are more distant from the rest of the sequences for the species might cause a lack of barcoding gap and might not be flagged as having any errors. Where it is a single sequence causing the lack of barcoding gap and the distance to the other sequences for the species is high, it would seem likely that the sequence is not accurate and removal of the sequence would enable more accurate identification of sequences. Analysis of sequences for *A. sulcicollis* (a species with a lack of barcoding gap and less than two percent divergence from the nearest neighbour) showed that species that include misidentifications can be

easily identified, and removal of the sequences can show the presence of a barcoding gap and enable accurate identification of both species.

The findings of this study are similar to those focused on DNA-based identification of Lepidoptera in Finland and Austria. This study found that nearly 20% of the species required further investigation of the sequences stored in BOLD to resolve issues relating to intraspecific distances that could be caused by errors and misidentifications. This is despite the fact that barcode effort and analysis for Lepidoptera has been more intensive than for any other insect order (Huemer et al. 2014). The level of curation necessary in such a well-studied group alongside the findings in this study suggest that errors and misidentifications in reference databases is a widespread issue and curation of reference sequences is critical for reliable DNA-based identification.

3.5.4 Geographic representation

The overall coverage of UK freshwater arthropod species from UK specimens is very low (5%) and even lower for species that are represented by at least five sequences from UK specimens (2%). Whether lack of sequences from UK specimens is an issue for DNA-based identification of freshwater arthropod species in the UK depends on the intraspecific variation within the species. Increased intraspecific variation over increased geographical scales has been predicted by theory and shown to be the case empirically (Bergsten et al. 2012). Increased intraspecific distance and decreased interspecific distance over larger geographical scales could reduce or remove the barcode gap making identification with DNA barcodes less effective (Bergsten et al. 2012). Most of the species present in the UK are not endemic to the UK. The sequences in BOLD are from specimens from countries all over the world. The large geographic scales that the sequences are from could cause a lack of a barcoding gap in some species that might have a clear barcoding gap if only sequences from UK specimens were used.

However, the low barcode coverage of many species makes it difficult to assess geographic variation in barcodes. Geographic variation within the three species analysed for intraspecific variation was analysed to explore whether geographic variation could impact measures of intraspecific variation. Although the three species do show a barcoding gap after the removal or sequences containing

errors and misidentifications, the maximum intraspecific variation remains relatively high (4.28-7.21%). If geographic variation impacts intraspecific variation, there might be situations where it would be easier and more effective to base analyses on references sequences originating from UK specimens only. For example, if geographic variation causes a lack of barcoding gap in some species when the specimens originate from a large geographic area, a database containing sequences from specimens from a smaller geographic area (e.g. UK plus neighbouring countries) might show a clear barcoding gap and enable more accurate identification of UK specimens. In addition, if specific assays were to be developed for species found in the UK, design could be easier and assays could have higher specificity if they are based on sequences from specimens from a reduced geographic area.

DNA-based identification of UK species with high geographic variation could be improved by having a database that only includes sequences from UK specimens (and specimens from countries that are similar to those from the UK). The reduction in intraspecific distance over the smaller geographical scales is more likely to provide a clear barcoding gap from other UK species. The geographic analyses of the three species analysed here show that geographic variation within species can be very variable. In all three species, the sequences form separate clusters with some sequences very similar to each other and others more distant. In the first two species, *I grammatica* and *L. fusca*, the sequences show clear geographically-related clustering, with sequences from individual countries tending to cluster together. The third species, *A. sulcicollis*, shows some clustering but sequences from Germany appear in different clusters rather than grouping closely together. Individual sequences that are positioned separately from all others could indicate errors (particularly where the genetic distance is high) but could also represent true variation within the species that is under-represented in BOLD public records so far. Where sequences form new clusters with other similar sequences, there is higher confidence that the dissimilarity is representing true variation (either within the species or to a very closely related species that is easily misidentified).

Sequences from UK specimens are available for two of these species (*I. grammatica* and *A. sulcicollis*). The sequences from UK specimens mostly cluster closely together in both species. In *I. grammatica*, sequences from the

UK group closely together (with the exception of one individual sequence which appears separately to all others) with sequences from Norway, France and Switzerland. A database containing these sequences would have lower maximum intraspecific distance and would provide accurate identification of UK specimens and enable the design of specific assays. In *A. sulcicollis*, the UK sequences cluster together with sequences from Norway and one of the sequences from Germany. For this species a database including just the sequences from the UK and Norway would provide accurate identification of UK specimens.

Specific primers are particularly used to screen samples for protected or non-invasive species. High intraspecific variation and low interspecific variation could prevent the design of primers that will only amplify the target species and not other closely related species. Designing primers only using sequences from specimens from the UK (and countries with similar sequences) would be easier and more effective. However, if targeting non-native species, it might be important to design assays based on sequences from a wider geographic area to ensure that assays can detect the sequences from individuals from other countries that might be introduced to the UK in the future. Horizon scanning for species that might arrive in the UK in the future (e.g. Roy et al. 2014) could provide a basis for which species need to include sequences from a wider area.

Where species do not have any representation from specimens from the UK (e.g. *L. fusca*), it is unknown if UK sequences would be similar to the sequences from any of the countries currently stored for that species or if they would cluster separately. The high intraspecific variation seen in some of these examples suggests that it is likely that sequences from one country might often be over two percent different from sequences from another country. If the geographic representation of barcodes for a species is very poor compared to the distribution of the species, specimens might not be accurately identified to species-level even though barcodes for that species are stored in BOLD. Ideally, each species would have at least one barcode stored from every country where it occurs. To ensure UK freshwater arthropods can be accurately identified using DNA-based methods, it is important that barcodes from UK specimens are stored in reference databases.

Studies on geographic variation in Arthropoda from different regions/habitats show varied conclusions. A study focused on spider (Araneidae, Lycosidae and Tetragnathidae) sequences from North America and Europe found that the barcode gap decreased with increasing geographic distance but that this did not prevent reliable identification of the species included in the study (Candek and Kuntner, 2014). Estimates of barcoding gaps in Odonata suggest that species identification might be limited by lack of barcoding gaps globally. However, this could be partly caused by errors and misidentifications in reference sequences and that barcoding gaps could be improved if targeting taxa at a local rather than global scale (Koroiva and Kvist, 2018). A study focused on aquatic diving beetles (Dytiscidae) throughout Europe found that the success of species identification decreased with increased geographical scale (Bergsten et al. 2012). It has been suggested that this decreased identification success in diving beetles could be due to them inhabiting discontinuous freshwater habitats that causes population fragmentation and high intraspecific differentiation (Huemer et al. 2014) so it is possible that geographic variation is more important for the UK freshwater taxa included in this study than for terrestrial taxa. The effects of geographic scale can only be accurately assessed with comprehensive and curated reference databases but there is clear evidence that geographical scale can impact the success of DNA-based identification in some taxonomic groups/habitats and therefore national barcoding initiatives are likely to provide increased accuracy in identification.

3.5.5 Conclusions

The overall barcode coverage of UK freshwater arthropods is good (73%) but the coverage is biased between classes and orders. Only 60% of species have coverage by public barcodes. Over a third of the species that have public barcodes stored have less than five sequences stored, limiting representation of intraspecific variation and the ability to curate stored sequences. There are very few species represented by sequences from specimens from the UK, so it is not possible to assess the similarity between sequences from the UK and other countries. Misidentifications of reference sequences stored in BOLD causes some species to lack barcode gaps. In addition, multiple species appear to have less than 2% divergence from each other and be placed within one BIN in BOLD.

Creation and curation of a UK barcode database for UK freshwater arthropods would enable prioritisation in UK sequencing projects for species with poor coverage; enable quality control of sequences to reduce the occurrence of misidentifications and errors that could impact accuracy of DNA-based identification; and lead to improved effectiveness of DNA-based identification of freshwater arthropods in the UK.

3.5.6 Recommendations

The creation and curation of a UK database of freshwater arthropods will enable UK biodiversity assessment and monitoring to benefit more fully from DNA-based identification methods. In order to ensure high quality references and prevent errors in public databases from perpetuating through the database, the UK database should be based on a core of 'gold standard' sequences where UK specimens have been identified by taxonomic experts, voucher specimens have been stored, and high-quality sequences have been produced (for example, as in FreshBase project sequencing).

Multiple sequences are needed for every species to capture intraspecific variation. Ideally, there should be specimens from different locations in the UK. Where sequences from specimens from other countries are similar to UK sequences, these sequences could also be included. It is also important to include sequences from other countries for non-native species so that new introductions can be easily identified and monitored.

Curation of UK freshwater arthropod sequences and prioritisation of UK sequencing is needed, and periodic re-evaluation is important to reanalyse barcoding gaps as more sequences become available. A suggested framework for curation of a database and prioritisation of barcoding UK freshwater arthropods is provided (Figure 3.9).

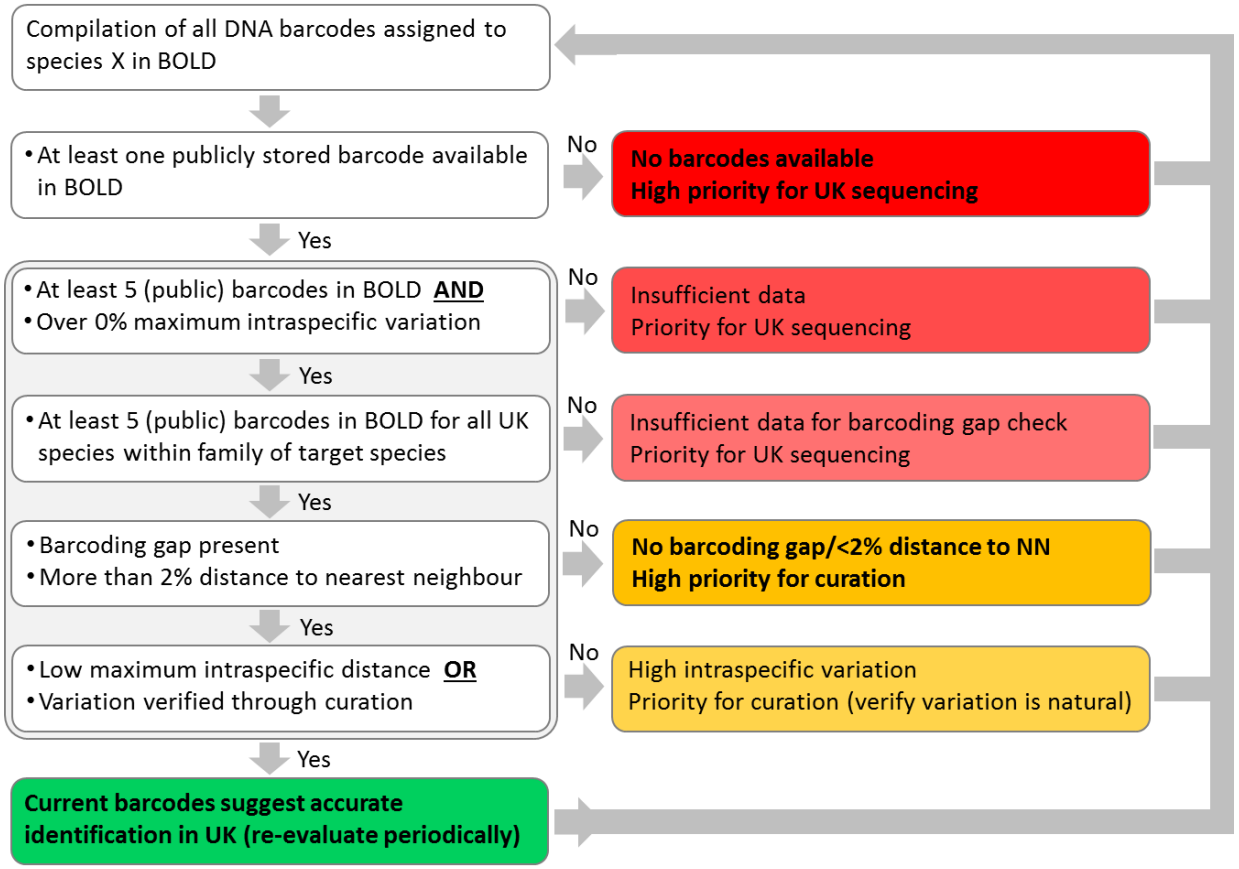

Figure 3.9 *Suggested framework for curation of a database and prioritisation of barcoding UK freshwater arthropods.*

4 Using metabarcoding to characterise freshwater zooplankton communities in lakes

4.1 Summary

In lentic freshwater ecosystems, zooplankton occupy a key central position in pelagic food webs, feeding on phytoplankton, bacteria, detritus and other zooplankton and being eaten by both invertebrate and vertebrate predators. Optimisation and validation of COI primers to enable characterisation of target freshwater zooplankton taxa could provide successful amplification and high taxonomic resolution for freshwater zooplankton biodiversity monitoring. This study aimed to develop and optimise a DNA-based methodology for monitoring freshwater zooplankton in the Lake District, United Kingdom (UK) through the optimisation of primers, validation of metabarcoding data and assessment of methodological decisions on ecological data.

This study optimised primers to amplify the target taxa and used them to provide very accurate and sensitive data on presence of zooplankton in bulk community samples. The methods enabled low abundance taxa, that could easily be overlooked in traditional morphological counts, to be detected and most species were identified to species-level, providing higher resolution than is usually obtained through morphological identification. Optimisation of bioinformatic analysis in combination with knowledge of the target taxa helped reduce false positives and negatives to produce more meaningful and reliable data. In addition, metabarcoding data provided some relative abundance data within samples.

Optimisation of primers, bioinformatic processing and data analysis can improve the detection of target taxa, reduce false positives and negatives, and improve relative abundance data. Validation of metabarcoding data against other data types for the samples can provide a better understanding of the strengths and weaknesses of metabarcoding data which can help ensure appropriate use of the data to enhance monitoring and assessment of freshwater biodiversity.

4.2 Introduction

4.2.1 DNA-based identification of freshwater biodiversity

Anthropogenic pressures on freshwater ecosystems are already increasing, and this is predicted to continue with rising human consumption alongside rapid environmental change (Darwall et al. 2018). As a result, freshwater biodiversity is in crisis (Reid et al. 2019). It is therefore essential that freshwater biodiversity change can be monitored effectively. Despite this critical need, most freshwater habitats have not been comprehensively assessed for their resident biodiversity, and many species are 'hidden' or difficult to observe, so there are large gaps in our current knowledge of these systems. It is vital that we develop methods that can provide rapid biodiversity assessments, monitor changes in biodiversity over time and in relation to anthropogenic impacts, and allow us to track the effects of management and restoration measures.

The use of DNA-based identification in freshwater ecosystems is increasing rapidly and can provide rapid assessments of the presence of a wide range of taxa from either environmental DNA (eDNA), or DNA extracted from bulk samples of for example, invertebrates. DNA-based methods show great potential for assessing and monitoring freshwater biodiversity but the results are highly dependent on multiple stages of the sampling and analytical process e.g. sampling design, primer choice, clustering of sequences, and taxonomic assignment. It is essential that methods are thoroughly optimised and validated for the target taxa so that the uncertainties surrounding each assessment can be fully understood and accuracy maximised by reducing false positives and negatives (Elbrecht and Leese 2017; Cristescu and Hebert 2018).

Research into using DNA-based identification for freshwater conservation has so far been biased towards vertebrates, especially fishes and amphibians (Belle et al. 2019b). However, freshwater invertebrates are potentially more important as diverse components of freshwater food webs and providers of essential ecosystem functions including energy transfer to higher trophic levels and influencing community organisation at lower trophic levels. Effective DNA-based identification of invertebrates from eDNA would therefore be valuable, but can be challenging as small taxa with hard exoskeletons shed less DNA into

the environment than large vertebrates (Tréguier et al. 2014; Harper et al. 2020). Many species also live in littoral or benthic habitats and might be under-represented in the surface water samples that are commonly used for eDNA analyses. DNA-based identification of invertebrates usually focuses instead on bulk samples of animals (Cristescu and Hebert 2018) that can be collected using the same methods as traditional invertebrate samples thus providing a more direct comparison to traditional monitoring. In addition, bulk samples avoid some of the many uncertainties with analysing eDNA – such as the origin, persistence and transportation of DNA (Dejean et al. 2011; Deiner and Altermatt 2014; Sansom and Sassoubre 2017) - that should make bulk sampling methods more tractable.

4.2.2 Zooplankton

The use of DNA-based methods could be particularly useful for zooplankton, especially using bulk samples. In lentic freshwater ecosystems, zooplankton occupy a key central position in pelagic food webs, feeding on phytoplankton, bacteria, detritus and other zooplankton and being eaten by both invertebrate and vertebrate predators. Zooplankton are an especially important food source for fishes, with most species feeding on zooplankton at some stage of life and some throughout their life (Moss 2010). Monitoring zooplankton communities can provide important data on environmental change because of their sensitivity to environmental conditions and value as aquatic ecosystem indicators (Jeppesen et al. 2011). However, morphological identification of zooplankton species is time-consuming and relies on taxonomic expertise. In contrast, DNA-based identification of bulk zooplankton samples could provide a more rapid method to monitor zooplankton communities and has the potential to provide higher taxonomic resolution than is usually gained through routine morphological identification. In order to use DNA-based identification for zooplankton community assessment, the methods need to be optimised and the results validated.

One of the key needs in developing DNA-based methods for freshwater zooplankton is the choice of genetic marker, which affects both the amplification success of different taxonomic groups and the taxonomic

resolution possible. In marine systems, many studies using DNA-based identification of zooplankton have used the nuclear 18S gene (e.g. Lindeque et al. 2013; Pearman et al. 2014; de Vargas et al. 2015; Chain et al. 2016) because it has more conserved binding sites for primers due to a relatively slower rate of evolution and therefore has high amplification success across the broad taxonomic groups found in planktonic samples. However, the slower evolution of the gene means that there is less interspecific variation, meaning that fewer taxa are likely to be resolved to species. In freshwaters, zooplankton are less diverse than those in marine habitats (Fernando 1994) so the broad taxonomic amplification afforded by 18S might not be necessary.

As an alternative to 18S, the most common marker for metabarcoding of animals is cytochrome c oxidase subunit I (COI) (Hebert et al. 2003b; Leray et al. 2013). The relatively faster rate of evolution of the mitochondrial COI gene provides high interspecific variation and therefore the potential to resolve more taxa to species-level. In addition, the reference databases for COI are comprehensive due to high barcoding effort for this marker (Ratnasingham and Hebert 2013) making taxonomic assignment of amplified sequences more accurate and precise with respect to taxonomic resolution. The high interspecific variation of this marker can make it difficult to amplify across taxonomic groups due to a lack of conserved regions for primer binding (Deagle et al. 2014) but it is suggested that these challenges can be overcome using well-designed degenerate primers (primers that enable more than one base possibility at a particular position) (Elbrecht and Leese 2017).

Amplification of freshwater copepods and cladocerans has been shown to be challenging using COI (Zhan et al. 2014) but zooplankton communities have been successfully characterised using COI in marine (Clarke et al. 2017) and freshwater systems (Yang et al. 2017b). These characterisations used a versatile primer pair, designed to amplify a broad range of metazoan taxa (Leray et al. 2013). Thorough evaluation of primers and development of improved primers for the target ecosystem or community can result in higher amplification success and taxonomic resolution (Elbrecht and Leese 2017). Optimisation and validation of COI primers to enable characterisation of target freshwater

zooplankton taxa could, therefore, provide successful amplification and high taxonomic resolution for freshwater zooplankton biodiversity monitoring.

4.2.3 Aims and hypotheses

The overall aim of this study was to develop and optimise a DNA-based methodology for monitoring freshwater zooplankton in the Lake District, United Kingdom (UK) in order to facilitate ecological inferences on community dynamics. The specific aims were to i) optimise primers for metabarcoding bulk samples of target zooplankton taxa using *in silico* analysis and PCR testing, ii) validate metabarcoding data using morphological count data to understand the strengths and weaknesses, and iii) assess the effects of primer choice and data filtering on both presence/absence data and read abundance data. This study tested the specific hypothesis that optimised and validated metabarcoding of bulk zooplankton samples can provide meaningful data on potential prey communities in lake ecosystems.

4.3 Methods

Target taxa for the lakes involved in this study were identified from long-term monitoring data from the UK Centre for Ecology and Hydrology (UKCEH). The dominant pelagic zooplankton genera were identified as: a dipteran: *Chaoborus* (Chaoboridae); four diplostracans: *Bosmina* (Bosminidae), *Ceriodaphnia* (Daphniidae), *Daphnia* (Daphniidae), *Leptodora* (Leptodoridae); two cyclopoid copepods: *Cyclops* (Cyclopidae) and *Mesocyclops* (Cyclopidae); and a calanoid copepod: *Eudiaptomus* (Diaptomidae).

4.3.1 Sample collection

Zooplankton samples for optimisation of primers were collected from four lakes in South Cumbria (UK) between July and November 2019: Blelham Tarn, Esthwaite Water, Loughrigg Tarn, Windermere South Basin and Windermere North Basin (Figure 4.1). Zooplankton were collected from each lake by triplicate vertical hauls using a plankton net (250 µm mesh). The purpose of these samples was to provide multiple specimens of the target genera for use in optimising the molecular methods so standardisation of depth, number of hauls, and position in the lake did not need to be standardised. Individuals of

each of the targeted genera were sorted and starved overnight in filtered (0.2 µm filter) lake water at room temperature. Starved individuals were rinsed and frozen at -80°C.

Zooplankton community samples were collected mid-month from the deepest point in Esthwaite Water from August to October 2018. The community samples were to be representative of the zooplankton community at specific times and depths so were standardised for all sampling points. Community samples were collected using a closing net (120 µm mesh). Three depths were sampled by triplicate vertical hauls: 0-4 m, 4-8 m, 8-12 m. Approximately 50 ml of carbonated water was added to each sample (5 ml at a time and swirled gently) and samples were placed on ice to narcotise the zooplankton in order to minimise predation during transportation. In the laboratory, each sample was re-suspended, thoroughly mixed, and split in half by transferring the sample repeatedly between two beakers and then ensuring each beaker contained half the total volume. One half was then stored for morphological taxonomic analysis by filtering the sample on to a 100 µm mesh filter, rinsing the zooplankton with distilled water, and then rinsing them into a universal tube with 70% ethanol. The other half was stored for DNA extraction by filtering the zooplankton on to a 100 µm mesh filter, rinsing the zooplankton with distilled water, and then filtering the sample on to a 40 µm nylon gauze filter using a vacuum pump system. The filter was then folded, transferred to a centrifuge tube, and frozen at -80°C.

Figure 4.1 *Map showing the locations of sampling sites.*

4.3.2 Morphological taxonomic analysis

Each sample was re-suspended in distilled water and a 5 ml sub-sample was taken using a Hensen-Stempel pipette. All zooplankton individuals in the sub-sample were identified and counted at genus-level (zooplankton are not often identified and counted at species-level for the Cumbrian Lakes monitoring scheme due to the time and level of taxonomic expertise necessary to achieve this for multiple samples). The sub-sample was then returned to the sample and this was repeated for a total of three 5 ml sub-samples. The mean of the taxon counts from the three sub-samples was taken for each sample and multiplied to provide an estimated count for the full sample (70 ml).

4.3.3 Optimisation: Primer selection, modification and in silico analysis

In order to optimise primers for the target zooplankton taxa: potential primers were selected from the literature; reference sequences for the target taxonomic groups were obtained from reference databases; selected primers were tested *in silico* against the reference sequences, and primers were modified in order to improve amplification success for the target groups.

Potential primers for amplifying the target freshwater zooplankton taxa were initially selected from those available in the literature by searching for primers that had been used to amplify freshwater invertebrates. Eighteen potential COI primers were identified from the literature for testing (Table 4.1). Primers were chosen that targeted amplicons within the Folmer region of the COI gene and that had successfully amplified invertebrate taxa in previous studies. These primers included the standard Folmer primers (LCO1490 and HCO2198) as well as other primers that target the same primer-binding sites (Figure 4.2). As the Folmer primers (or other primers targeting these primer-binding sites) are usually used in generating barcodes, these regions are not usually included in reference databases (as the use of ambiguity codes in primers used to generate the barcode sequence can cause the inclusion of potentially incorrect bases). This causes a challenge for *in silico* testing of many of the previously used primer pairs as they mostly use one of the Folmer primer-binding sites.

Table 4.1 *COI primers selected for testing in this study.*

Primer name	Strand	Sequence	Citation
LCO1490	F	GGTCAACAAATCATAAAGATATTGG	Folmer et al. (1994)
HCO2198	R	TAAACTTCAGGGTGACCAAAAAATCA	Folmer et al. (1994)
BF1	F	ACWGGWTGRACWGTNTAYCC	Elbrecht and Leese (2017)
BF2	F	GCHCCHGAYATRGCHTTYCC	Elbrecht and Leese (2017)
BF3	F	CCHGAYATRGCHTTYCCHCG	Elbrecht et al. (2019)
BR2	R	TCDGGRTGNCCRAARAAYCA	Elbrecht and Leese (2017)
ArF5	F	GCICCIGAYATRKCITTYCCICG	Gibson et al. (2014)
mlCO1intF	F	GGWACWGGWTGAACWGTWTAYCCYCC	Leray et al. (2013)
mlCOIintF-XT	F	GGWACWRGWTGRACWITITAYCCYCC	Wangensteen et al. (2018)
Fol-degen-rev	R	TANACYTCNGGRTGNCCRAARAAYCA	Yu et al. (2012)
jgHCO2198	R	TAIACYTCIGGRTGICCRAARAAYCA	Geller et al. (2013)
LepF1	F	ATTCAACCAATCATAAAGATATTGG	Hebert et al. (2004)
MLepF1-Rev	R	CGTGGAAAWGCTATATCWGGTG	Brandon-Mong et al. (2015)
230_R	R	CTTATRTTRTTTATICGIGGRAAIGC	Gibson et al. (2015)
fwhF2	F	GGDACWGGWTGAACWGTWTAYCCHCC	Vamos et al. (2017)
fwhR2n	R	GTRATWGCHCCDGCTARWACWGG	Vamos et al. (2017)
ZBJ-ArtF1c	F	AGATATTGGAACWTTATATTTTATTTTTGG	Zeale et al. (2011)
ZBJ-ArtR2c	R	WACTAATCAATTWCCAAATCCTCC	Zeale et al. (2011)

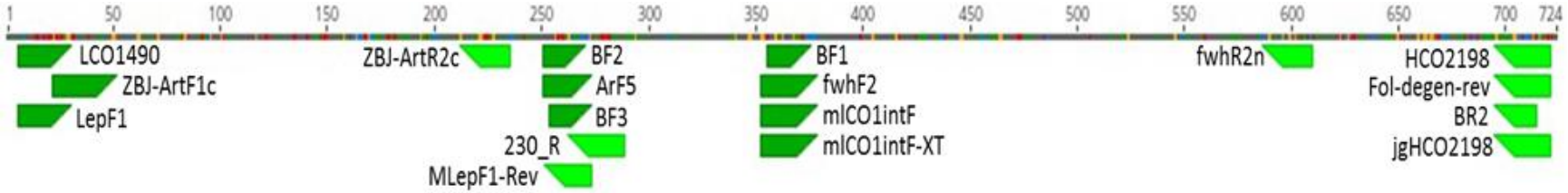

Figure 4.2 *Positions of the eighteen selected primers on the Folmer barcoding region of the COI gene (image created in Geneious Prime).*

In order to obtain sequences to test the full set of selected primers, reference databases: Barcode of Life Data Systems (BOLD v3 (Ratnasingham and Hebert 2013)) and GenBank (Benson et al. 2012) were searched (October 2019) for full COI sequences so that the primer-binding regions would be present in the sequences. Full COI sequences were not available for most of the target genera so databases were searched for any taxa within the same orders as taxa that had been found in UKCEH long term monitoring of zooplankton (Anomopoda, Ctenopoda, Calanoida, Cyclopoida and Diptera). In addition, sequences for any taxa within two orders of rotifers (Bdelloidea and Monogononta) and parasitic crustaceans (Arguloida) were also downloaded as these taxa were also found in recent zooplankton samples. The downloaded COI sequences were grouped into orders, aligned and a consensus sequence was generated for each of eight orders using MAFFT v7 (Katoh and Standley 2013) in Geneious Prime (Version 2019.2).

The selected primers (Table 4.1) were then tested *in silico* against the consensus sequences in Geneious Prime (Version 2019.2). Primer binds were tested allowing three mismatches overall but no mismatches within three bp of the 3' end (Dieffenbach et al. 1993). The number of target orders (Anomopoda, Ctenopoda, Calanoida, Cyclopoida, Monogonota, Bdelloidea, Diptera and Arguloida) that each primer was able to bind to and the total number of mismatches that occurred across all target order consensus sequences were recorded (Table 4.2). Primers that were able to bind to all eight of the order consensus sequences and had less than ten total mismatches across orders were taken forward for further testing (see highlighted primers in Table 4.2).

Table 4.2 The number of order consensus sequences that each primer binds to and the total number of mismatches across all primer binding positions in *in silico* tests. Primers chosen for further testing are highlighted.

Name	Strand	Number of orders with primer binds	Total number of mismatches
LCO1490	F	2	6
HCO2198	R	6	13
BF1	F	8	6
BF2	F	7	7
BF3	F	8	4
BR2	R	8	1
ArF5	F	8	13
mlCO1intF	F	6	9
mlCOIintF-XT	F	6	8
Fol-degen-rev	R	8	0
jgHCO2198	R	3	4
LepF1	F	4	11
MLepF1-Rev	R	3	5
230_R	R	6	7
fwhF2	F	8	8
fwhR2n	R	8	7
ZBJ-ArtF1c	F	3	5
ZBJ-ArtR2c	R	4	11

The chosen six primers (Table 4.3) were checked against the consensus order sequences and local sequences where possible (where primer positions were internal to the standard Folmer primer binding positions). Four of the primers (BF3, fwhF2, fwhR2n and Fol-degen-rev) were modified to reduce mismatches with target taxa. Modifications involved increasing degeneracy and/or reducing primer length. Degeneracy was reduced and the length was shortened for the reverse primer, Fol-degen-rev, as the higher degeneracy appeared to be unnecessary for the target taxa. The binding positions of the six primers are shown in Figure 4.3. The nine possible primer pairs created using the chosen forward and reverse primers target amplicons between 205 and 418 base pairs in length (Table 4.4).

Table 4.3 The six selected primers and their characteristics (modifications in red).

Name	Strand	Sequence (5' to 3')	Primer length	Degeneracy
BF3A	**F**	CCHGAYATRGCHTTYCCNCG	20	288
fwhF2B	**F**	ACWGGDTGAACWGTWTAYCCNCC	23	192
BF1	**F**	ACWGGWTGRACWGTNTAYCC	20	128
fwhR2nB	**R**	GTRATWGCHCCNGCTARWACNGG	23	768
BR2	**R**	TCDGGRTGNCCRAARAAYCA	20	192
Fol-degen-revA	**R**	~~TAN~~ACYTCHGGRTGNCCRAARAAYCA	23	384

Figure 4.3 *Positions of the six primers chosen for further testing on the Folmer barcoding region of the COI gene (standard Folmer primers LCO1490 and HCO2198 shown for reference) (image created in Geneious Prime).*

Table 4.4 *The nine possible primer pairs created using the three forward and three reverse primers chosen from* in silico *testing and the targeted amplicon length of each pair.*

Pair	Forward primer	Reverse primer	Product length
1	fwhF2B	fwhR2nB	205 bp
2	BF1	fwhR2nB	208 bp
3	BF3A	fwhR2nB	310 bp
4	fwhF2B	BR2	313 bp
5	fwhF2B	Fol-degen-revA	313 bp
6	BF1	BR2	316 bp
7	BF1	Fol-degen-revA	316 bp
8	BF3A	BR2	418 bp
9	BF3A	Fol-degen-revA	418 bp

4.3.4 Optimisation: DNA extraction

Prior to DNA extraction, zooplankton individuals were thawed and rinsed three times in double-distilled water (ddH2O). Multiple rinsed individuals for each genus were pooled and DNA was extracted from the sample using Qiagen DNeasy Blood and Tissue kits, following the manufacturer's protocol in order to generate a high-quality DNA template for use in testing primer pairs. This was repeated for each of the following genera: *Chaoborus, Bosmina, Ceriodaphnia, Daphnia, Leptodora, Cyclops, Mesocyclops,* and *Eudiaptomus*. The concentration of extracted DNA was assessed using a Nanodrop 1000 Spectrophotometer. DNA concentrations were equalised at approximately 20 ng/µl.

Following low amplification success with *Bosmina* DNA tests, the DNA template concentrations were re-tested with a Qubit 3 Fluorometer. Extracted *Bosmina* DNA was below the threshold of the Qubit (<0.5 ng/ml) suggesting the low success for *Bosmina* primer tests may be due to low concentration of DNA template rather than primer-template mismatch. This low DNA concentration could have been missed originally due to over-estimation of DNA from species with chitinous carapaces when measured with absorbance-based quantification methods due to the high refractive index of chitin at 260 nm (Athanasio et al. 2016). Athanasio et al., (2016) compared extraction methods for obtaining high-quality DNA from species with chitinous carapaces and recommended MasterPure Complete DNA & RNA Purification kits.

DNA was extracted from 24 *Bosmina* individuals using a MasterPure Complete DNA & RNA Purification kit following the manufacturer's protocol. The concentration of DNA was assessed using Qubit 3 Fluorometer and provided a value of 4.4 ng/ml.

In order to assess general amplification success and choose annealing temperatures for each of the nine primer pairs, a mock community using DNA template from abundant local zooplankton genera from three different orders was created. Equal volumes of DNA template from the genera: *Chaoborus*, *Daphnia* and *Eudiaptomus* were combined for use in an initial gradient PCR.

4.3.5 Optimisation: Gradient PCR and single-taxon primer tests

Mock community gradient PCRs for the nine primer pairs were run on a Veriti 96-well Thermal Cycler. PCRs were set up with 2x Amplitaq Gold 360 Master Mix, 0.5 µM of each primer, 2 µl of DNA template, and molecular grade water to make up to a total volume of 20 µl. One positive control (using LCO1490 and HCO2198 (standard Folmer primers) and one negative control (no template DNA) were included. The following thermocycling protocol was used: initial denaturation at 95°C for 10 min, then 40 cycles of: 95°C for 1 min, a gradient of annealing temperatures from 46-56°C for 45 seconds, then extension at 72°C for 1 min, followed by a final extension of 72°C for 7 min. PCR success was determined by visualising amplicons on a 0.8% agarose gel.

Single-taxon PCRs for the eight zooplankton genera with each of the nine primer pairs were run on a Veriti 96-well Thermal Cycler. PCRs were set up as for the gradient PCR and the same thermocycling protocol was used except that the annealing temperature was fixed at 48°C for 45 seconds. PCR success was determined by visualising amplicons on a 0.8% agarose gel.

A single-taxon PCR for *Bosmina* with the nine primer pairs was repeated using the new MasterPure extracted *Bosmina* DNA template. The PCR was set up identically to the previous PCRs. PCR success was determined by visualising amplicons on a 0.8% agarose gel.

4.3.6 Community samples: DNA extraction

Prior to DNA extraction, community samples for metabarcoding were thawed and zooplankton were transferred from the nylon filter into a 1.5 ml tube using a sterile spatula. The sample was then homogenised using a plastic pestle and any material attached to the pestle was returned to the sample using a sterile needle. DNA was extracted from the community sample using MasterPure Complete DNA and RNA Purification kits following the manufacturer's protocol. The concentration of DNA was assessed using a Qubit 3 Fluorometer.

4.3.7 Community samples: Metabarcoding

Primer pairs 1 and 5 (targeting amplicons of 205 bp and 313 bp) were selected for DNA metabarcoding (see Results 4.3.1 for selection details). The First Step

PCRs were run on a Veriti 96-well Thermal Cycler. The PCRs were set up with 2x Amplitaq Gold 360 Master Mix, 0.5 µM of each primer, 2 µl of DNA template, and molecular grade water to make up a total volume of 25 µl. One positive control, using *Chaoborus* (single taxon) DNA template, and one negative control were used in each PCR. The following thermocycling protocol was used: initial denaturation at 95°C for 10 min, then 35 cycles of: 95°C for 1 min, 49°C for 45 seconds, then extension at 72°C for 1 min. Followed by a final extension of 72°C for 7 min. PCR success was determined by visualising amplicons on a 0.8% agarose gel.

First step PCR product was cleaned up using a ZR-96 DNA Clean-up Kit (Zymo) following manufacturers protocol. MiSeq adapters and 8nt dual-indexing barcode sequences were added during a second step of PCR amplification. 1 µl of cleaned DNA was used in the second round PCR. The PCR was set up with 0.25 µl HiFi Taq Q5 NEB, 5 µl reaction buffer, 5 µl high GC, 0.5 µl dNTPs, 5 µl index primers, 1 µl DNA template, and molecular grade water with a total volume of 25 µl. Two single-taxon (*Chaoborus* and *Mesocyclops*) DNA templates (primer pair 1) separate and combined were used as positive controls and three clean-up blanks were included. The following thermocycling protocol was used: initial denaturation at 95°C for 2 min, then 8 cycles of: 95°C for 15 seconds, 55°C for 30 seconds, then extension at 72°C for 30 seconds, followed by a final extension at 72°C for 10 min. PCR success was determined by visualising amplicons on a 1.5% agarose gel.

The following sequencing steps were carried out by Tim Goodall (UKCEH Wallingford). Libraries were normalised using SequalPrep Normalization Plate Kit (Thermo Fisher Scientific) and quantified using Qubit dsDNA HS kit (Thermo Fisher Scientific). The pooled library was further purified by gel extraction (QIAquick, Qiagen) and diluted to achieve 400 pM with 7.5% Illumina PhiX. Denaturation of each library was achieved with addition of 10% final volume of 2N NaOH, incubated at room temperature for 5 minutes followed by neutralisation with an equal volume of 2N HCl. The library was then diluted to its load concentration with Illumina HT1 Buffer. A final denaturation was performed by heating to 96°C for 2 minutes followed by cooling in crushed ice. Sequencing was performed on Illumina MiSeq using V3 600 cycle reagents.

4.3.8 Community samples: Bioinformatic processing

Pre-processing of raw Illumina MiSeq paired-end reads was done using the MetaWorks v1.8.1 pipeline available from https://github.com/terrimporter/MetaWorks (Porter and Hajibabaei 2020a). MetaWorks is an automated SnakeMake pipeline that runs in a conda environment. Specifications are set by editing the configuration file. The demultiplexed paired-end reads from Illumina MiSeq were merged using SEQPREP v1.3.2 from bioconda using the default MetaWorks settings of: minimum Phred quality score of 13 in the overlap region and at least a 25 bp overlap. Primers were trimmed based on their sequences using CUTADAPT v3.2 from bioconda. The forward primer is trimmed first and the output from this step is used as the input for trimming the reverse reads. The MetaWorks default settings were used for the minimum Phred quality score at the ends (≥20) and the allowance of no more than 3 Ns. The minimum length of trimmed reads was kept at the default setting of 150 bp. Reads were dereplicated, using VSEARCH v2.15.2 from bioconda, only retaining unique sequences. Exact sequence variants (ESVs) were generated using the unoise algorithm and rare clusters (clusters containing less than three reads) were removed with the uchime3_denovo algorithm. Putative chimeric sequences were removed using the uchime3_denovo algorithm in VSEARCH.

Taxonomic assignment of ESVs was done using BOLDigger v1.2.5 available from https://github.com/DominikBuchner/BOLDigger (Buchner and Leese 2020). BOLDigger is a python program that queries fasta files against the BOLD Systems databases (including private and early-release data) and returns taxonomic assignment and additional metadata for each sequence. BOLDigger also uses algorithms to help choose the best hit and flags suspicious hits. Best hit options were compared and checked manually to determine the best option. The BOLDigger best hit option was chosen. This option uses thresholds (98% species-level, 95% genus level, 90% family level, 85% order level, <85% class level) and chooses the most common hit above the threshold and flags suspicious hits so they can be manually checked.

The BOLDigger best hit table was joined to the ESV table that was produced using the MetaWorks pipeline. Positive and negative controls were checked for unexpected sequence reads. ESVs were filtered for target taxa only (all taxa in the phylum Arthropoda). ESVs were clustered manually using taxonomic assignments for each dataset because the use of one sequence similarity threshold for clustering across taxonomic groups cannot capture the different levels of inter- and intra-specific variation that exist for different taxa (Goldstein et al. 2000; Porazinska et al. 2010) (see Chapter 3 for analysis of variation within and among species). Using a fixed threshold across taxa can lead to over- and/or under-clustering of ESVs (Bonin et al. 2021). As the aim of clustering is to absorb artefactual sequences caused by PCR or sequencing error and to provide species-like units of biodiversity, over- and under-clustering of sequences results in false representations of the traditional unit used in biodiversity studies (Alberdi et al. 2018; Porter and Hajibabaei 2020b; Antich et al. 2021).

The focus here was on comparing metabarcoding data to genus-level morphological taxonomic data so ESVs were filtered to remove sequences that were less than 95% similar to reference sequences. A manual check of these ESVs showed they were mostly very low read numbers identified to family-level and 2-4 (short and long amplicon datasets respectively) ESVs that were only assigned to class-level. ESVs assigned to a single species showed a clear 'head-tail structure' (a cluster of ESVs assigned to a particular species consisted of one or two ESVs with high read abundance and high similarity to a reference sequence (99-100% similarity) and multiple ESVs with low read abundance and lower similarity) (Porazinska et al. 2010; Macheriotou et al. 2019). These 'tail' ESVs are likely to be caused by intraspecific variation and PCR/sequencing errors and artificially inflate measures of diversity if not clustered together. In order to produce species-like units that could be compared to morphological taxonomic data, these low abundance 'tail' ESVs were manually clustered together with the dominant 'head' ESV(s). This ensured that the pattern was consistent across samples and ESVs were not clustered if the pattern was inconsistent.

The metabarcoding data were not filtered using a set filter threshold as part of the bioinformatic processing as detection of small and low abundance taxa was of interest and the effect of filtering was examined during data analysis.

4.3.9 Community samples: Data analysis

Read abundance data were converted to percentage arthropod read abundance within the sample. Percentage read abundance was also recalculated as a percentage of all crustacean reads rather than all arthropod reads.

For read abundance data, a taxon was counted as present in a sample if it had >0% reads for unfiltered data or ≥0.05% reads for filtered data. For morphological count data, a taxon was counted as present in a sample if it had >0 individuals counted in the sample. Read abundance data and morphological count data were said to 'agree' for a sample if a taxon was counted as present in both data types or absent in both data types. The data types were said to 'disagree' for a sample if a taxon was only present in either the read abundance data or the morphological count data.

Read abundance was not used as a measure of absolute abundance of individuals as many factors bias the number of reads e.g. relative sizes of the taxa, extraction methods, and primer bias (Elbrecht and Leese 2015; Pinol et al. 2015). However, metabarcoding data can provide abundance-like data (Piñol et al. 2019) so relationships between percentage read abundance and morphological count data were explored here to assess whether relative abundance of taxa within or among samples can be inferred from percentage read abundance.

4.4 Results

4.4.1 Optimisation: Gradient PCR and single-taxon primer tests

All primer pairs successfully amplified the mock community DNA and 48°C was chosen as the optimum temperature to run a PCR with all primer pairs in together (see Appendix S2: Figure S2.1 for electrophoresis gel photos showing results of gradient PCR for each of the nine primer pairs).

Primer tests with single-taxon DNA template from the eight target genera showed low success for the *Bosmina* DNA template (Appendix S2: Figure S2.2). The PCR was repeated with a higher-quality *Bosmina* DNA template to get a more accurate representation of primer success for this taxon (Appendix S2: Figure S2.3). Gel electrophoresis results showed primer pairs 4 and 5 to be most successful for the target taxa (Table 4.5). Primer pairs 1, 2, 3, 8, 9 failed to amplify the targeted amplicon from some target taxa. Although, pairs 6 and 7 amplified all target taxa, they appeared to amplify a slightly larger amplicon from *Daphnia* DNA and so were less consistent than pairs 4 and 5. Pairs 4 and 5 use the same region and target amplicons of 313-316 bp in length so are likely to have similar results. Primer pair 5 was chosen from this group to test metabarcoding bulk community samples. In addition, primer pair 1 was chosen as it performed very well on all but one genera (*Bosmina*) in PCR tests and would be useful in metabarcoding more degraded eDNA as it targets a shorter amplicon.

Table 4.5 *Number of target taxa (total of 8 target taxa) amplified by the nine primer pairs (primer pairs chosen for metabarcoding community samples are highlighted).*

Primer pair	Product length	Number of target taxa successfully amplified
1	205 bp	7
2	208 bp	5
3	310 bp	7 (5 very faint bands)
4	313 bp	8
5	313 bp	8
6	316 bp	8 (*Daphnia* band slightly bigger than expected)
7	316 bp	8 (*Daphnia* band slightly bigger than expected)
8	418 bp	7 (*Ceriodaphnia* band slightly smaller than expected)
9	418 bp	7 (*Ceriodaphnia* band slightly smaller than expected)

4.4.2 Validation: Overall primer pair comparison

Reads from both the long (313 bp) and short (205 bp) amplicons were primarily (93%) from the target phylum, Arthropoda. The majority of the non-target amplification for both amplicons was from the phylum Rotifera. Thirteen arthropod genera were amplified and assigned using both amplicons (Figure 4.4). Four genera (*Bosmina, Diaphanosoma, Acanthocyclops and Thermocyclops*) could only be amplified and assigned using the long amplicon and one genus (*Eucyclops*) could only be amplified and assigned by the short amplicon. Apart from *Bosmina* (which PCR tests showed would not be amplified by primer pair 1 (short amplicon)), all the target genera were amplified and assigned using both amplicons.

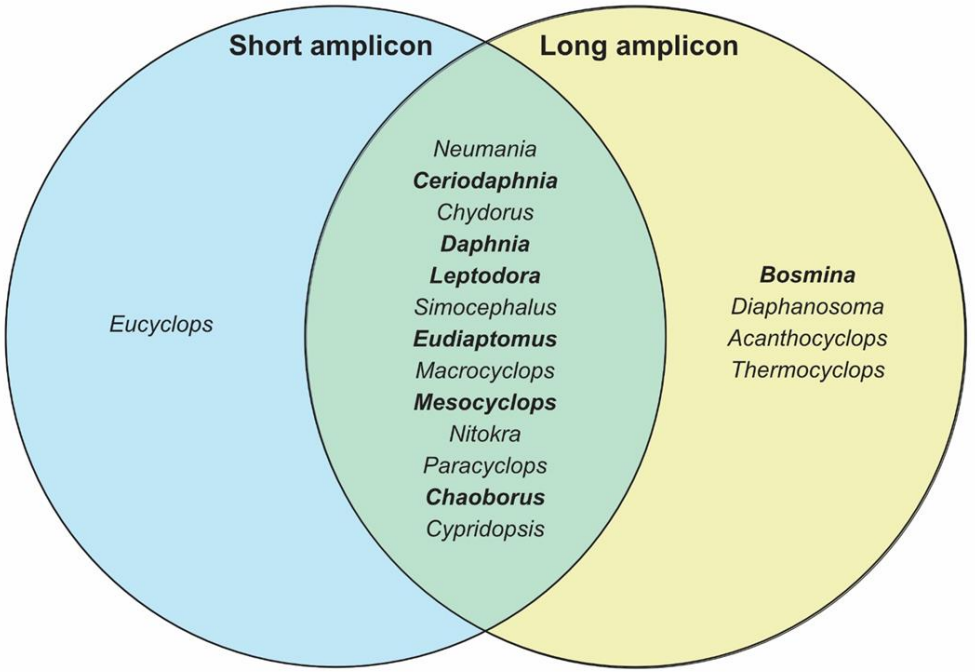

Figure 4.4 *Genera amplified by primer pair 1: short amplicon (205 bp) and primer pair 5: long amplicon (313 bp) that were chosen for metabarcoding from primer tests. The seven target genera from the primer tests that are present in the community samples are shown in bold font (Cyclops was not found in the community samples from metabarcoding or morphological taxonomic analysis).*

4.4.3 Validation: Presence/Absence

Detecting the presence of each of the target genera in the nine samples from Esthwaite Water using metabarcoding showed an overall agreement with morphological counts of 87% for the long amplicon and 78% for the short amplicon (long amplicon: 81% if *Bosmina* points are excluded) (Figure 4.5, a and b). Three genera (*Daphnia, Eudiaptomus* and *Mesocyclops*) showed consistent agreement between reads from both amplicons and the morphological count data. *Bosmina* and *Ceriodaphnia* were not always detected by the read data when they were present in the count data (three samples for the long amplicon and five for the short amplicon (these primers were unable to amplify *Bosmina*). *Leptodora* and *Chaoborus* were present in the read data but not the count data for five samples for the long amplicon and nine samples for the short amplicon.

When the metabarcoding read data were filtered to remove reads below 0.05% of target (arthropod) reads in the sample, the overall agreement in presence/absence of the target genera remains the same using the long amplicon and increases by 3% using the short amplicon (Figure 4.5, c and d). Although the overall agreement using the long amplicon remained at 87%, some identities of the detections changed. The presence of *Bosmina* in two samples, based upon count data, was not detected by metabarcoding after filtering. Prior to filtering, metabarcoding detections of *Bosmina* in these samples were in agreement with the count data. The presence of *Chaoborus* was not detected in two samples where it had been detected using the unfiltered data but was not present in the count data (improving the agreement with the count data). After filtering the short amplicon data, *Leptodora* and *Chaoborus* detections were lost for three samples where they were also absent in the count data. This improved overall levels of agreement between metabarcoding and morphological counts. However, lack of detection of *Eudiaptomus* in one sample where it was present in the count data led to a new disagreement point.

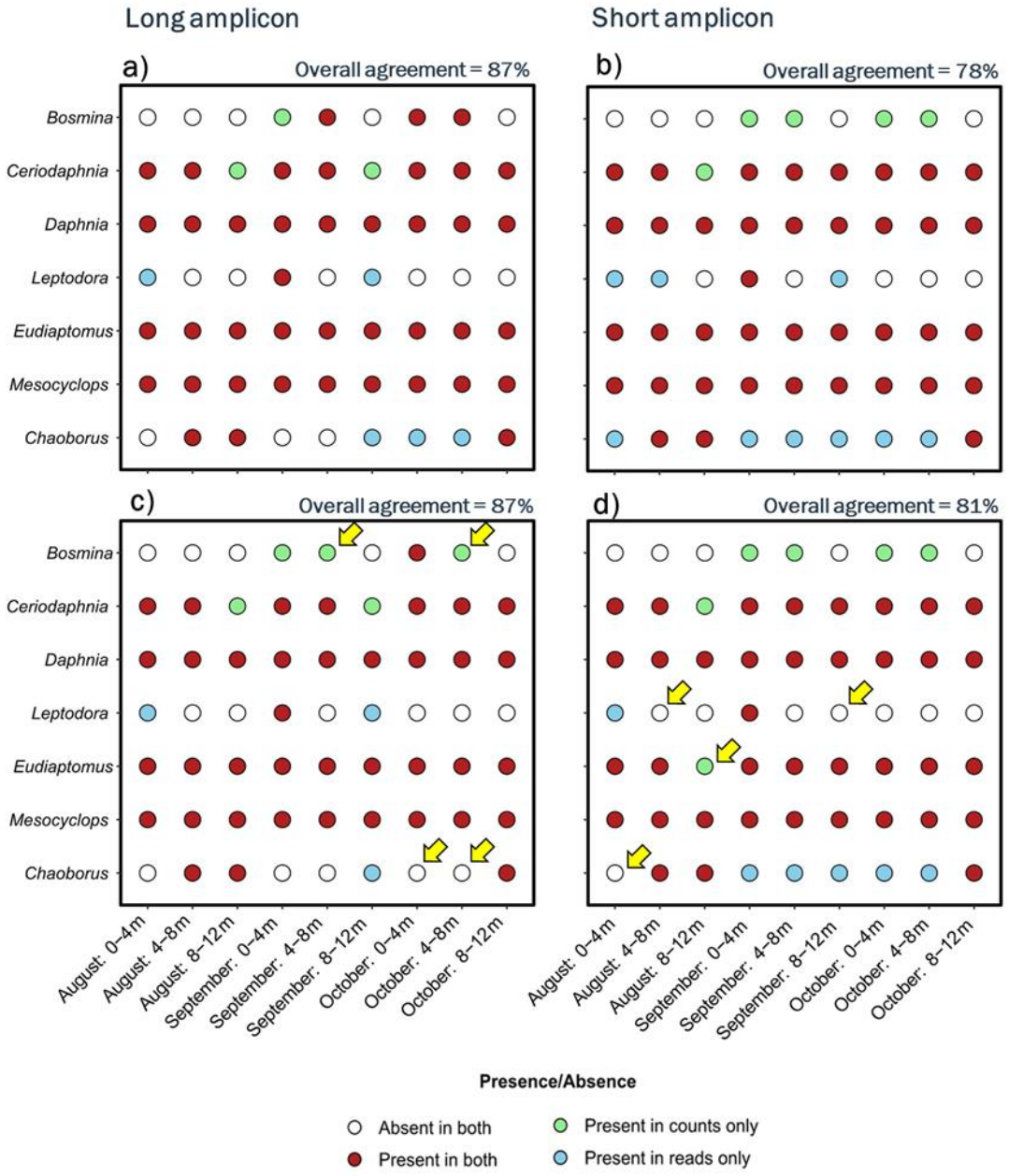

Figure 4.5 *Agreement in presence/absence of the target genera between metabarcoding and morphological count data for the long and short amplicons. No filtering of metabarcoding data (a and b) and metabarcoding reads filtered at 0.05% (c and d). Arrows on (c) and (d) highlight the effect of the 0.05% filter on the presence/absence of the target genera.*

The accuracy with which target genera were detected using metabarcoding data varied depending on which amplicon was used and whether very low reads were filtered out of the dataset (Figure 4.6). Detection of *Mesocylops* and *Daphnia* presence were unaffected by amplicon choice or filtering.

Detection of *Ceriodaphnia* was affected by amplicon choice (the short amplicon detected *Ceriodaphnia* in one more sample than the long amplicon) but the detections with both amplicons were unaffected by filtering out low abundance reads. In contrast, the detection of *Eudiaptomus* was unaffected by amplicon choice in the unfiltered data but filtering out low abundance reads caused *Eudiaptomus* detections to be lost using the short amplicon. *Bosmina* could only be detected with the long amplicon and was detected in fewer samples than in the count data. Filtering out very low read abundance further reduced the number of samples with positive detections for *Bosmina*.

Detections of *Chaoborus* and *Leptodora* were affected by both amplicon choice and filtering. Both were detected in more samples when using metabarcoding than in the morphological count data, for both amplicons. The short amplicon detected both genera in more samples than the long amplicon in unfiltered data. Filtering out very low read abundances reduced the number of samples with detectable *Chaoborus* for both amplicons but only reduced the number of samples with *Leptodora* presence for the short amplicon data. The number of samples with positive detections using read data remained higher than that of detections using count data with 0.05% filtering.

The three sequenced negative controls (short amplicon) each contained very low reads (maximum of seven) for the two taxa that were used as positive controls (*Chaoborus* and *Mesocyclops*) but no other taxa. This low-level cross contamination between the positive and negative controls suggests that a very small number of reads (e.g. <10) from *Chaoborus* and *Mesocyclops* in the samples could be caused by contamination from the positive controls rather than from DNA in the samples. In most samples, these taxa had very high read abundance (mean = 12033, 8978 respectively) so an addition of <10 reads would have a very small effect on the overall percentage read abundance. However, three samples had much lower read abundance for *Chaoborus* (August 0-4 m =

9 reads, September 4-8 m = 9 reads, October 4-8 m = 12 reads) and, although these are all slightly higher than the read abundance seen in the negative controls, it is possible that they are caused by cross-contamination from the positive controls rather than from DNA in the samples. These are all samples where *Chaoborus* was detected by metabarcoding but not by the morphological counts but only one (August 0-4 m) was affected by filtering the data at 0.05%.

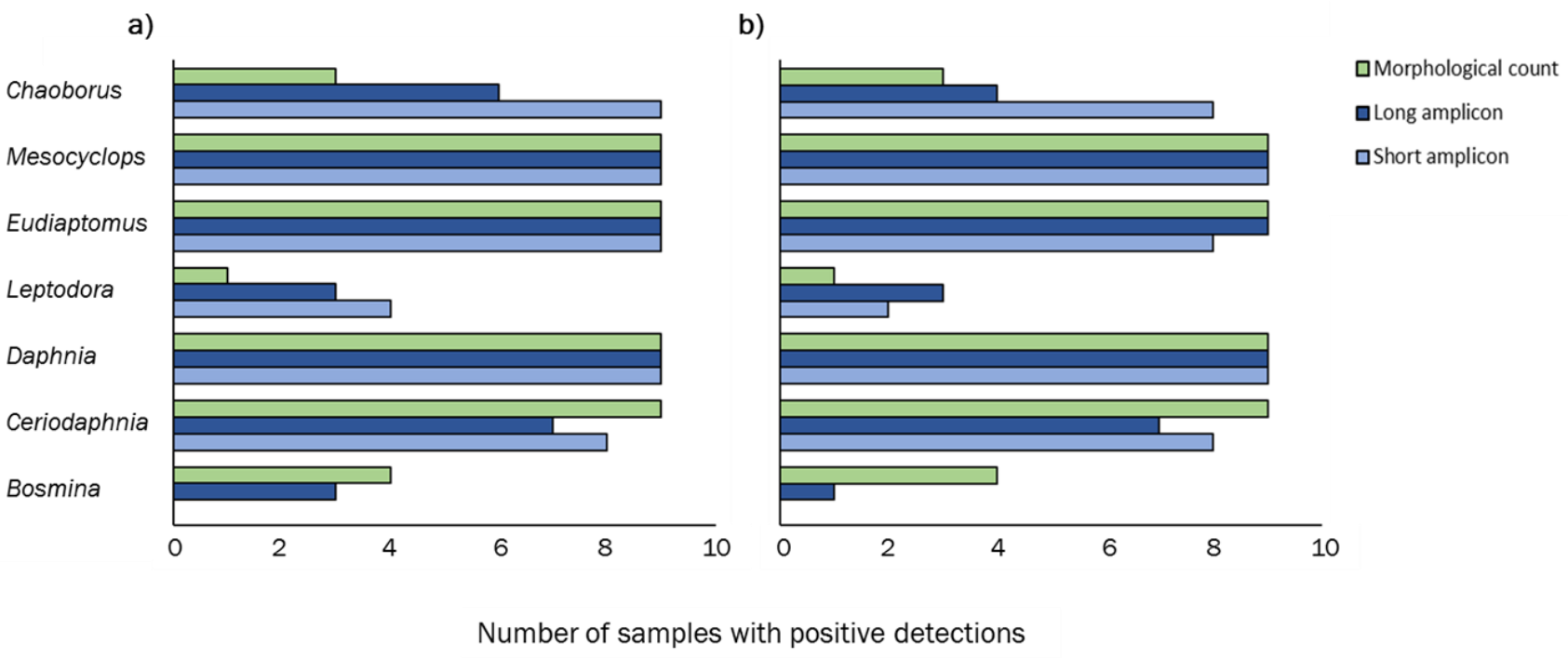

Figure 4.6 *The number of samples that show a positive detection for each genus from morphological counts, long amplicon reads, and short amplicon reads using unfiltered read data (a) and read data filtered to remove reads less than 0.05% of arthropod reads per sample (b).*

4.4.4 Validation: Relative Abundance

4.4.4.1 *Percentage read abundance*

Using percentage target (arthropod) read abundance, the two amplicons showed different patterns of relative abundance in the samples (Figure 4.7). The short amplicon reads were dominated by *Chaoborus* and *Mesocyclops* and showed higher percentages of *Ceriodaphnia* than the long amplicon reads. In contrast, the long amplicon reads show higher percentages for *Daphnia* than the short amplicon reads. The higher percentages of *Chaoborus* reads in the short amplicon data result in one sample (August: 8-12 m) showing very low percentages of any other taxa. Furthermore, reads for *Eudiaptomus* fall below the 0.05% filter threshold affecting the presence/absence of the taxon if filters are used. The differences in percentage read abundance that occur with amplicon choice can affect apparent species dominance e.g. in the September: 4-8 m sample, *Daphnia* is the most abundant taxon according to the long amplicon reads but *Mesocylops* is the most abundant according to the short amplicon reads.

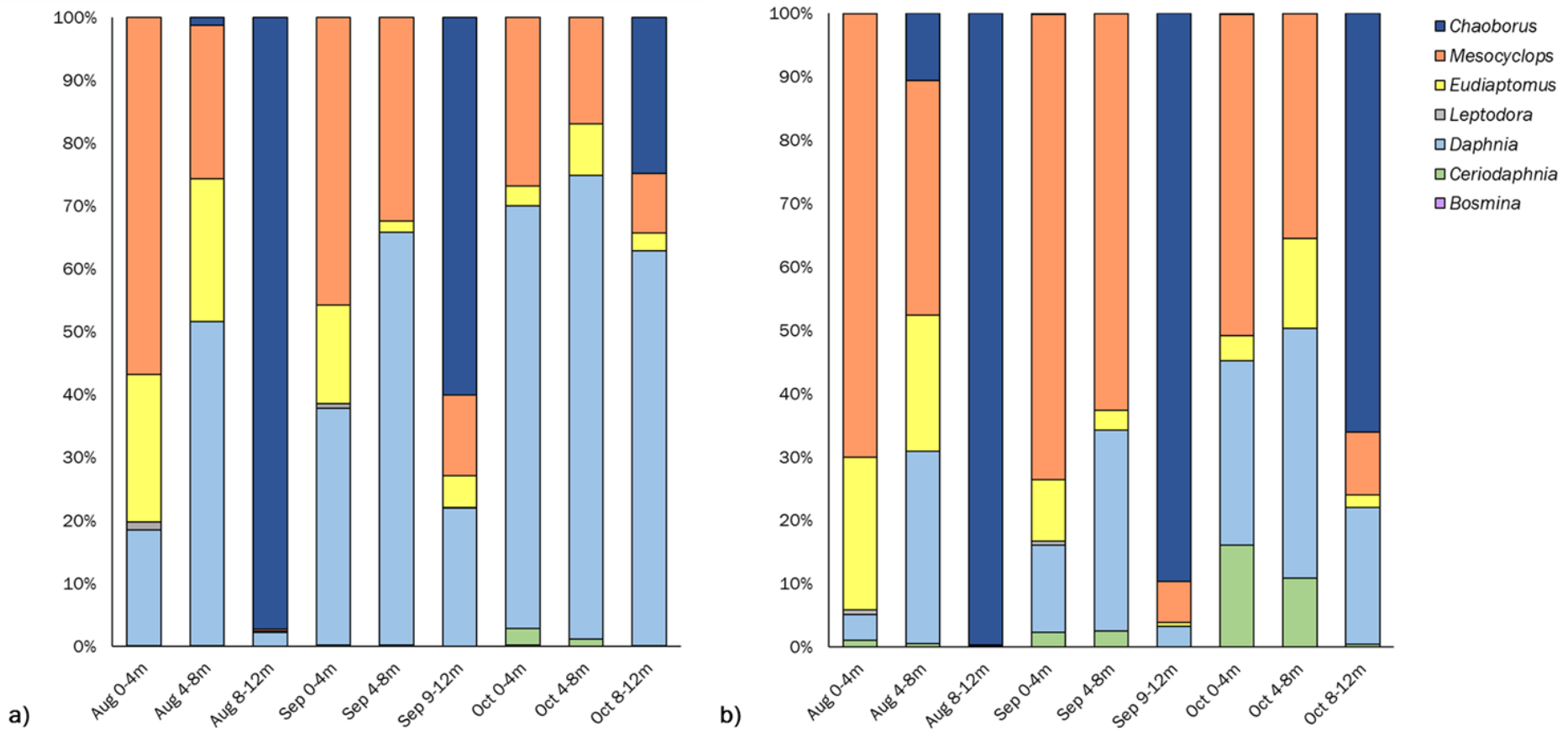

Figure 4.7 *Percentage read abundance for the seven target genera from the long amplicon (a) and the short amplicon (b).*

4.4.4.2 *Primer bias*

Comparison of the percentage read abundance from both amplicons shows the different amplicons show bias towards different taxa (Figure 4.8). In comparison, the long amplicon shows bias towards *Daphnia* and the short amplicon shows bias towards *Chaoborus*, *Mesocyclops* and *Ceriodaphnia*. In contrast, percentage reads for *Eudiaptomus* do not show a consistent bias from either amplicon.

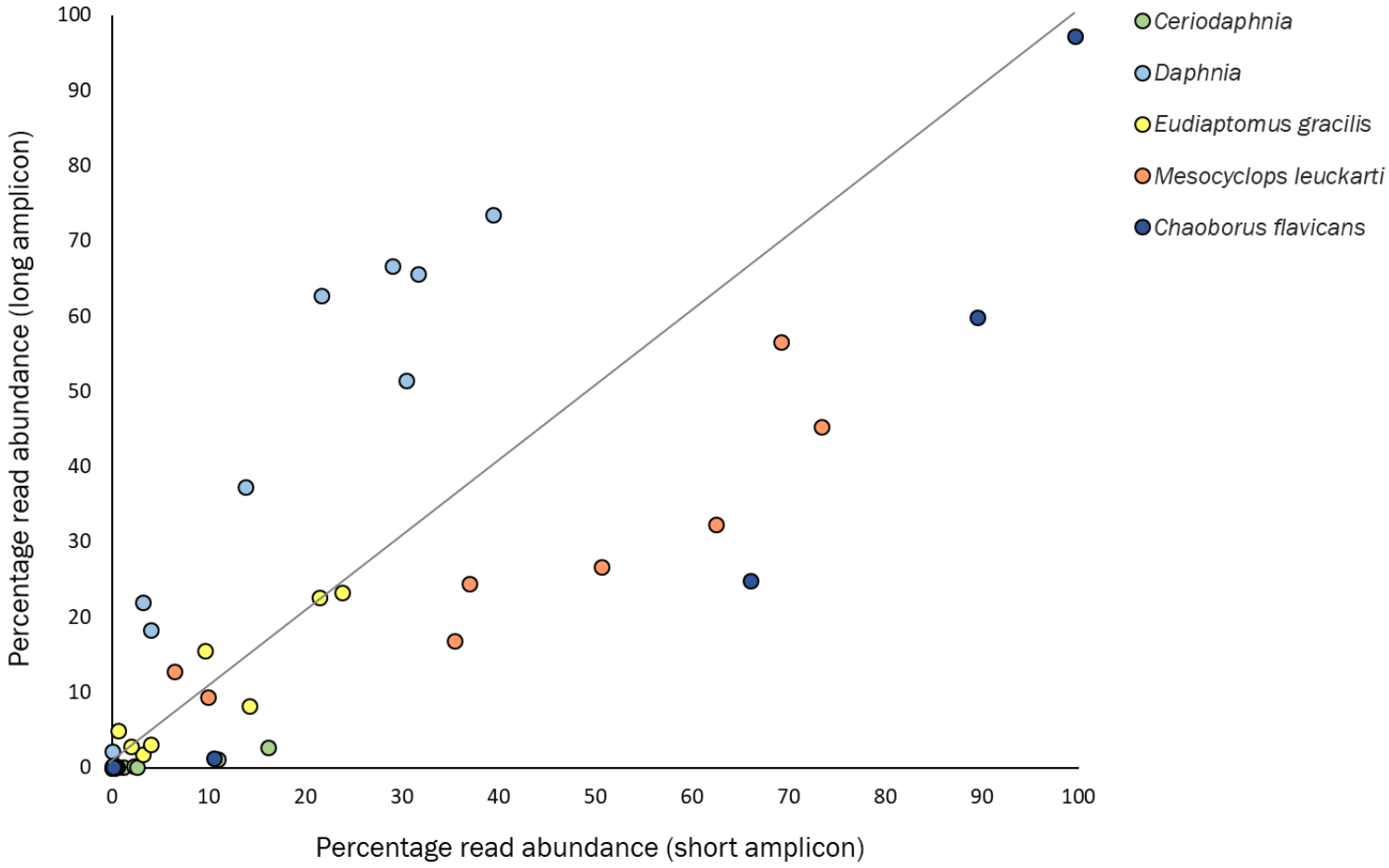

Figure 4.8 *Percentage long amplicon reads against percentage short amplicon reads for all samples for each genus.*

4.4.4.3 *Percentage arthropod read abundance and number of individuals*

Overall comparison between percentage arthropod read abundance and microscopically-derived numbers of individuals showed a positive correlation for both the long amplicon (Spearman = 0.64, $p < 0.001$, $n = 45$) and the short amplicon (Spearman = 0.57, $p < 0.001$, $n = 45$) (Figure 4.9).

These comparisons showed variation among samples (Figure 4.9 a and c) and within samples (Figure 4.9 b and d). However, the sample size is too small to evaluate whether there are significant correlations either among or within samples (n = 9 and 5 respectively). The comparisons showed that the deep water samples (8-12 m) show poor relationships using either amplicon. These are the samples that are dominated by *Chaoborus* reads (Figure 4.7). Both amplicons are strongly biased towards *Chaoborus*, with single *Chaoborus* individuals resulting in the majority of reads within a sample.

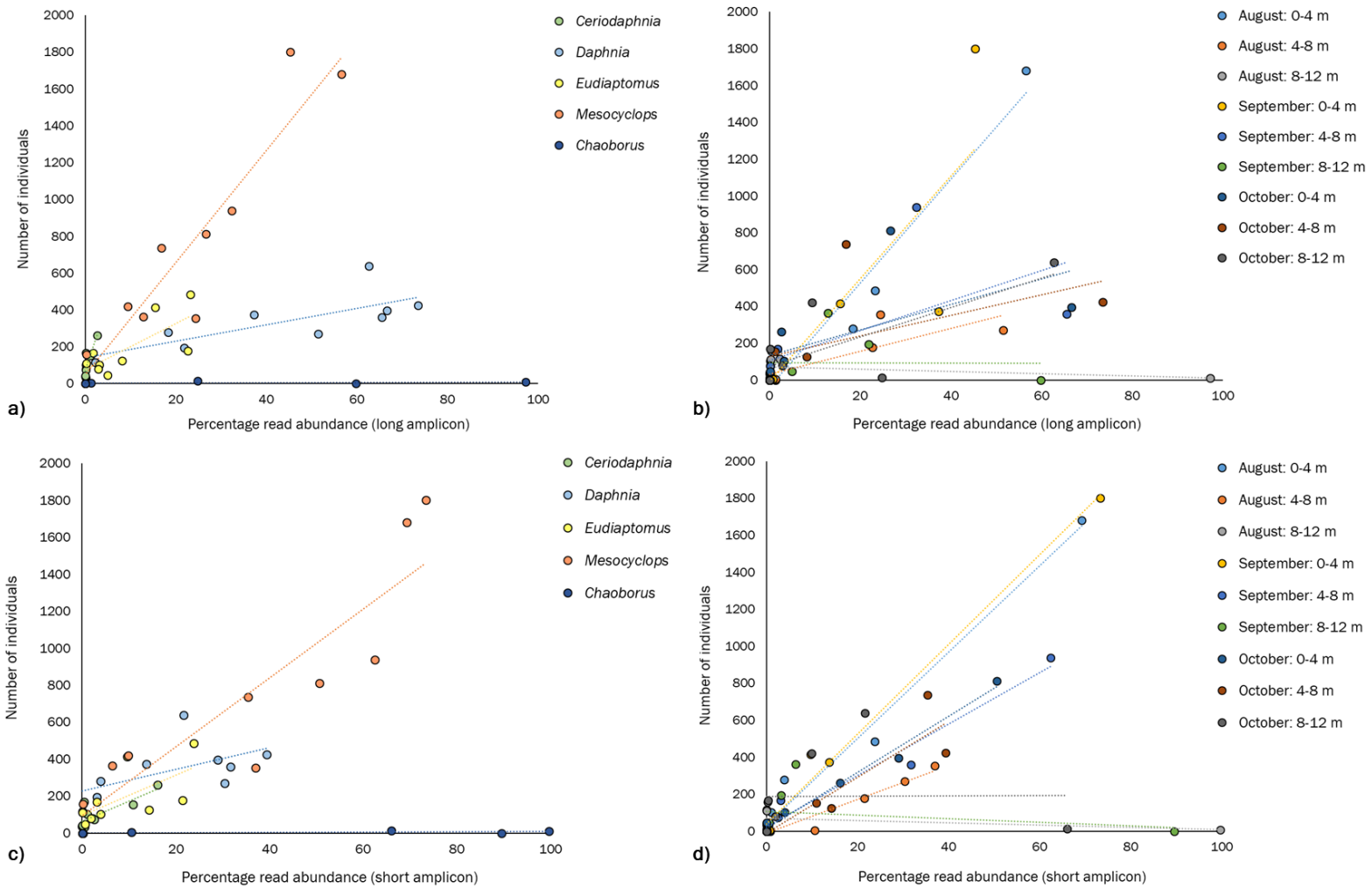

Figure 4.9 *Percentage arthropod read abundance against number of individuals for the long amplicon: among samples (a) and within samples (b); and the short amplicon: among samples (c) and within samples (d).*

4.4.4.4 *Percentage crustacean read abundance and number of individuals*

The dominance of sample reads when very few *Chaoborus* are present, (caused by primer bias and/or the size difference between *Chaoborus* and the other taxa) causes other taxa in the sample to be under-represented in percentage sample reads (and potentially go undetected). Recalculating the percentage read abundance using only taxa in the sub-phylum Crustacea as the target taxa might provide a more representative relative abundance of these taxa, excluding *Chaoborus* (sub-phylum: Hexapoda). This cannot account for taxa that were not detected at all due to the bias towards *Chaoborus* (e.g. *Ceriodaphnia* and *Eudiaptomus* in the filtered dataset only) but could provide a better measure of relative abundance of crustacean taxa within samples.

Overall comparison between percentage crustacean read abundance and numbers of individuals showed stronger positive correlations for both the long amplicon (Spearman = 0.73, $p < 0.001$, n = 36) and the short amplicon (Spearman = 0.79, $p < 0.001$, n = 36) (Figure 4.10). Although the sample size is too small to evaluate whether there are significant correlations either among or within samples (n = 9 and 4 respectively), the comparison using the short amplicon suggests that these data might provide some useful indications of relative abundance within samples (Figure 4.10 d).

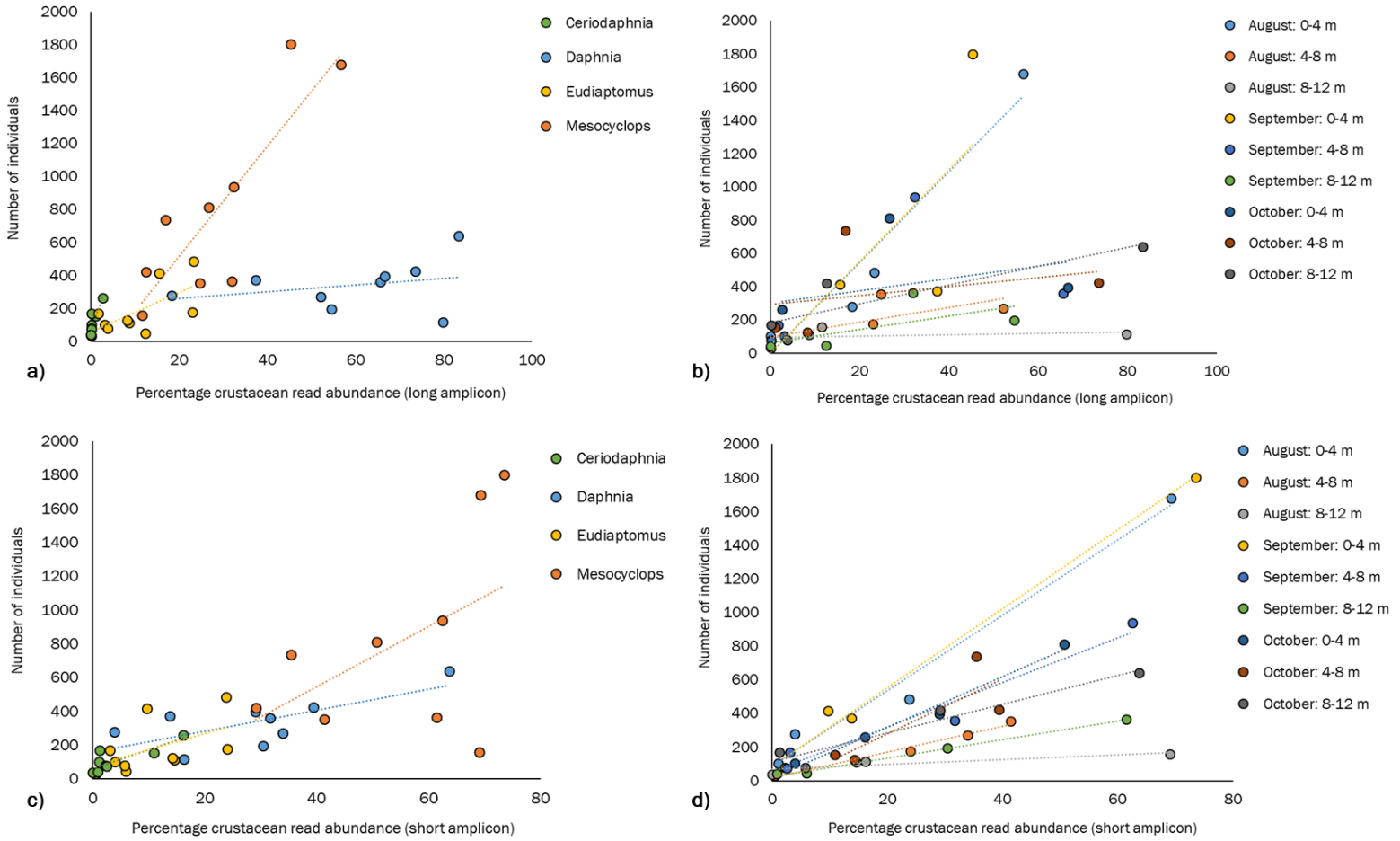

Figure 4.10 *Percentage crustacean read abundance against number of individuals for the long amplicon: among samples (a) and within samples (b); and the short amplicon: among samples (c) and within samples (d).*

The relationship between the number of individuals and the percentage read abundance was better using the short amplicon reads so these data are focused on for relative abundance.

The bias towards *Chaoborus* distorted relative taxon abundance in comparison with the number of individuals in the sample when the percentage arthropod read abundance was used (Figure 4.11).

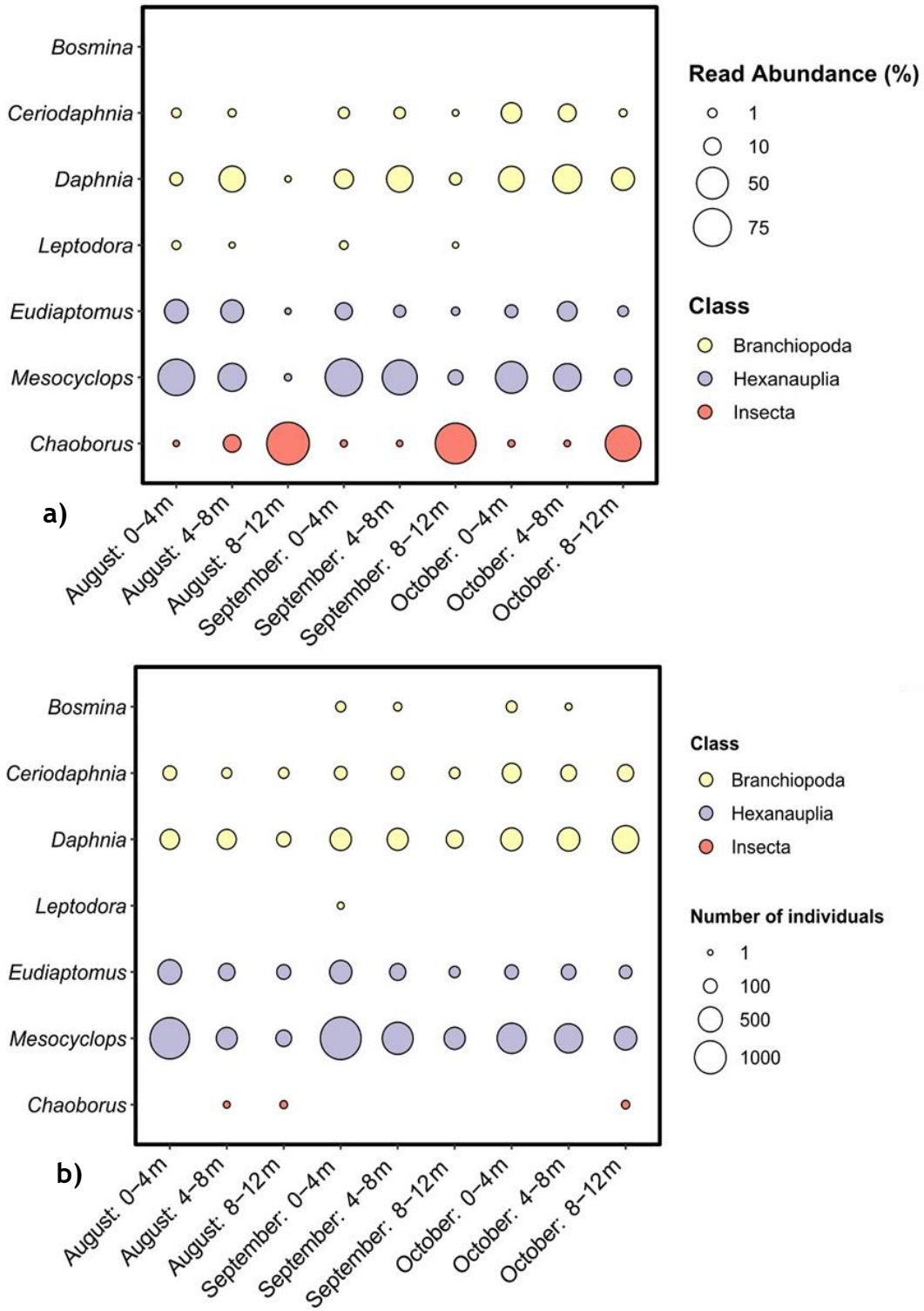

Figure 4.10 *Presence and abundance of target genera per sample according to short amplicon read abundance (percentage of arthropod taxa) (a) and morphological counts (b).*

The percentage crustacean short amplicon read abundance provided a better representation in comparison with the number of individuals from counts (Figure 4.12) enabling read abundance data to be used to understand more about the relative abundance of taxa within samples.

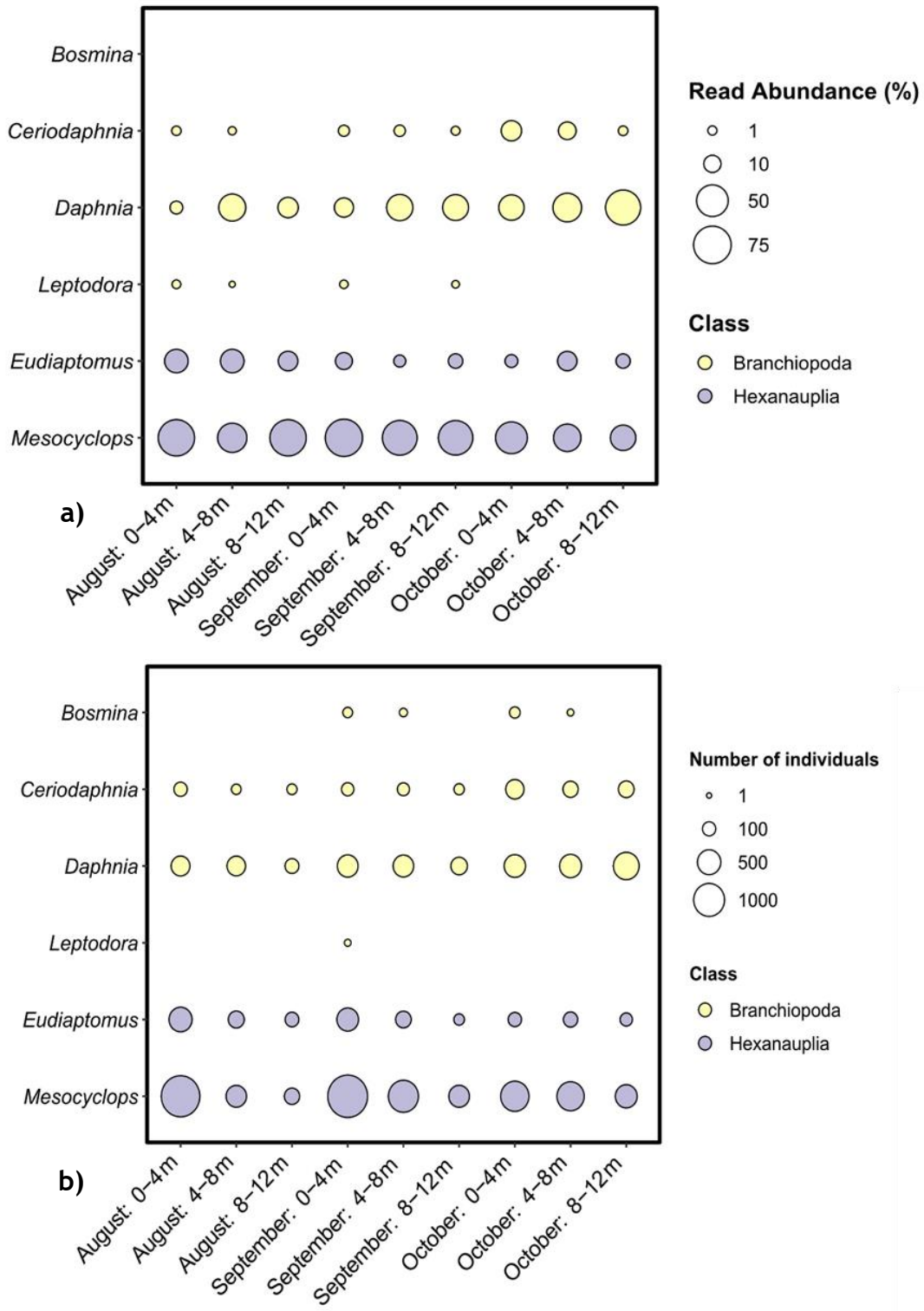

Figure 4.11 *Presence and abundance of target genera per sample according to short amplicon read abundance (percentage of crustacean taxa) (a) and morphological counts (b).*

4.4.5 Validation: Community composition

Comparison of read abundance and morphological count data enables validation of metabarcoding data and reveals potential biases. Metabarcoding also provides other data that might not be gained from morphological counts. Metabarcoding detected other arthropod taxa that were not the dominant target taxa of the pelagic samples. The other arthropods that were detected showed some variation between the two amplicons. *Chydorus brevilabris* was detected in two samples with the long amplicon but only one with the short amplicon. *Macrocyclops albidus* was detected in only one sample with the long amplicon but in five with the short amplicon (Figure 4.13). In addition, most of the sequences could be assigned to species-level (compared to genus-level morphological data). All but one taxon (*Ceriodaphnia*) could be identified to species using the long amplicon data and all but two taxa (*Ceriodaphnia* and *Daphnia*) using the short amplicon data (Figure 4.13). These data can show differences in community composition at different depths and across time at high taxonomic resolution and are able to detect rarer taxa.

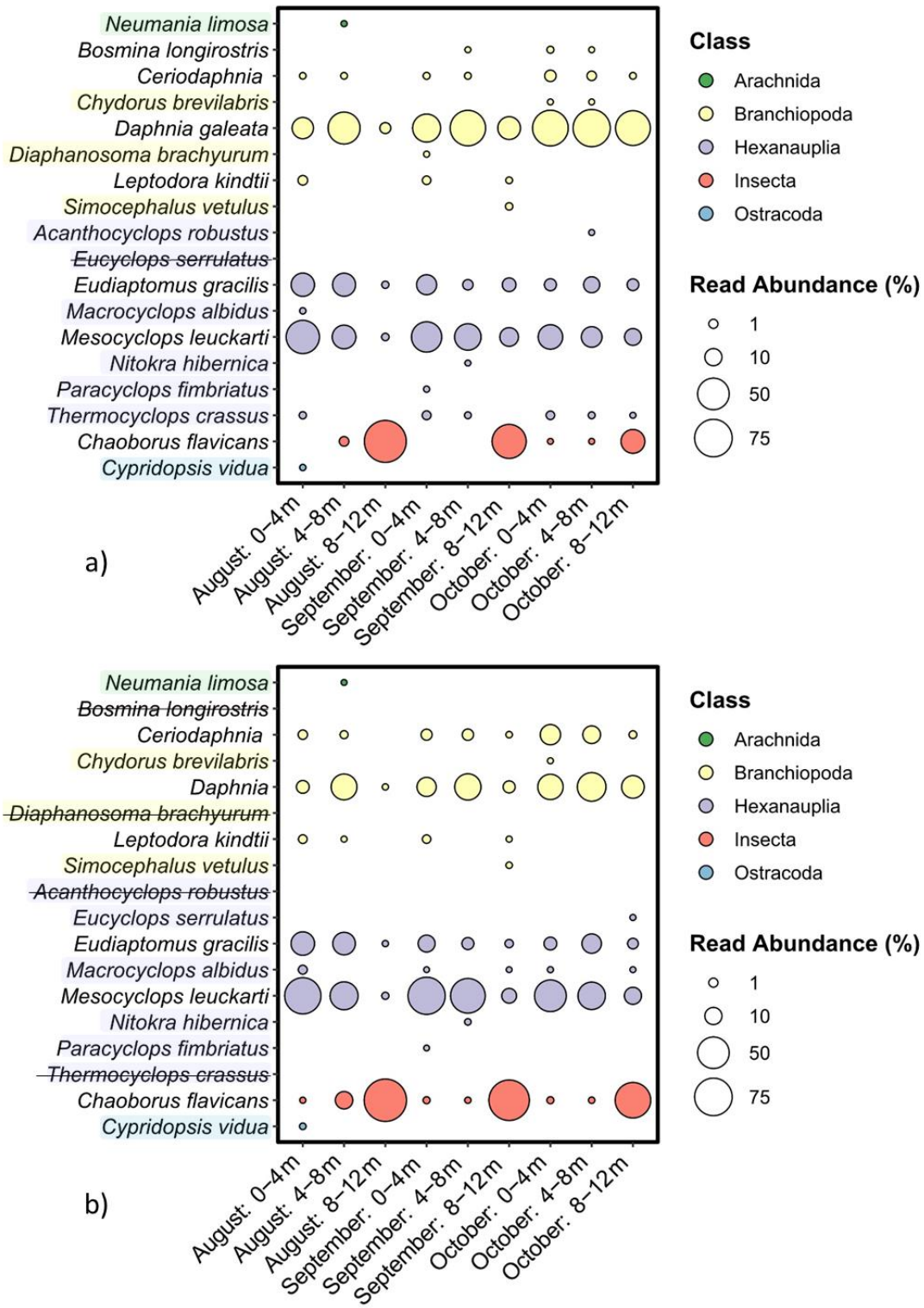

Figure 4.12 *Presence and abundance of target genera per sample according to read abundance (percentage of arthropod taxa) for the long amplicon (a) and the short amplicon (b). Arthropod taxa detected in addition to the dominant target taxa are highlighted. Taxa detected by only one amplicon are crossed out where not detected.*

4.5 Discussion

When specific taxa are of interest, careful marker choice and thorough optimisation are essential so that false positive and false negative detections can be minimised, and data limitations can be understood (Elbrecht and Leese 2017; Cristescu and Hebert 2018). The best choice of marker for a study depends on the identities of the target taxa and the required taxonomic resolution. While many studies recommend multiple markers to improve amplification success and taxonomic resolution (Pompanon et al. 2012; Taberlet et al. 2012), this increases costs. This study aimed to optimise and validate a metabarcoding approach for bulk samples of pelagic crustacean zooplankton taxa from lakes in the Lake District, UK. The diversity of the target taxa in this study was very low in comparison to marine zooplankton samples (Fernando 1994). This meant that optimisation of a single, high-resolution marker could potentially provide successful amplification and species-level resolution of all the target taxa without the need for multiple markers. In order to ensure that the target taxa could be amplified, and assess the strengths and limitations of metabarcoding for zooplankton community assessment, this study aimed to optimise primers and validate the resulting metabarcoding data in comparison with morphological taxonomic data.

4.5.1 Optimisation

Many metabarcoding studies have focused on the importance of marker choice (e.g. Deagle et al. 2014; Elbrecht et al. 2016; Clarke et al. 2017; Elbrecht and Leese 2017; Alberdi et al. 2018; Zhang et al. 2018) as there are no truly universal markers that will amplify all taxa in a sample and provide the genetic variation necessary for species-level taxonomic resolution. High interspecific variation within the mitochondrial COI gene and comprehensive reference databases (Ratnasingham and Hebert 2013) provide the potential for species-level resolution and identification of taxa. However, this high variation among species also causes a lack of conserved primer binding regions, making it challenging to design primers that will successfully amplify many taxa, especially if the taxa are not closely related (Deagle et al. 2014). The use of degenerate primers has been proposed as a potential solution to this challenge

but careful design and thorough optimisation are necessary (Elbrecht and Leese 2017).

The design of primers for different taxonomic groups has led to a proliferation of primers in the literature that have been successfully used for metabarcoding. These published primers are often used in subsequent studies without optimisation for different target taxa (Elbrecht and Leese 2017). In this study, COI primers from the literature that have been used to successfully metabarcode freshwater invertebrates were used as the starting point for optimising primers for the target freshwater zooplankton taxa. Initial *in silico* tests of all the primers provided useful data on the numbers of target orders the primers were likely to bind to and numbers of mismatches with those sequences. This information enabled the range of potential primers to be narrowed down and enabled the chosen primers to be modified to further increase the chances of successful amplification of the target taxa. Thorough PCR testing of a large number of primers is not practical, so although some variation between *in silico* tests and PCR amplification is expected (Alberdi et al. 2018), it is an effective starting point for the selection and modification of primers (Clarke et al. 2014).

PCR tests of selected primers showed variation in amplification success for the target taxa. The variation in the number of target taxa amplified by the primer pairs was surprising given that the nine pairs consisted of just six different primers and that two of the forward primers (BF1 and fwhF2) and two of the reverse primers (BR2 and Fol-degen-rev) targeted the same primer binding regions but varied in primer length and degeneracy (Figure 4.3 and Table 4.3). This variation in amplification success on a small number of target taxa demonstrates the importance of PCR testing with the target taxa and the effects that even very small changes in primer design can have on amplification success.

In addition to the number of target taxa amplified by the different primer pairs, PCR tests also suggested that some of the primer pairs might be susceptible to primer slippage i.e. primers binding a few base pairs upstream or downstream of the designed binding site due to homopolymer regions in the flanking region

on the target template (adjacent to the 3' end of the primer) (Elbrecht et al. 2018). Four of the primer pairs tested by PCR showed slight variation in the length of the amplicons for particular taxa. Primer pairs 6 and 7 amplified slightly longer amplicons than expected, while pairs 8 and 9 amplified slightly shorter amplicons than expected. This taxon-specific variation in amplicon length might be caused by primer slippage. One of the primers (BF1) was found to be susceptible to slippage with some taxa in the study that identified this issue (Elbrecht et al. 2018). Further assessment of the target template sequences in this study would help to understand the cause of the amplicon length variation seen here. Although primer slippage is more likely with degenerate primers, steps to reduce it during primer design/modification should be taken so as not to cause an artificial inflation of sequence diversity in metabarcoding datasets. The four primer pairs producing variation in amplicon length for some target taxa were therefore less suitable for metabarcoding zooplankton community samples in this study.

Although primer pairs 4 and 5 were assessed to be equally successful in amplifying all the target taxa, these primers were likely to yield very similar metabarcoding results since the only difference was in the length and degeneracy of the reverse primer. Only one of these primer pairs was therefore chosen for use in zooplankton community metabarcoding (primer pair 5). The use of shorter amplicons may be advisable in studies where DNA is likely to be more degraded (Clare 2014) so primer pair 1 (targeting 205 bp) was also chosen for use in community metabarcoding even though it failed to amplify one of the target taxa (*Bosmina*).

4.5.2 Validation

Metabarcoding data showed that primers targeting both the long (313 bp) and short (205 bp) amplicons amplified mostly only arthropod taxa with some amplification of rotifers. Although optimisation for this study was focused on the pelagic arthropods, primers were initially also assessed *in silico* using sequences from two orders of rotifers (Bdelloidea and Monogononta) as rotifers are abundant in pelagic lake habitats but not always included in traditional zooplankton counts. In this study, the plankton net and filtering mesh sizes

would not retain most rotifers and so it was expected that rotifers would be limited within these samples. Amplification of some taxa within the phylum Rotifera suggests these primers might be useful for metabarcoding of rotifers but further validation would be necessary in order to assess this.

The primer pairs for both the long and short amplicons successfully amplified six of the target taxa from zooplankton community samples and the amplified sequences could be assigned to at least genus-level so both amplicons could be useful in metabarcoding zooplankton community samples from Lake District locations. As predicted by single-taxon PCR tests, only the long amplicon amplified *Bosmina* from the community samples so the short amplicon would not be suitable if *Bosmina* detection is required. Both amplicons also detected other taxa that are not usually abundant in pelagic habitats but can be found in low numbers in these samples. The majority of these taxa were amplified by both primer pairs but four taxa were only amplified by either the primer pair for the long or short amplicons. Although the primers used in this study were designed to amplify freshwater invertebrates, and chosen and modified for use with zooplankton, there is variation in which taxa they detect in community samples. Prior knowledge of which taxa can be amplified by a primer pair is essential when particular taxa are of interest and is important in all studies to better understand false negatives in detections.

The detection of target taxa in community samples was more consistent between the morphological data and the long amplicon metabarcoding data than with the short amplicon data. Although false negatives for the smallest taxa (*Ceriodaphnia* and *Bosmina*) was a problem with both amplicons, they occurred less frequently using the long amplicon. However, this difference in agreement was mainly caused by the lack of detection of *Bosmina* by the short amplicon, which was predicted by single-taxon PCR tests. False negatives for *Ceriodaphnia* were less frequent using the short amplicon. False positives for the larger taxa (*Chaoborus* and *Leptodora*) occurred less frequently using the long amplicon. The short amplicon might be more sensitive to amplifying very small amounts of DNA or degraded DNA present in the samples. This higher sensitivity could cause both the reduction in false negatives of *Ceriodaphnia* (as it is sensitive to the small amounts of DNA present from low numbers of this

small taxon) and also the increase in false positives of larger taxa (where remains of these taxa would provide small amounts of DNA in the sample). For three samples, the very low read abundance for *Chaoborus* could be caused by cross-contamination from the positive control but this cannot account for all the false positives so additional sensitivity with the short amplicon is likely. This additional sensitivity could be useful when very low amounts of DNA are expected (e.g. low abundance taxa or degraded DNA from dietary samples) but is less accurate in assessing the presence/absence in samples containing varying sizes of zooplankton. It should also be noted that the community samples were split in half for comparisons between the number of individuals and read abundance so would not be identical in composition. This is more likely to cause an issue for low abundance taxa where the low numbers of individuals present could end up only in just one half of the sample.

False positives for the larger taxa can be reduced by filtering out reads below a particular threshold of the total read abundance for the sample (e.g. 0.01-0.05%). However, this filtering of low read abundance from the datasets also increased the false negatives for *Bosmina* (long amplicon) and *Eudiaptomus* (short amplicon). Filtering low reads in metabarcoding data is often a standard part of bioinformatics pipelines, but whether or not data are filtered, and at what threshold, has important effects on false positives and false negatives and should ideally be done with knowledge from validation of which taxa are most likely to be affected and in what way (Corse et al. 2017; Leray and Knowlton 2017). Read abundance in bulk samples is dependent on the abundance of the taxa and the size of the taxa (Leray and Knowlton 2017) so blanket filtering across a dataset will not increase the accuracy of detections for all taxa in all samples.

If only presence/absence detections are required, the long amplicon provided better agreement with morphological data, was able to detect all the target taxa, and could detect more additional taxa than the short amplicon. However, if gaining an element of relative abundance information from the metabarcoding data is important, the short amplicon data showed better relationships between read abundance and morphological data. The short amplicon data provided some relative abundance data within samples where

higher read abundance for a taxon related to higher numbers of individuals within a sample. However, the sample size for these relationships was very small due to the number of different taxa within each sample so these relationships should be treated with caution and further validation is needed.

High read abundance for *Chaoborus* appeared to mask the relative abundance information that existed in the read abundance for the other taxa. Very low numbers of *Chaoborus* individuals in a sample were represented by very high proportions of the reads (e.g. 9 individuals were represented by 97.2-99.7% of the reads for the August 8-12 m sample (long and short amplicon respectively) (Figure 4.9). This domination of reads might be caused by both primer bias towards *Chaoborus* and the large size of *Chaoborus* in comparison with the prey taxa. As *Chaoborus* is an insect and the prey taxa are crustaceans, primers that target the crustacean taxa and block amplification of *Chaoborus* could provide better abundance information on the prey community and prevent the associated false negative detections caused by domination of reads by a single taxon. Recalculation of percentage read abundance using only crustaceans as the target taxa provided a better relationship between percentage read abundance and the number of individuals within samples (but cannot account for any taxa that were not detected at all due to primer bias towards *Chaoborus*). Variation in how percentage read abundance is calculated (as a percentage of all amplified sequences, all taxa minus those considered contamination, all 'target' taxa etc.) affects the relationships between the percentage read abundance and the abundance of individuals. Where read abundance data can be validated with numbers of individuals in samples, it might be possible to adjust the analysis of metabarcoding data to improve the relative abundance information provided by the metabarcoding data. In contrast, if this validation is not done, bioinformatics and data analysis decisions cannot be optimised for the particular primers, taxa and sample type and confidence in the resulting data will therefore be lower.

It should be noted that the findings in this study about both the detection of zooplankton taxa and the relative abundance of those taxa are dependent on the specific mix of sequences in the community samples. In zooplankton samples with higher diversity/different taxa, different biases might be seen

that change the likelihood of detecting particular species and the relative read abundances. It is therefore very difficult for findings about primer optimisation/validation to be generalised to samples from other habitats even if the communities are relatively similar in taxonomic composition. This highlights the need for preliminary studies when metabarcoding is to be used to draw ecological conclusions.

4.5.3 Conclusions

Thorough optimisation and validation of methods for metabarcoding freshwater zooplankton samples is essential if DNA-based approaches are to provide an effective method to enable quick and reliable assessments of zooplankton communities for monitoring purposes. The methods used in this study provide very accurate and sensitive data on presence of zooplankton taxa and were able to detect low abundance taxa that could easily be overlooked in traditional morphological counts, where methods often involve counting sub-samples of the total sample. False positives in this study were for larger taxa and were likely caused by remains of the taxa caught up in samples. These false positives can be reduced by filtering low abundance reads from the dataset but this should be done carefully as it is also likely to cause some false negatives for smaller/lower abundance taxa. Where the size differences between taxa are known to be large, filtering could be used for just the larger taxa rather than across the whole dataset to reduce false positives and negatives and provide more accurate presence data.

Using metabarcoding data to understand the absolute abundance of taxa is currently not accurate enough due to biases throughout the process (Elbrecht and Leese 2015; Pinol et al. 2015; Luo et al. 2022). However, metabarcoding data can provide some relative abundance information for bulk samples. Validation of read abundance data against microscopically-determined numbers of individuals in samples can help to optimise the bioinformatics and data analysis for relative abundance and understand the limitations of the data. The bioinformatics and data analysis for the zooplankton data in this study could be further optimised to provide more accurate relative abundance data by applying correction factors to account for the relative sizes of the taxa and

apparent primer biases. The bias towards *Chaoborus* in this study is strong enough to prevent the amplification of some taxa in the sample causing false negatives. For these zooplankton samples, it would be beneficial to explore whether this is due to primer bias and if so design/modify primers to be less biased towards *Chaoborus*. However, differences in providing relative taxon abundance were seen between the primer pairs in this study so designing primers that are less biased and also provide relative taxon abundance data may be challenging. An alternative method of preventing the bias in read abundance towards *Chaoborus* would be to remove *Chaoborus* individuals from samples prior to DNA extraction. The large size of late instar larvae makes this a tractable alternative.

To understand interactions among zooplankton, and especially how these interactions change with prey composition, more accurate relative abundance data (within samples) from metabarcoding would be informative. For metabarcoding to be used for monitoring zooplankton communities, data on changes in taxa abundance over time and space (among samples) are needed. In this study, relationships between read abundance and variations in the absolute number of individuals among samples are poor, but could potentially be improved using correction factors for relative sizes of taxa and primer bias. Deriving abundance data from metabarcoding is a very active area of research currently (Luo et al. 2022) and further developments in this area are likely to enable metabarcoding of bulk zooplankton samples to become a quick and accurate alternative to morphological counts for monitoring of zooplankton communities in the future.

An important benefit afforded by metabarcoding is that it was possible to identify most taxa to species-level. Morphological identification to species-level is extremely time-consuming and depends on specialised taxonomic expertise, where taxonomy as a field is in decline (Hopkins and Freckleton 2002) and morphological identification is likely to become more challenging over time. Morphological counts are therefore often done to genus-level, which is likely to miss differences in the true biodiversity among samples, making samples from different depths, sites and times look more homogenous than they are in reality. Capturing these differences in biodiversity over time and space

is critical to monitoring change, particularly in low diversity systems such as freshwater zooplankton. Metabarcoding alongside morphological counts to genus-level would enhance the capacity of current zooplankton monitoring even before improvements in DNA-based abundance data are achieved.

This study highlights the importance of optimisation and validation of methods for metabarcoding when ecological inferences are to be made from the metabarcoding data. Although the use of published primers and standard bioinformatic packages can be used to generate metabarcoding data, results from these assessments must be treated with caution as any primer biases, false positives and false negatives are likely to be unknown if the methods have not been optimised for the specific target taxa and sample types. In this study, even very small changes in primer design caused differences in which taxa amplified successfully from single-taxon template DNA. In addition, in community samples of mixed taxa, biases towards some taxa prevented the amplification of taxa that were present in the sample and could be successfully amplified by the primers. Optimisation of primers, bioinformatic processing and data analysis for the target taxa enabled false positives and negatives to be understood and reduced to provide a more accurate assessment of the zooplankton community. This understanding of the strengths and limitations of metabarcoding data is essential to ensure the data are used appropriately to enhance monitoring and assessment of freshwater biodiversity.

5 DNA-based analysis of the diet of phantom midge, *Chaoborus flavicans*, larvae in a lake ecosystem: combining community metabarcoding and dietary screening to analyse interaction strengths.

5.1 Summary

Monitoring of the ecological interactions between species could provide a more sensitive method of monitoring changes in ecosystems. DNA-based identification provides the opportunity to resolve interactions between organisms in communities and monitor how these interactions change over time and in response to environmental change. In pelagic freshwater habitats, zooplankton occupy a key central position. Phantom midge (*Chaoborus flavicans*) larvae are voracious predators of zooplankton that can become very abundant in mesotrophic-eutrophic lakes and can have a significant impact on zooplankton communities.

This study aimed to demonstrate the potential for combined community metabarcoding and individual gut content screening methods for identifying dynamics in small body-size zooplankton predator-prey interactions. Using the bulk metabarcoding methods optimised in Chapter 4, the spatio-temporal dynamics in the potential prey community were characterised. Specific assays to detect prey taxa in the gut contents of *Chaoborus* were then developed and used to screen *Chaoborus* individuals for the prey taxa. Interaction strengths between *Chaoborus* and the prey taxa were then analysed.

The optimised methods for metabarcoding bulk zooplankton samples provided sequences for assay design and optimisation, data on the behaviour and habitat use of *Chaoborus in Bleham Tarn,* and enabled the potential prey taxa available to *Chaoborus* individuals in the selected samples to be identified. Optimised specific assays enabled the gut contents of *Chaoborus* individuals to be screened for the target prey taxa, *Daphnia* and *Bosmina*. Individual-level dietary data enabled a measure of interaction strength to be compared across samples.

Optimisation of metabarcoding methods provided essential data on the zooplankton community composition and the prey availability for *Chaoborus* in Blelham Tarn and enabled conclusions to be drawn about the behaviour and habitat use of *Chaoborus*. Assays, specific to the target prey taxa, provided an efficient and cost-effective method of obtaining individual-level interaction data across multiple samples. The individual-level data showed that diet varied between sizes of *Chaoborus* individuals. Understanding this ontogenetic shift in diet enabled interaction strengths to be compared over time and space, providing a powerful, quantitative method for monitoring community changes that are likely to precede species turnover and loss.

5.2 Introduction

5.2.1 Importance of resolving interactions in lake ecosystems

Freshwater biodiversity is widely considered to be in crisis (Reid et al. 2019; Tickner et al. 2020) and the pressures on freshwaters are predicted to increase in the future as human consumption continues to increase alongside rapid environmental change (Darwall et al. 2018). To monitor changes in and impacts on biodiversity, we need methods that are sensitive to the early changes in communities that precede changes in species presence. Understanding anthropogenic impact requires not only the knowledge of which species are present or absent, but also an understanding of the ecological processes that occur within ecosystems and how changes in these processes affect ecosystem functions and services (McCann 2007). A very important aspect of biodiversity that could enable more sensitive monitoring of changes in ecosystems is the ecological interactions between species (Tylianakis et al. 2008; Valiente-Banuet et al. 2015). Interactions between organisms underpin ecosystem functioning and stability (Stouffer 2010; Thompson et al. 2012a; Staudinger et al. 2021) and are impacted by changes in the environment.

Studies in lake ecosystems have shown how environmental change impacts interactions between organisms. Changes in lake temperature caused by climate change cause cold water adapted fish to shift their habitat use and foraging patterns within the lake, resulting in changes in their feeding

interactions (Bartley et al. 2019). Long-term monitoring of Windermere (UK) has shown how environmental change (eutrophication and warming) can cascade through the food web from the phytoplankton and zooplankton to the fish populations at higher trophic levels (Staudinger et al. 2021). Differential phenological shifts in phytoplankton, *Daphnia*, and perch (*Perca fluviatilis*) led to trophic mismatch between perch and *Daphnia*, which could affect fish survival and impact ecosystem functioning (Thackeray et al. 2013; Ohlberger et al. 2014).

Temporal and spatial variation in environmental conditions and resource availability within a habitat cause changes in the behaviour of individuals (Beckerman et al. 2010). Adaptive changes in habitat use and foraging behaviour of individuals result in changes in the presence and strength of trophic interactions among species (Berlow et al. 2004; McMeans et al. 2016). Monitoring changes in interaction strengths could therefore provide a more sensitive method of monitoring changes in ecosystems and provide early warning signals of problems before they result in the loss of species and ecosystem functions (Bartley et al. 2019).

5.2.2 Challenges of resolving planktonic food webs

Accurately resolving predator-prey interactions can be challenging (Clare et al. 2009). Resolution of freshwater trophic interactions traditionally involves evidence from morphological analysis of stomach or faecal contents and/or experimental feeding trials (Woodward et al. 2010). Morphological identification of partially digested prey individuals is very time consuming, relies on expert knowledge of prey taxonomy and morphological diversity, and can be biased towards prey species with hard body-parts that are more difficult to digest (Thompson et al. 2012b). Furthermore, once a trophic interaction has been resolved in a study once, it is often assumed that the species will interact when they co-occur but interactions are not static and these snapshots of interactions cannot resolve how interactions change in space and time and in response to environmental change. These methodological challenges have therefore limited the resolution of freshwater trophic interactions detectable in the past.

5.2.3 Opportunities using DNA-based methods

DNA-based identification provides the opportunity to resolve interactions between organisms in communities and monitor how these interactions change over time and in response to environmental change. This represents a major advance on the well-used approach of inferring interactions from published studies and databases, implicitly assuming that those interactions do not change over time or with environmental conditions.

One measure of interaction strength between species is the frequency of occurrence of an interaction or the frequency of consumption (Berlow et al. 2004). DNA-based identification provides a very sensitive method for detecting the presence of prey DNA in dietary samples (e.g. stomach contents, regurgitates or faecal samples) (Symondson 2002; King et al. 2008; Pompanon et al. 2012). The number of individuals that have consumed a particular prey species can provide a measure of interaction strength, making monitoring in this way quantitative without the challenge of obtaining accurate abundance data from DNA-based methods. High quality data for interaction strengths is dependent on accurate prey detections for multiple consumer individuals making accuracy and cost efficiency important considerations in method choice. Screening dietary samples using diagnostic PCR can provide more consistent results (Rennstam Rubbmark et al. 2019) and be more cost-effective than metabarcoding. In addition, it can solve other challenges relating to metabarcoding of dietary samples (e.g. large amounts of predator DNA "swamping" that from prey taxa).

Screening dietary samples using diagnostic PCR requires detailed knowledge of the consumer's potential diet *a priori* so that specific assays can be developed and validated. Metabarcoding of community samples can provide data on what species are present in the community, co-occurrence of predator and prey species, habitat use, and behaviour. In addition, metabarcoding provides sequences of the target species which can be used in development and optimisation of the assays for the target prey. As such, the combination of community metabarcoding and screening techniques within a single study is a

powerful approach to identifying and quantifying predator-prey interactions in a field setting.

5.2.4 Phantom midge (Chaoborus flavicans) larvae

Zooplankton occupy a key central position in pelagic habitats. Taxa that graze on phytoplankton and are an important food source for fish, e.g. *Daphnia*, are often particularly well-studied. Although many zooplankton taxa do graze on phytoplankton, their communities are both functionally and taxonomically diverse, including heterotrophic protists, rotifers, cladoceran and copepod crustaceans and larval insects. The different species have different feeding ecology (herbivory, omnivory and carnivory) and different predator avoidance behaviours, so their "central" position in the food web is complex.

Phantom midge (*Chaoborus flavicans*) larvae are voracious predators of zooplankton and can become very abundant in mesotrophic-eutrophic lakes (Weisser et al. 2018). *Chaoborus* predation on other zooplankton can affect the structure of the zooplankton community (Jäger et al. 2011). *Chaoborus* can therefore be both in competition with and prey for zooplanktivorous fish.

Blelham Tarn is a small meso-eutrophic lake that drains into the north basin of Windermere (UK). *Chaoborus* larvae are found in high abundance in Blelham Tarn (Maberly et al. 2016) and further eutrophication could cause an increase in abundance (Tang et al. 2018). As the impact of *Chaoborus* on zooplankton communities can be significant even at moderate densities (Jäger et al. 2011), there is the potential for *Chaoborus* to play an important role in the Blelham Tarn food web. Understanding *Chaoborus* interactions is therefore ecologically important, and monitoring changes in interactions with *Chaoborus* could provide a sensitive indicator of environmental changes within Blelham. These insights are relevant to other small, productive lakes across the world.

5.2.5 Aims and hypotheses

This study aimed to evaluate the potential for combined community metabarcoding and individual gut content screening methods for identifying dynamics in small body-size zooplankton predator-prey interactions. Using the bulk metabarcoding methods optimised in Chapter 4, it characterises the

spatio-temporal dynamics in potential zooplankton community (potential prey) in Blelham Tarn at three depths every two weeks (July to September 2019 (day and night-time samples)). The metabarcoding data were then used to optimise specific assays to detect prey taxa in the gut contents of *Chaoborus flavicans* individuals. The specific assays were then used to screen *Chaoborus* individuals from three time points for the prey taxa. Interactions strengths between *Chaoborus* and prey taxa were then analysed. The specific hypotheses tested in this study were: optimised and validated metabarcoding of bulk zooplankton samples can provide meaningful data on potential prey communities; and DNA-based screening of individual predator diets can provide a sensitive method for monitoring changes in communities.

5.3 Methods

5.3.1 Main sampling site

Blelham Tarn is a small, shallow (approximately 14 m maximum depth), meso-eutrophic lake in the Lake District (UK) that drains into the north basin of Windermere (Figure 5.1). The fish community mainly consists of Northern pike (*Esox lucius)*, European perch (*Perca fluviatilis*) and roach (*Rutilus rutilus)*. The Tarn is classified as being in a Moderate ecological state based on the Water Framework Directive (WFD) classification and has shown signs of deterioration in water quality (Maberly et al. 2016). Blelham Tarn has been a focus of long-term monitoring since the 1940s by the Freshwater Biological Association (FBA) and UK Centre for Ecology & Hydrology (UKCEH).

5.3.2 Zooplankton samples for barcoding

To provide reference sequences for potential *Chaoborus* prey, zooplankton individuals for Sanger sequencing were collected from four lakes in south Cumbria (UK) between July and November 2019. Samples were collected from the focal site, Blelham Tarn, and from lakes in close proximity to the focal site within the same catchment: Esthwaite Water, Loughrigg Tarn, Windermere south basin and Windermere north basin (Figure 5.1). Zooplankton were collected using vertical hauls with a plankton net (250 µm mesh). Samples were identified morphologically to genus-level and transferred to separate beakers

for each genus. All genera were starved overnight in filtered (0.2 µm filter) lake water at room temperature. Starved individuals were identified to species (where possible) under a high-powered stereomicroscope, rinsed, and frozen at -80°C.

5.3.3 Zooplankton community and predator samples

To provide information on temporal changes in the prey community, zooplankton community samples were collected from Blelham Tarn every two weeks between July and September 2019. Daytime samples were taken at approximately 10:30 am and night-time samples were taken approximately 36 hours later, one hour after sunset on the relevant dates to allow for potential vertical migration of zooplankton.

Zooplankton community samples and predator samples were collected from the deepest point in the lake using a closing net (120 µm mesh). For community samples, three depths were sampled by triplicate vertical hauls: 0-4 m, 4-8 m, 8-12 m. For predator samples, the same three depths were sampled by six vertical hauls to increase the number of individuals collected (as the predators were less abundant than potential prey). In order to minimise predation during transportation, approximately 50 ml of carbonated water was added to each sample (5 ml at a time and swirled gently) and samples were then placed on ice. Additional samples were collected using vertical hauls with a plankton net (250 µm mesh) through the full depth (12 m) to collect additional *Chaoborus* individuals for use in method optimisation.

In the laboratory, the predator samples were screened (visually) and each *Chaoborus* individual was transferred to an Eppendorf tube and frozen at -80°C. Each of the community samples was re-suspended, thoroughly mixed, and split in half by transferring the sample repeatedly between two beakers and then ensuring each beaker contained half the total volume. One half was then stored for morphological identification by filtering the sample on to a 100 µm mesh filter, rinsing the zooplankton with distilled water, and then rinsing them into a universal tube with 70% ethanol. The other half was stored for DNA extraction by filtering it on to a 100 µm mesh filter, rinsing the zooplankton with distilled water, and then filtering the sample on to a 40 µm nylon gauze filter using a

vacuum pump system. The filter was then folded, transferred to a centrifuge tube, and frozen at -80°C.

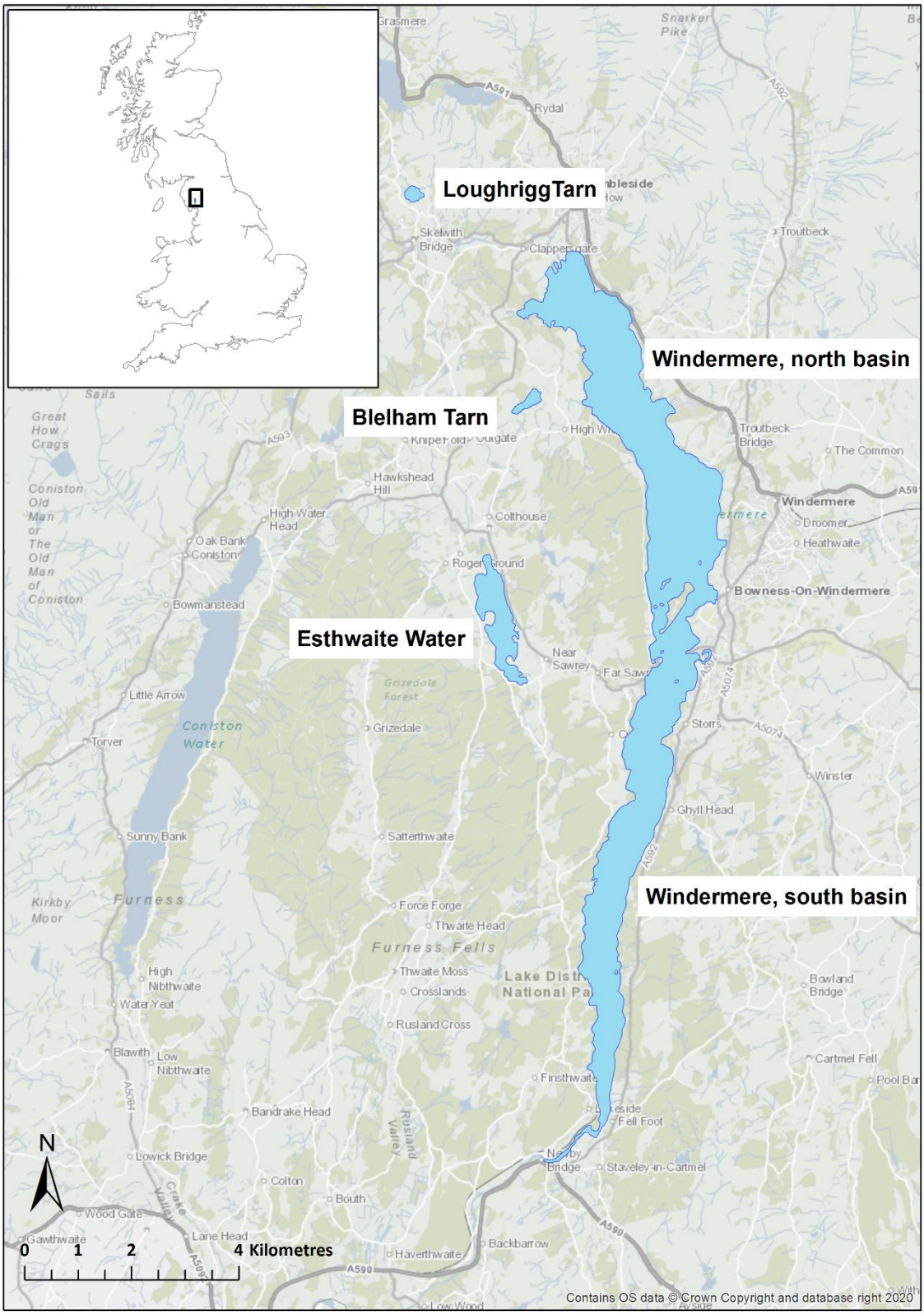

Figure 5.1 *Map showing location of Blelham Tarn (main sampling location) and surrounding lakes in the catchment (locations of sampling for reference sequences).*

5.3.4 Barcoding

To provide reference sequences for potential *Chaoborus* prey, DNA was extracted from zooplankton individuals for Sanger sequencing. Initially, DNA was extracted from single zooplankton individuals using a column-based method (Qiagen DNeasy Blood and Tissue kits) following the manufacturer's protocol. Tests using single and multiple individuals showed only the larger taxa amplified successfully when only single individuals were extracted (PCR success was determined by visualising amplicons on a 0.8% agarose gel). Extraction methods and PCR conditions were therefore tested and optimised in order to maximise DNA yield from very small individual zooplankton taxa. DNA was extracted from zooplankton individuals using a Sigma Extract-N-Amp Tissue PCR kit and a modified protocol (Madhu and Gumienny 2016). Briefly, each zooplankton individual was rinsed three times in ddH_2O and then homogenised in Extract-N-Amp Extraction Mixture using sterile needles (dipped in ethanol and flamed). The mixture was then transferred to a PCR tube and centrifuged for 2-3 seconds (≤6000 rpm). Tubes were then placed in a thermocycler at 55°C for 10 min, then 95 °C for 3 min, centrifuged, and then neutralised using the Extract-N-Amp Neutralisation Solution. Volumes of solutions were reduced for the smallest zooplankton taxa in order to increase the DNA concentration of the final template. Total final extraction volumes ranged from 4.5 to 13.5 µl.

PCRs for zooplankton individuals were run on a Veriti 96-well Thermal Cycler. PCRs were set up with 20 µl of Extract-N-Amp PCR Reaction Mix, 0.5 µM of each primer, 4-8 µl of DNA template (depending on size of zooplankton taxa), and molecular grade water with a total volume of 40 µl. One positive control (using LCO1490 and HCO2198 (standard Folmer primers)) and one negative control (no template DNA) were included. The following thermocycling protocol was used: initial denaturation at 95°C for 10 min, then 40 cycles of: 95°C for 1 min, annealing temperature of 48°C for 45 seconds, then extension at 72°C for 1 min. Followed by a final extension of 72°C for 7 min. PCR success was determined by visualising amplicons on a 0.8% agarose gel. DNA templates were sent for Sanger sequencing (in both directions) to the Molecular Biology Laboratories at the Natural History Museum, London. The following genera were sent for sequencing (number of individuals in brackets): *Chaoborus* (7), *Bosmina*

(10), *Ceriodaphnia* (8), *Chydorus* (1), *Daphnia* (14), *Leptodora* (5), *Cyclops* (4), *Mesocyclops* (7), *Arctodiaptomus* (1), *Eudiaptomus* (7). Numbers of individuals were higher for genera that included multiple species in the target lakes. Only single individuals of *Chydorus* and *Arctodiaptomus* were sent for sequencing because they were not dominant in the pelagic samples.

5.3.5 Community metabarcoding

To provide information on temporal changes in the prey community, zooplankton community samples were metabarcoded. Prior to DNA extraction, community samples were thawed and zooplankton were transferred from the nylon filter into a 1.5 ml tube using a sterile spatula. The sample was then homogenised using a plastic pestle and any material attached to the pestle was returned to the sample using a sterile needle. DNA was extracted from the community sample using MasterPure Complete DNA and RNA Purification kits following the manufacturer's protocol. The concentration of DNA was assessed using a Qubit 3 Fluorometer.

The two primer pairs selected for DNA metabarcoding in the previous chapter (targeting amplicons of 205 bp and 313 bp) were used for metabarcoding Blelham Tarn community samples. The First Step PCRs were run on a Veriti 96-well Thermal Cycler. The PCRs were set up with 2x Amplitaq Gold 360 Master Mix, 0.5 µM of each primer, 2 µl of DNA template, and molecular grade water with a total volume of 25 µl. One positive control, using *Chaoborus* (single taxon) DNA template, and one negative control were used in each PCR. The following thermocycling protocol was used: initial denaturation at 95°C for 10 min, then 35 cycles of: 95°C for 1 min, 49°C for 45 seconds, then extension at 72°C for 1 min. Followed by a final extension of 72°C for 7 min. PCR success was determined by visualising amplicons on a 0.8% agarose gel.

First step PCR product was cleaned up using a ZR-96 DNA Clean-up Kit (Zymo) following the manufacturer's protocol. MiSeq adapters and 8nt dual-indexing barcode sequences were added during a second step of PCR amplification. 1 µl of cleaned DNA was used in the second round PCR. The PCR was set up with 0.25 µl Taq Q5 NEB, 5 µl reaction buffer, 5 µl high GC, 0.5 µl dNTPs, 5 µl index primers, 1 µl DNA template, and molecular grade water with a total volume of

25 µl. Two single-taxon DNA templates (primer pair 1) separate and combined were used as positive controls and three clean-up blanks were included. The following thermocycling protocol was used: initial denaturation at 95°C for 2 min, then 8 cycles of: 95°C for 15 seconds, 55°C for 30 seconds, then extension at 72°C for 30 seconds. Followed by a final extension at 72°C for 10 min. PCR success was determined by visualising amplicons on a 1.5% agarose gel.

The following sequencing steps were carried out by Tim Goodall (UKCEH Wallingford). Libraries were normalised using SequalPrep Normalization Plate Kit (Thermo Fisher Scientific) and quantified using Qubit dsDNA HS kit (Thermo Fisher Scientific). The pooled library was further purified by gel extraction (QIAquick, Qiagen) and diluted to achieve 400 pM with 7.5% Illumina PhiX. Denaturation of each library was achieved with addition of 10% final volume of 2N NaOH, incubated at room temperature for 5 minutes followed by neutralisation with an equal volume of 2N HCl. The library was then diluted to its load concentration with Illumina HT1 Buffer. A final denaturation was performed by heating to 96°C for 2 minutes followed by cooling in crushed ice. Sequencing was performed on Illumina MiSeq using V3 600 cycle reagents.

Brief methods for bioinformatics processing and data analysis are shown here. Further details and justification of method choice can be found in the Methods section of Chapter 4.

Pre-processing of raw Illumina MiSeq paired-end reads was done using the MetaWorks v1.8.1 pipeline available from https://github.com/terrimporter/MetaWorks (Porter and Hajibabaei 2020a). The demultiplexed paired-end reads from Illumina MiSeq were merged using SEQPREP v1.3.2 from bioconda using the default MetaWorks settings of: minimum Phred quality score of 13 in the overlap region and at least a 25 bp overlap. Primers were trimmed based on their sequences using CUTADAPT v3.2 from bioconda. The forward primer is trimmed first and the output from this step is used as the input for trimming the reverse reads. The MetaWorks default settings were used for the minimum Phred quality score at the ends (≥20) and the allowance of no more than 3 Ns. The minimum length of trimmed reads was kept at the default setting of 150 bp. Reads were dereplicated, using VSEARCH

v2.15.2 from bioconda, only retaining unique sequences. Exact sequence variants (ESVs) were generated using the unoise algorithm and rare clusters (clusters containing less than three reads) were removed with the uchime3_denovo algorithm.

Taxonomic assignment of ESVs was done using BOLDigger v1.2.5 available from https://github.com/DominikBuchner/BOLDigger (Buchner and Leese 2020). The BOLDigger best hit table was joined to the ESV table that was produced using the MetaWorks pipeline. Positive and negative controls were checked for unexpected sequence reads. ESVs were filtered for target taxa only (all taxa in the phylum Arthropoda) and filtered to remove sequences that were less than 95% similar to reference sequences. ESVs were clustered manually using taxonomic assignments for each dataset. The metabarcoding data were not filtered so as to retain low abundance/rare taxa. Read abundance data were converted to percentage arthropod read abundance within the sample. Percentage read abundance was also recalculated as a percentage of all crustacean reads rather than all arthropod reads.

Percentage crustacean read abundance showed a better relationship with morphological abundance (within samples) due to *Chaoborus* dominating the percentage arthropod read abundance and obscuring the relative abundance of the crustacean taxa (see Chapter 4). Percentage crustacean read abundance data were therefore used to infer relative abundances in Blelham Tarn.

5.3.6 Specific primer design and optimisation

To develop assays to detect the target prey in the gut contents of *Chaoborus* individuals, specific primers targeting the five dominant zooplankton prey taxa (*Bosmina*, *Ceriodaphnia*, *Daphnia*, *Eudiaptomus* and *Mesocyclops*) were designed using alignments of the sequences from individual barcoding (Folmer region: 658 bp), the sequences from metabarcoded community samples (313 and 205 bp), and reference sequences downloaded from NCBI and BOLD (>500 bp). Alignments of the sequences and specific primer design were done using the MAFFT v7 (Katoh and Standley 2013) plugin in Geneious Prime (Version 2021.2). Reference sequences that showed very low similarity to the local sequences were checked and removed if there was a possibility that they were

either database errors or caused by high geographic variation within the taxon. Specific primers were designed to produce products of varying lengths between 100 and 350 bp so that the bands could be distinguished on electrophoresis gels if used together in multiplex reactions. The melting temperatures (Tm) of the primers were designed to be within 5°C of each other to allow for the possibility of them being used together in multiplex reactions. Other requirements for primers were: primer length of 18-25 bp, GC content of 40-60%, inclusion of a GC clamp, a maximum three Gs/Cs in the last five bases at the 3' end. Values for hairpins, self-dimers, and hetero-dimers for all designed primers were checked using the IDT OligoAnalyzer™ Tool (https://eu.idtdna.com/pages/tools/oligoanalyzer). Where values suggested undesirable features, e.g. stable hairpins, primers were redesigned.

Primers that matched the above criteria were then tested *in silico* against all the target and non-target sequences including the dominant zooplankton taxa and other taxa found in community samples through metabarcoding analyses. Each of the primers tested *in silico* was found to only bind to its target taxa and not to the other targets or to the non-target taxa from the community samples.

To ensure DNA template used for optimisation of specific primers was of high concentration and volume for multiple tests, nested PCR was done using DNA template extracted from multiple individuals within each target genera using Qiagen DNeasy Blood and Tissue kits following the manufacturer's protocol (see Chapter 4 for details on these extractions)

The first round of the nested PCR amplified the full length Folmer region of the COI gene using two sets of primers for the same region (the original Folmer primers (LCO1490 and HCO2198) (Folmer et al. 1994) and primers modified to better amplify zooplankton taxa (ZplankF1 and ZPlankR1) (Prosser et al. 2013). PCRs were run on a Veriti 96-well Thermal Cycler. The PCRs were set up with 2x Amplitaq Gold 360 Master Mix, 0.5 µM of each primer, 6 µl of DNA template, and molecular grade water with a total volume of 20 µl. One negative control was used. The following thermocycling protocol was used: initial denaturation at 95°C for 10 min, then 40 cycles of: 95°C for 1 min, 48°C for 45 seconds, then extension at 72°C for 1 min. Followed by a final extension of 72°C for 7 min.

PCR success was determined by visualising amplicons on a 0.8% agarose gel. The PCR product with the stronger band on the gel was chosen for use in the second round of nested PCR. PCR product from the Folmer primers was used for *Chaoborus, Cyclops, Mesocyclops,* and *Eudiaptomus* and PCR product from the Prosser primers was used for *Daphnia*, *Ceriodaphnia*, and *Bosmina*. The PCR products were cleaned using Zymo DNA Clean and Concentrator-100 kit and diluted for use in the second round PCR.

The second round of nested PCR then tested whether the new specific primers (designed in this study) amplified the target taxa and determined optimum annealing temperatures. PCRs were set up as above but used 2 µl of DNA template. DNA template from two dilutions of the first round PCR (10^{-2} and 10^{-3}) and a gradient of annealing temperatures was used (48-70°C). A positive control (*Chaoborus* DNA template with Folmer primers) and a negative control for each of the specific primer pairs were used. PCR success was determined by visualising amplicons on a 1% agarose gel. Following guidance for increasing detection sensitivity in prey detection studies 'not to use the highest annealing temperature that allows an amplification to be obtained, but instead to decrease it to a level where the specificity for the assay still is assured' (Sint et al. 2011), annealing temperatures just below the highest annealing temperature with a band on the gel were chosen.

Specific primers were then tested against non-target taxa using the chosen annealing temperatures. PCRs were set up as above but each primer pair was only run at a fixed annealing temperature (*Bosmina* and *Daphnia* at 62°C, *Mesocyclops* at 66°C, and *Ceriodaphnia* and *Eudiaptomus* at 68°C). DNA template for the target taxa was used for a positive control for each primer pair. PCR success was determined by visualising amplicons on a 2% agarose gel (increased percentage to gain clearer bands for shorter length products). Where primers were found to be non-specific at the chosen temperature, further gradient PCRs were run (following the above protocol) to assess specificity at higher temperatures.

5.3.7 Chaoborus dietary analyses

To prepare *Chaoborus* individuals for dietary analysis, individuals from the predator samples were thawed, rinsed in molecular grade water, and the length of the head capsules were measured immediately prior to DNA extraction. The gut contents of *Chaoborus* individuals were extracted using the modified Extract-N-Amp protocol (used for barcoding individual zooplankton) in a total final volume of 4.5 µl. The protocol varied slightly to maximise the DNA from the gut contents rather than the *Chaoborus* individual: the abdomen was split open in several places and the gut contents were squeezed out into the extraction mixture with sterile needles. The remaining *Chaoborus* tissue was then removed from the mixture prior to transfer into a PCR tube for lysis.

Primer pairs that were specific to their target taxa were tested using DNA extracted from *Chaoborus* gut contents. PCRs were set up as above using fixed annealing temperatures (*Bosmina* and *Daphnia* at 62°C, *Mesocyclops* at 68°C, and *Eudiaptomus* at 70°C). Reaction volumes and DNA template concentrations were optimised for the low concentration of the DNA template from *Chaoborus* gut contents.

Primer pairs that successfully amplified DNA from *Chaoborus* gut contents were tested in multiplex PCR reactions so that the total DNA extracted from the gut contents of an individual *Chaoborus* could be tested for multiple prey taxa in a single reaction. Two different PCR master mixes were used to assess whether a master mix designed for multiplex reactions provided better results than the master mix used in previous single reactions. Multiplex reactions were tested against each target taxa individually, a mock community of the two target taxa, a mock community of all the taxa, and a mock community of all the non-target taxa. Both PCRs used up to 6 µl of DNA template, and molecular grade water with a total volume of 20 µl. One negative control and single reaction positive controls (using the specific primers and their target taxa DNA template) were used. PCRs were set up following the manufacturers' protocols for each of the master mixes. The first PCRs were set up as above using 2x AmpliTaq Gold 360 Master Mix. The second PCRs were set up using 2x Qiagen Multiplex Master Mix with 0.4 µM of each primer and the following thermocycling protocol was used:

initial denaturation at 95°C for 15 min, then 35 cycles of: 94°C for 30 seconds, 62°C for 90 seconds, then extension at 72°C for 1 min. Followed by a final extension of 60°C for 30 min. PCR success was determined by visualising amplicons on a 2% agarose gel.

Following optimisation, gut contents extracted from twenty *Chaoborus* individuals from each of three sampling dates (13th August, 27th August and 10th September) were analysed for the presence of *Daphnia* and *Bosmina* (the two target taxa included in the optimised multiplex assay). PCRs were set up using Qiagen Multiplex Master Mix as above but using the total volume of DNA extracted from individual gut contents (4.5 µl) and a total reaction volume of 13.5 µl for 40 cycles. Positive and negative controls were used for each PCR.

5.3.8 Data Analysis

Data analyses were carried out in Microsoft Excel and the bubble and Sankey plots were plotted in R Statistical Software v4.0.2 (Core Team, 2020).

5.4 Results

5.4.1 Barcoding

Barcode sequences received from the Natural History Museum, London were checked for sequence quality (Geneious Prime (Version 2019.2) and compared against reference sequences in NCBI (nt) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and BOLD v4.0 (www.boldsystems.org). Poor quality sequences and sequences that did not match to expected reference sequences were discarded from further analyses. High quality sequences for the following genera were retained (number of individuals shown in brackets): *Chaoborus* (7), *Bosmina* (2), *Ceriodaphnia* (6), *Chydorus* (1), *Daphnia* (11), *Leptodora* (5), *Cyclops* (2), *Mesocyclops* (7), *Arctodiaptomus* (1), *Eudiaptomus* (7).

5.4.2 Community metabarcoding

Percentage arthropod read abundance (for both short and long amplicons) for all community samples from Blelham Tarn provide an initial assessment of the composition of the prey community (short amplicon Figure 5.2, long amplicon

Figure 5.3). In total, 19 genera were detected in the Blelham Tarn zooplankton samples. Five genera (*Ceriodaphnia*, *Daphnia*, *Eudiaptomus*, *Mesocyclops* and *Chaoborus)* were consistently present throughout the sampling period. Twelve genera were only occasionally present, and usually at low read abundance. These genera were therefore not considered to be dominant members of the zooplankton community and some of these genera are also known to usually inhabit littoral or benthic habitats rather than pelagic habitats. Two genera (*Bosmina* and *Macrocyclops*) had low read abundance but showed some consistency in presence throughout the sampling period. *Bosmina* were not detected using the short amplicon in general primer optimisation (see Chapter 4) (barcoding and metabarcoding results suggest there might be differences in the *Bosmina* species present in the different lakes) and, given their small body size, low read abundance relative to larger taxa (e.g. approximate average sizes: Bosmina ~0.3 mm, Daphnia ~1.3 mm, Mesocyclops ~1.5 mm (Brooks and Dodson 1965)) is expected. In contrast, *Macrocyclops* is a large copepod, so low read abundance of this genus suggests it is not likely to be abundant in these samples.

Community metabarcoding also provided information on the behaviour of *Chaoborus*. Percentage read abundance for *Chaoborus* in day-time samples shows a pattern of lower read abundance for the shallow (0-4 m) samples and higher read abundance for the deeper samples (Figures 5.2 and 5.3). In contrast, percentage read abundance for *Chaoborus* in night-time samples shows either consistent read abundance across depths or higher read abundance in the shallower samples. This pattern is consistent with personal observation of samples from Blelham Tarn that suggest *Chaoborus* exhibit diel vertical migration (DVM) and that 0-4 m night-time samples are preferable for preliminary dietary analyses.

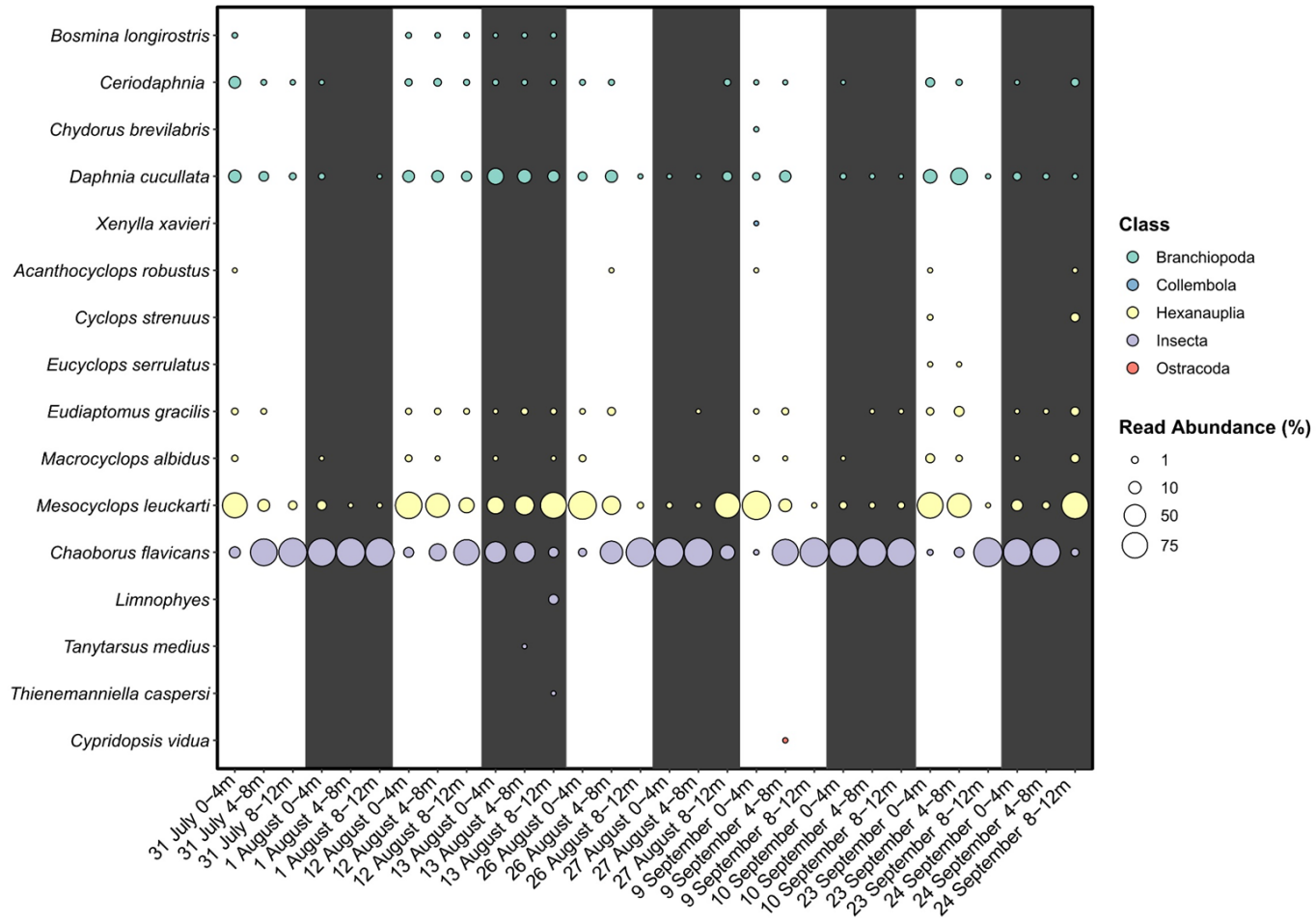

Figure 5.2 *Percentage arthropod read abundance (short amplicon) for Blelham Tarn zooplankton community samples (July to September 2019). Samples taken at three depths (0-4 m, 4-8 m, 8-12 m) on each date. Day-time samples shown with a white background, night-time samples shown with a grey background.*

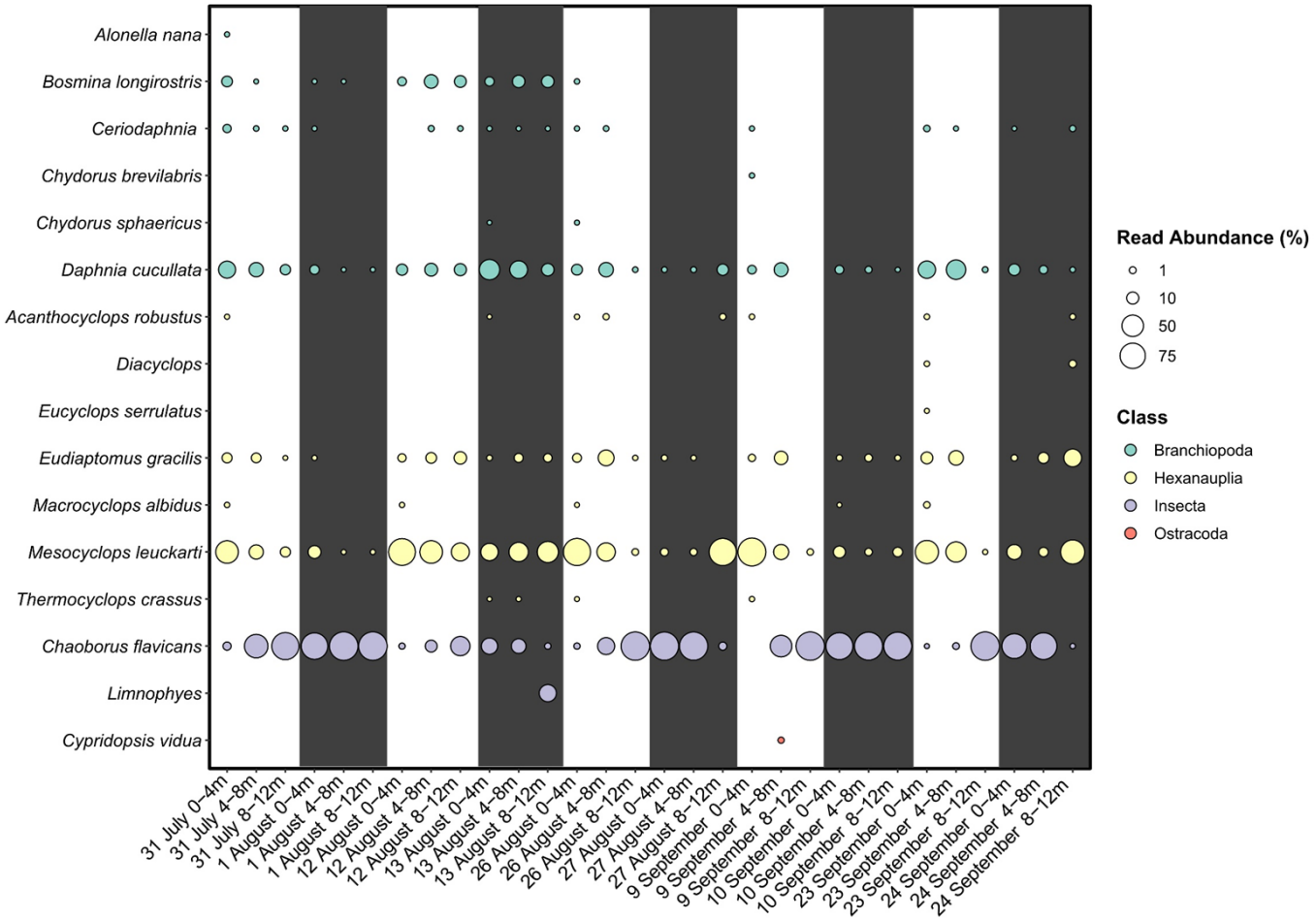

Figure 5.3 *Percentage arthropod read abundance (long amplicon) for Blelham Tarn zooplankton community samples (July to September 2019). Samples taken at three depths (0-4 m, 4-8 m, 8-12 m) on each date. Day-time samples shown with a white background, night-time samples shown with a grey background.*

Percentage arthropod read abundance from both amplicons for night-time samples collected at 0-4 m (Figure 5.4) suggest that the potential diet for *Chaoborus* in these samples was predominantly composed of *Mesocyclops* and *Daphnia*. Although the percentage read abundance for *Ceriodaphnia* and *Bosmina* is relatively low, the agreement between the two amplicons suggests they were present in these samples and the low read abundance is likely to be due to their small size and the difficulty in detecting *Bosmina* with the short amplicon. As seen in primer optimisation (Chapter 4), the high read abundance for *Chaoborus* (which both amplicons showed strong bias towards) can cause false negatives for smaller/low abundance taxa and it is noteworthy that both amplicons show reads for *Bosmina* where the percentage read abundance for *Chaoborus* is relatively lower. In contrast, the inconsistent and low percentage read abundance of the larger taxa, *Eudiaptomus* and *Macrocyclops*, suggest these taxa were likely to only be present at very low abundance in these samples.

Percentage crustacean read abundance provided a better representation of abundance than percentage arthropod read abundance due to the bias towards *Chaoborus* (Chapter 4). The percentage crustacean read abundance for the night-time, 0-4 m samples (Figure 5.5) suggests that the most abundant potential prey taxa in all samples were *Mesocyclops* and *Daphnia*. *Eudiaptomus* (larger taxon) is likely to have been present at lower abundance than *Mesocyclops* and *Daphnia* in all but one of the samples (27th August). The samples with reads for the smaller taxa, *Ceriodaphnia* and *Bosmina*, might be the only samples where these taxa were present but false negatives are more likely for these taxa.

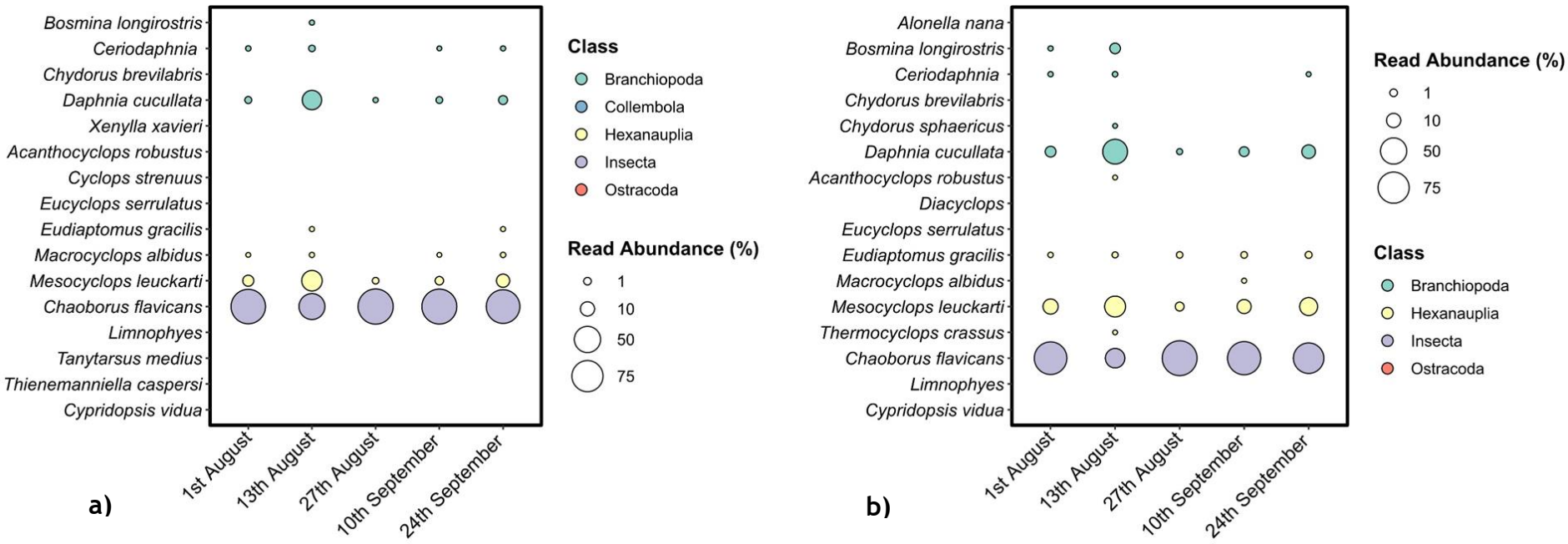

Figure 5.4 *Percentage arthropod read abundance (a: short amplicon, b: long amplicon) for Blelham Tarn zooplankton community samples (night-time, 0-4m samples from July to September 2019).*

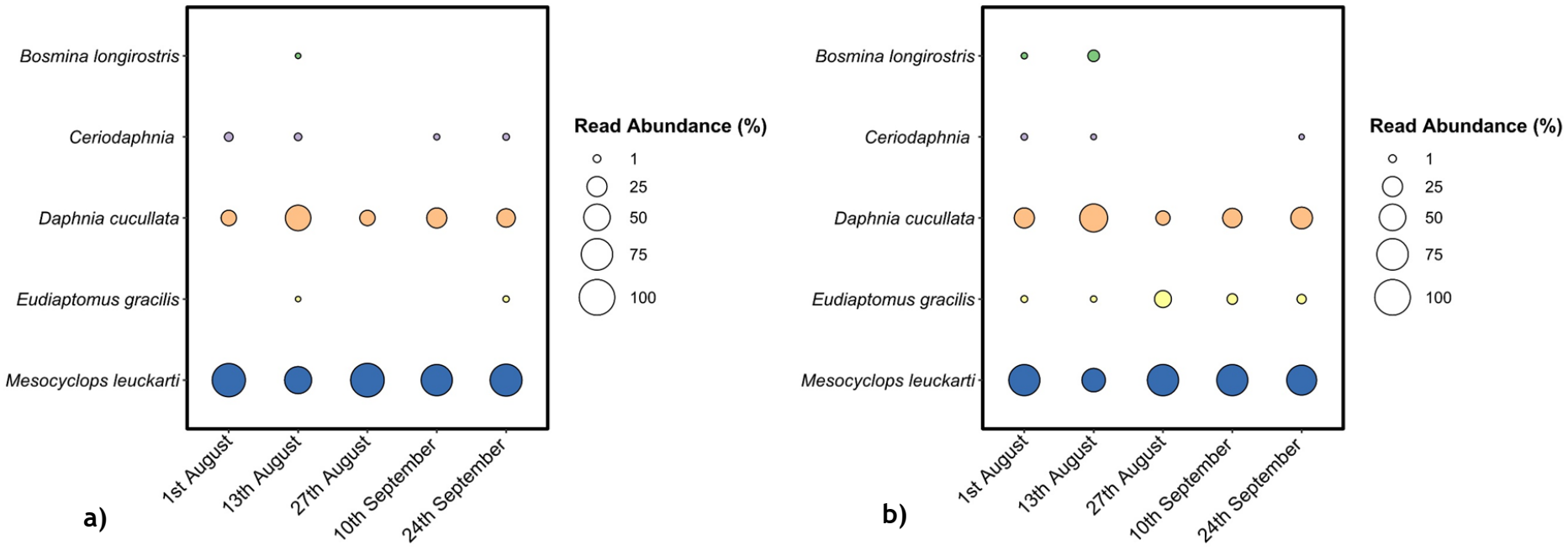

Figure 5.5 *Percentage crustacean read abundance (a: short amplicon, b: long amplicon) for Blelham Tarn dominant zooplankton taxa (night-time, 0-4m samples from July to September 2019).*

5.4.3 Specific primer design and optimisation for potential prey

Design of specific primers for the potential prey taxa present in the night-time 0-4 m samples resulted in ten new primers (five assays) (Table 5.1). Although these primers are likely to detect multiple species within each genus, they have been optimised and validated to detect the species found in Blelham Tarn and so would need further validation to assess their suitability for use in other locations. The assays were designed to amplify different product lengths to enable their use in multiplex reactions.

All specific assays amplified the target taxa in single assay tests and optimal annealing temperatures for each assay were selected from gradient PCR results (Table 5.2). Specificity tests showed that the assays for *Bosmina* and *Daphnia* were specific at the chosen optimal temperatures but the assays for *Ceriodaphnia*, *Eudiaptomus* and *Mesocyclops* also amplified other taxa present in the samples. Raising the annealing temperatures for those assays (within the limits for target amplification shown in gradient PCRs) resulted in specific assays for *Eudiaptomus* and *Mesocyclops* but the *Ceriodaphnia* assay still amplified *Daphnia* at the maximum annealing temperature for target amplification. Dietary sample tests (using DNA extracted from *Chaoborus* gut contents) for the four specific assays resulted in positive detections for *Daphnia* and *Bosmina*. No positive detections were found for *Eudiaptomus* or *Mesocyclops*. This could be because the *Chaoborus* individuals used in these tests had not consumed these taxa. However, some visual evidence of copepods in *Chaoborus* gut contents during the DNA extraction process suggests that copepods are consumed. It is more likely that the designed assays are not sensitive enough to detect the low concentration of DNA found in individual *Chaoborus* gut contents in comparison with the higher concentration of the DNA template used in assay testing, especially as the annealing temperatures had to be raised for these assays in order to achieve specificity.

The positive detections of *Bosmina* and *Daphnia* in the gut contents of *Chaoborus* showed that these assays were sensitive enough to detect the low prey DNA concentrations found in the gut contents. The very low volume and concentration of DNA extracted from *Chaoborus* gut contents meant that using

all the DNA from an individual in one multiplex reaction was necessary in order to optimise the detection sensitivity. The assays for *Bosmina* and *Daphnia* also had identical optimal annealing temperatures and so were more likely to be successful in multiplex reactions. Multiplex reaction tests resulted in positive detections for both *Bosmina* and *Daphnia* from DNA extracted from *Chaoborus* gut contents, including some individuals that tested positive for both taxa. The validation tests resulted in one multiplex assay for *Bosmina* and *Daphnia* for analysing dietary samples from *Chaoborus* individuals from Blelham Tarn.

Table 5.1 Specific primers designed to amplify target zooplankton genera found in Blelham Tarn. Range of product lengths to enable primers to be used in multiplex reactions. The presence of a GC clamp at the 3' end of the primers to help promote specific binding is shown in red.

Primer name	**Sequence** (5' to 3')	**Primer length** (18-25 bp)	**Product length** (100-350 bp)	**Tm** (Oligo conc. 0.5 uM)
Daphnia COI F	CAGGGATCTCATCAATTCTTGG	22	126	54.1
Daphnia COI R	GTAGGAGTGCGGTGATTCC	19		56.3
Ceriodaphnia COI F	TTGACTAGTGCCTTTAATGTTAGGG	25	307	55.5
Ceriodaphnia COI R	CGGAATTCGATCTAAAGTTATCCC	24		54
Bosmina COI F	TGGAACTGGGTGAACTGTTTACC	23	104	58
Bosmina COI R	AAGAAATACCCGCCAAATGTAAGG	24		56.4
Eudiaptomus COI F	CGGCACTAATCAATTTCCAAACC	23	185	55.7
Eudiaptomus COI R	GAGCTTGGTCAGGCATAGTCG	21		58.6
Mesocyclops COI F	AGACACACCCGCTAAATGAAGG	22	156	58.1
Mesocyclops COI R	TTAGTGCCTGCCTTGTTTATGC	22		56.8

Table 5.2 *Results of specific assay tests to assess suitability for use in analysing the diet of* Chaoborus*. Each assay was tested using gradient PCR reactions using DNA template for the target and non-target taxa that are abundant in the community. Optimal annealing temperatures were determined. Each assay was then tested with the optimal annealing temperatures for amplification and specificity using DNA from* Chaoborus *gut contents. Two assays were then tested in a multiplex reaction using target and non-target DNA template and DNA from* Chaoborus *gut contents.*

Assay	**Single assay: target taxa**	**Optimal annealing temperature: target taxa**	**Specific at optimal temperature?**	**Temperature required for specificity**	**Dietary sample tests**	**Multiplex tests**
Bosmina	✓	62°C	✓	-	✓	✓
Ceriodaphnia	✓	68°C	×	×	-	-
Daphnia	✓	62°C	✓	-	✓	✓
Eudiaptomus	✓	68°C	×	70°C	×	-
Mesocyclops	✓	66°C	×	68°C	×	-

5.4.4 Chaoborus dietary analyses

Preliminary dietary analyses were carried out using *Chaoborus* individuals from the middle three night-time, 0-4 metre depth samples (13th August, 27th August, and 10th September 2019).

The proportion of *Chaoborus* individuals (20 individuals from each sampling date) testing positive for only *Bosmina*, only *Daphnia*, both *Bosmina* and *Daphnia*, or neither taxa showed variation across the three dates (Figure 5.6). The proportion of *Chaoborus* individuals that tested positive for *Bosmina* was highest (70% in total) in the 13th August sample and dropped to 30% in the following two samples. The proportion of *Chaoborus* individuals that tested positive for *Daphnia* was 40% in the first sample (13th August), rose to 65% in the middle sample (27th August) and was lowest in the final sample (10th September) at 35%. A higher proportion of *Chaoborus* individuals consumed both prey taxa on the 13th August (30%) compared to the 27th August and 10th September samples (10 and 5% respectively). The proportion of *Chaoborus* individuals testing negative for both taxa was low in the first two samples (20 and 15%) but rose to 40% in the 10th September sample.

The proportion of *Chaoborus* individuals testing positive for only *Bosmina*, only *Daphnia*, both *Bosmina* and *Daphnia*, or neither taxa also showed variation across the different size groupings of *Chaoborus* individuals (Figure 5.7). Fifty percent of the smallest size grouping (0.25 mm head capsules) tested positive for *Bosmina* and none tested positive for *Daphnia*. In the next size grouping (0.5 mm), 52% tested positive for *Bosmina* and 32% tested positive for *Daphnia*. In both the larger size groupings (0.75 mm and 1 mm), the proportion of individuals testing positive for *Bosmina* was lower (38% and 44% respectively), whilst the proportion testing positive for Daphnia was larger (>50%) than in the smaller size groups.

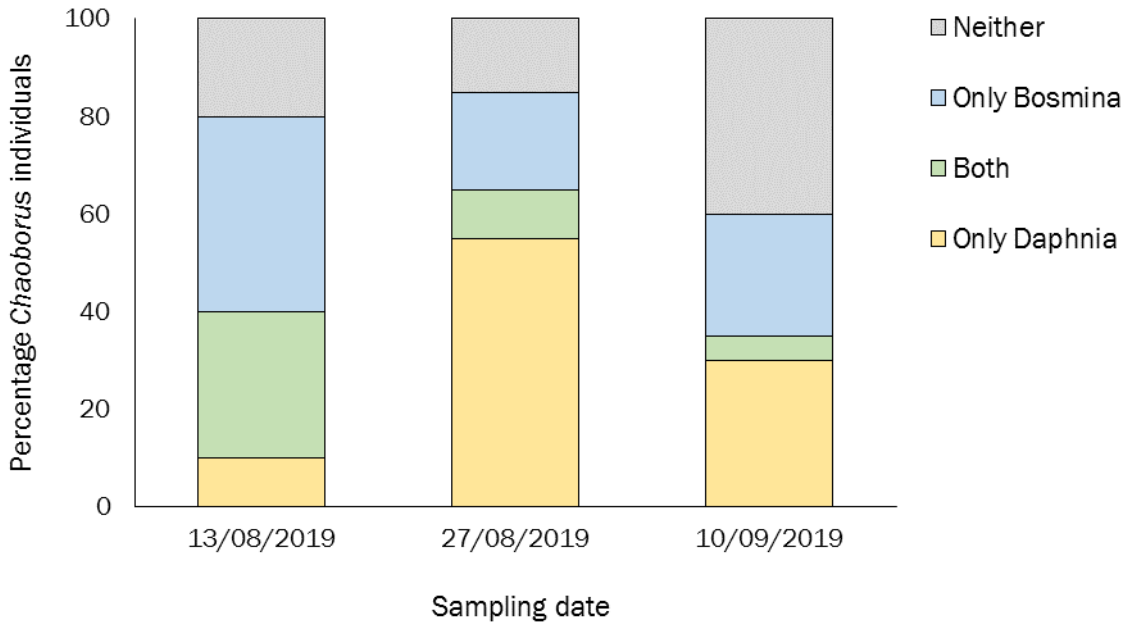

Figure 5.6 *Proportion of all Chaoborus individuals that had a positive PCR result for: only Bosmina, only Daphnia, both Bosmina and Daphnia (samples: night-time, 0-4 m from 13/08/2019, 27/08/2019, 10/09/2019).*

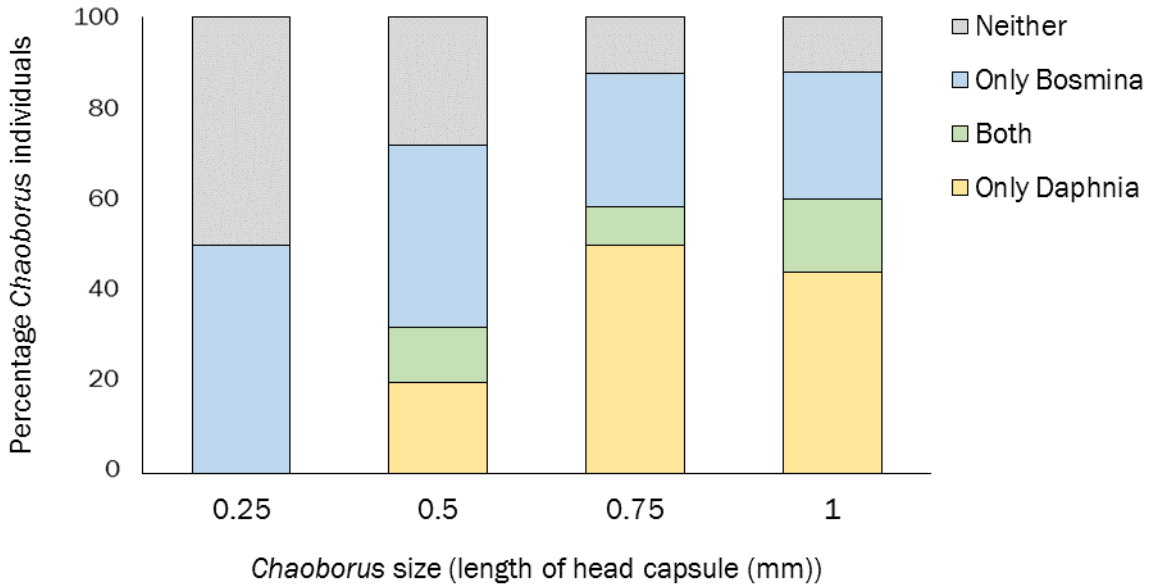

Figure 5.7 *Proportion of four sizes of Chaoborus individuals that had a positive PCR result for: only Bosmina, only Daphnia, both Bosmina and Daphnia (Chaoborus individuals pooled from night-time, 0-4 m samples from 13/08/2019, 27/08/2019, 10/09/2019).*

Treating the different sizes of *Chaoborus* as the nodes of an interaction network enabled the changes in interaction strengths with the two prey taxa (proportion of predators for which a specific interaction was detected) to be visualised. Pooling all the interaction data across sampling dates smooths out some of the stochastic variation caused by differences among sampling days to produce a composite picture of the predator-prey interactions (Figure 5.8), while separating the interaction data by sampling date shows how some of the differences between dates are likely to be caused by changes in the sizes of *Chaoborus* individuals present in those samples as the summer progressed (Figure 5.9).

The first sample (13/08/19) had a much larger proportion of small *Chaoborus* individuals than the two later samples. The smallest *Chaoborus* size group (0.25 mm) was only present in this first sample and the next size group (0.5 mm) was absent in the second sample (27/08/19) and accounted for only a very small proportion (6.7%) of individuals in the last sample (10/09/19). The smallest individuals (0.25 mm) consumed only *Bosmina*. The majority of 0.5 mm individuals tested positive for *Bosmina* (66.7%).

In contrast, the majority of the larger individuals (0.75 mm and 1 mm) tested positive for *Daphnia* (overall 65% and 63.2% respectively). The proportion of these larger *Chaoborus* individuals testing positive for *Daphnia* varied slightly between sampling dates. The proportion of 0.75 mm individuals was very low in the 13/08/19 sample and 50% of these individuals tested positive for *Daphnia*. In the 27/08/19 and 10/09/19 samples where there were more 0.75 mm individuals, 75% and 60% tested positive for *Daphnia* respectively. Similarly, the proportion of 1 mm individuals was low in the 13/08/19 sample and 50% of these individuals tested positive for *Daphnia*. In the 27/08/19 and 10/19/19 samples where there were more 1 mm individuals, 63.6% and 75% tested positive for *Daphnia* respectively.

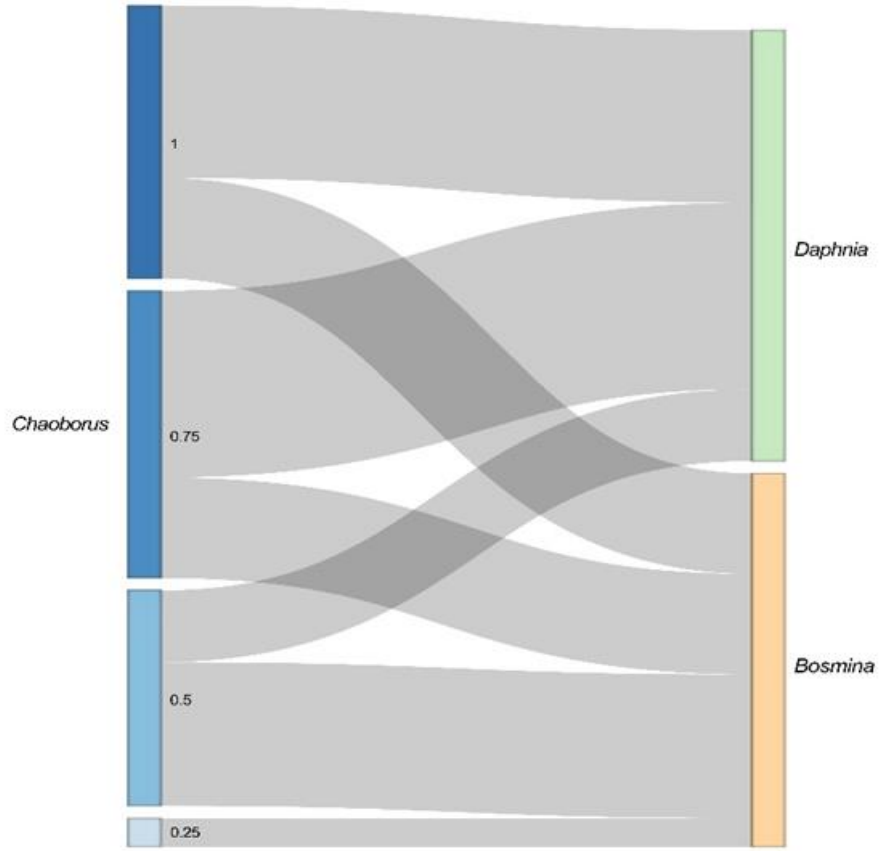

Figure 5.8 *Composite interaction strengths between four Chaoborus sizes and two prey genera (Daphnia and Bosmina) in Blelham Tarn (pooled samples: night-time, 0-4 m, 13/08/2019, 27/08/2019, 10/09/2019).*

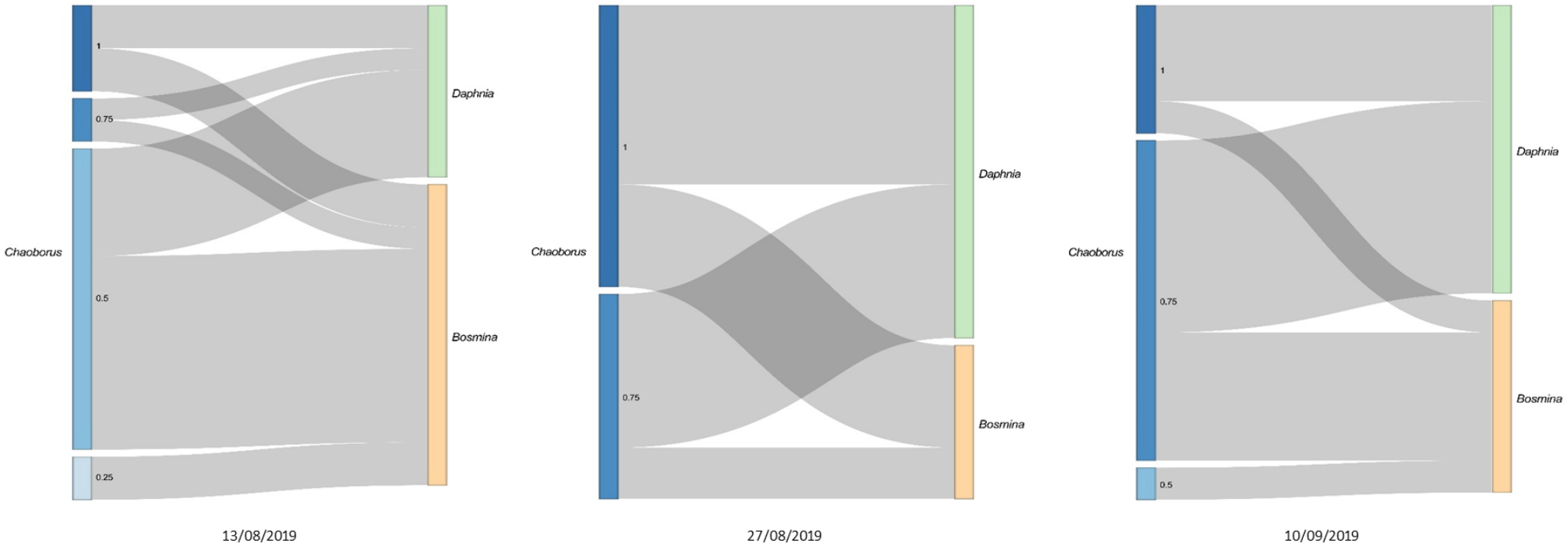

Figure 5.9 *Interaction strengths between four Chaoborus sizes and two prey genera (Daphnia and Bosmina) in Blelham Tarn (samples: night-time, 0-4 m, 13/08/2019, 27/08/2019, 10/09/2019).*

5.5 Discussion

In order to assess trophic interactions using DNA-based identification methods it is recommended that the potential diet is known *a priori* as it is usually not possible to get a complete diet assessment, or choose the best method or reference library, without some knowledge of the consumer's feeding behaviour and the available resources in the habitat (Nielsen et al. 2018). Metabarcoding of community samples can provide data on what species are present in the community, co-occurrence of predator and prey species, habitat use, and behaviour. In addition, metabarcoding provides sequences of the target species which can be used in the development and optimisation of assays for the target prey. With this *a priori* knowledge and optimised assays, the interactions that underpin ecosystem functioning and stability can be assessed more accurately and efficiently. This study aimed to develop an approach for dietary assessment of the predatory zooplanktivore *Chaoborus* by using metabarcoding to provide zooplankton community assessments for Blelham Tarn; taking this community information to optimise diet analyses to detect the potential prey of *Chaoborus* in dietary samples; and analyse the diet of *Chaoborus* individuals to resolve interactions with potential prey taxa.

5.5.1 Zooplankton community metabarcoding

Percentage arthropod read abundance (for both the short and long amplicons) for community samples from Blelham Tarn provide data on community composition that is informative for developing and using DNA-based methods for subsequent dietary analyses.

Firstly, community metabarcoding data provided information on the potential diet and essential information for assay design and optimisation. A total of 19 genera were detected with five of these consistently detected throughout the sampling period. Twelve genera were only occasionally present and usually at low read abundance, so were not considered to be dominant members of the pelagic community. Some of these taxa were known to inhabit littoral or benthic habitats rather than pelagic, suggesting that the low read abundance and occasional detections were due to very small numbers of individuals

occasionally being found in the pelagic samples. Two genera had low read abundance but showed some consistency in detection frequency (*Bosmina* and *Macrocyclops*). Prior knowledge of their body size and the low amplification success of *Bosmina* during optimisation (see Chapter 4) enabled better judgement of whether these taxa were likely to be dominant members of the community. Knowledge of the dominant potential prey taxa is essential so that assays can be developed that can detect the presence of these taxa. In addition, knowledge of which other taxa might occasionally be found in the samples is essential to ensure that specific assays will not amplify these non-target taxa.

Secondly, community metabarcoding provided information on the behaviour and habitat use of *Chaoborus,* enabling samples for preliminary dietary analyses to be chosen. Although variation in the percentage read abundance must be treated with caution due to the biases in the metabarcoding process, *Chaoborus* reads showed a consistent pattern that is indicative of diel vertical migration (DVM). Within-species, across-sample read abundance did not show a strong relationship with abundance of individuals for *Chaoborus* in optimisation of these primers (see Chapter 4) but this pattern is consistent with expected behaviour of *Chaoborus* (Christjani and Von Elert 2015; Weisser et al. 2018) and personal observation. This pattern in abundance might not have been evident during primer optimisation as the samples were from Esthwaite Water where *Chaoborus* abundance was lower.

DVM is thought to be an adaptation to reduce predation risk by visual predators (fish) during the day (Voss and Mumm 1999). When fish are present, *Chaoborus* is thought to migrate to deeper water during the day when predation risk is high and move to shallower waters at dusk to feed when predation risk by is lower (Weisser et al. 2018). The pattern shown by *Chaoborus* read abundance suggested that DVM could be occurring in Blelham Tarn and therefore enabled preliminary dietary analyses to focus on the samples where *Chaoborus* are more likely to be actively foraging. Further dietary analyses, after optimisation, of *Chaoborus* individuals from the different depths and times would provide further evidence of this behaviour.

Finally, community metabarcoding enabled some inferences to be made about the most abundant potential prey taxa available to *Chaoborus* individuals in the selected samples. Read abundance data cannot be used to provide absolute taxon abundance, but validation of the assays and knowledge of taxon body sizes enabled some inferences of relative taxon-abundance within samples (see Chapter 4). It is likely that *Mesocyclops* and *Daphnia* were abundant prey taxa in all the selected samples and that *Eudiaptomus* was present at lower abundance in most samples. It is likely that the smaller taxa (*Ceriodaphnia* and *Bosmina*) were abundant in samples where they were detected but might also be present in samples where they were not detected (false negatives). Primers that are less biased towards *Chaoborus*, and the application of correction factors for body size, could enable better inferences of relative abundance for the different sizes of prey taxa.

5.5.2 Specific primer design and optimisation for potential prey

Sequences gained via community metabarcoding of bulk samples enabled cross-referencing with public reference databases. In some cases, high intraspecific variation in all the available reference sequences for the taxon made design of specific primers challenging due to lack of conserved primer binding regions. The comparison of sequences gained from metabarcoding from Blelham Tarn enabled publicly available reference sequences to be selected based on similarity. Reference sequences that showed very low similarity were checked and removed if they were possible database errors or caused by high geographic variation within the taxon. This enabled specific primers to be designed to target the variation present in the local samples. *In silico* tests showed that the primers are likely to also amplify other congeneric species, but further validation would be needed to ensure these primers are suitable for use in other locations.

Design, optimisation and validation resulted in one multiplex assay for *Bosmina* and *Daphnia* that was sensitive enough to detect the very small amounts of DNA present in the gut contents of single *Chaoborus* individuals. Although PCR tests showed the primers to be specific to the target taxa, sequencing of the PCR product amplified from *Chaoborus* gut contents could provide an additional

quality check to ensure specificity. Development of specific assays for this purpose is challenging and time consuming but enables a very quick, efficient and cost-effective method to screen individual dietary samples for the presence of the potential prey taxa without many of the biases and limitations of metabarcoding.

5.5.3 Chaoborus dietary analyses

Results from this study showed the potential for DNA-based approaches to reveal the non-static and life-history dependent nature of trophic interactions. Analysis of three night-time, 0-4 m samples showed that the proportion of *Chaoborus* individuals that tested positive for the two potential prey taxa varied between samples and between different sizes of *Chaoborus* individuals. Treating the different sized *Chaoborus* individuals as a single node of an interaction network, masks differences in diet between the different instars. Ontogenetic shifts in diet have been shown to be important in aquatic food webs (Woodward et al. 2010) as there are often substantial changes in diet as organisms grow. Incorporating body size into food web models has been shown to produce a more accurate representation of ecosystem structure and dynamics (Woodward et al. 2005; Thierry et al. 2011). Treating each of the four size categories of *Chaoborus* in this study as separate nodes in a simple interaction network is therefore likely to separate diet changes driven by predator size from diet changes caused by environmental differences (prey availability/abundance/habitat use).

Pooling interaction data across sampling dates provides an overview of the diet of the different size categories of *Chaoborus*. The smallest individuals (head capsule: 0.25 mm) only tested positive for *Bosmina* and the majority (66.7%) of the next size group (0.5 mm) tested positive for *Bosmina*. In contrast, the majority of the larger individuals (0.75 mm and 1 mm) tested positive for *Daphnia* (65% and 63.2% respectively). These pooled results suggest there might be an ontogenetic shift in diet from the smaller cladoceran (*Bosmina*) in early instars to the larger cladoceran (*Daphnia*) as *Chaoborus* grow and their gape-size allows the capture of larger prey (Moore 1988). Changes in the abundance of different sizes of *Chaoborus* were seen across sampling dates that affect the

frequency of overall interactions with the two prey taxa. This means that different patterns of interaction strengths with different prey species are likely as each generation grows and develops.

Bosmina is detected in the diet of *Chaoborus* throughout all three sampling dates even though metabarcoding only detected their presence in the first of these dates (13/08/2019). There might be several explanations for this result. Firstly, the *Bosmina* specific assay is likely to be much more sensitive in detecting *Bosmina* than the general metabarcoding assay and the presence of larger, more abundant taxa in the community samples is likely to swamp the metabarcoding reads so that small, low abundance taxa are less likely to be detected. The lack of evidence of *Bosmina* in the later community samples, when they were detected in the earlier community samples, might suggest *Bosmina* were less abundant in these samples and so less likely to be detected. Lower abundance of *Bosmina* in the community samples could be caused, in part, by the increased abundance of large *Chaoborus* individuals feeding on them. The patterns seen in this study supports the suggestion that predator diet provides a particularly sensitive method of detecting the presence of prey taxa.

The dietary data suggest that small cladocerans, such as *Bosmina*, are a particularly important food source for early instars of *Chaoborus*, which might be unable to handle larger prey. Although there is a shift towards a preference for *Daphnia* in the larger sizes of *Chaoborus*, *Bosmina* remains a part of the diet for the largest size class even though the metabarcoding data suggest that *Daphnia* were abundant throughout the sampling period. Another factor affecting the interactions between these species is changes in the size of the prey taxa. The sizes of *Daphnia* in the samples were particularly varied (personal observation) and it is possible that *Daphnia* individuals reach a size that is too large for even the largest *Chaoborus* to handle. This would explain why *Bosmina* remain an important food source even when *Daphnia* are available. Although molecular methods of detection cannot provide data on the size of the prey eaten by individuals, measurements of individuals in the community samples at the different sampling points would enable some inference of how prey size affects the interactions.

When differences in the size of *Chaoborus* are measured and accounted for, any remaining changes in interaction strengths over time are likely to be caused by changes in the abundance and size of the prey taxa and the habitat use of predator and prey species which are, in turn, related to changes in the environmental conditions. It is these changes in interaction strengths over time that can be monitored and used as indicators of changes in communities. Pooling data across the sampling dates provides a time-averaged, composite snapshot of interaction strengths from August and September 2019 that smooths out some of the stochastic variation caused by differences between sampling days. Changes in interaction strengths between composite snapshots (in subsequent years) could be used to infer and understand the effects of interannual environmental variation on trophic interactions.

5.5.4 Conclusions

Although the metabarcoding process involves potential biases at multiple stages, an understanding of these biases and careful optimisation can allow the production of meaningful ecological data. In this study community metabarcoding provided essential data that formed a basis for understanding community composition and prey availability, provided indications of the behaviour and habitat use of *Chaoborus*, and provided sequences for developing specific assays. Although the development of these assays was complex and time consuming, once the methods were optimised, screening provided an efficient and cost-effective method of obtaining individual-level interaction data across multiple samples. The collection of interaction data at the individual-level in this study showed the importance of ontogenetic shifts in diet in exploring changes in diet over time. Individual-level data from multiple individuals enabled interaction strengths to be compared over time and space, providing a powerful, quantitative method for monitoring community changes that are likely to precede species turnover and loss.

6 General discussion

6.1 Overview

The overall aim of this thesis was to explore how DNA-based resolution of predator diets can contribute to improving the assessment and monitoring of freshwater biodiversity, and to identify the method developments necessary to achieve this improvement.

The studies in this thesis focused on the key research needs and knowledge gaps that were identified by a quick-scoping review of the current state-of-the-art (Chapter 2). These included the need for: comprehensive, high-quality reference databases for the target taxa being investigated (Chapter 3); studies focused on the development and application of DNA-based dietary analysis for freshwater invertebrate taxa (Chapters 3, 4 and 5), especially in standing water habitats (Chapters 4 and 5); careful and question-driven method choice and optimisation of DNA-based methodology (metabarcoding and screening methods) (Chapters 4 and 5); improved understanding of the potential predator-prey interactions through prior knowledge of the prey community, in order to facilitate method development (Chapter 4 and 5); and studies to demonstrate how the diets of individual predators can be resolved and used to identify changes in predator-prey interactions over time, illustrating the potential for monitoring important drivers of change in communities (Chapter 5).

The results of the first two main chapters (Chapters 2 and 3) support the overall hypothesis that molecular research on quantifying biodiversity has so far focused more on using DNA-based methods to replicate existing monitoring approaches, rather than explore the potential for DNA-based approaches to expand our understanding of whole-system biodiversity and predator-prey interactions. Studies using DNA-based resolution of predator diets are biased towards vertebrate taxa, especially specific taxa of conservation concern, or non-native taxa. Publicly stored reference sequences for freshwater arthropods are biased towards taxa that are already monitored by traditional means or are of interest due to conservation concern or their non-native status. DNA-based

approaches can provide more cost-effective and efficient methods of gathering similar types of data to those we have obtained with traditional methods (e.g. species occurrence information over space and time), but it is clear that thorough validation is still needed to understand the strengths and weaknesses of these more novel methods. However, beyond this, DNA-based approaches provide the opportunity to characterise the diverse and vital components of biodiversity that have been less well-studied in the past due to the limitations of traditional methods (e.g. insects, crustaceans and rotifers). Furthermore, these approaches have the potential to resolve individual-level interactions between organisms and potentially provide new, more sensitive methods of monitoring changes in communities.

The studies in this thesis aimed to improve understanding of how DNA-based identification methods can be used to characterise freshwater arthropod communities, a poorly studied group using these techniques, and to resolve trophic interactions within those communities, which have a small body size and are difficult to investigate using traditional methods. Together, they provide the basis for the development of approaches capable of generating new data for monitoring changes in freshwater biodiversity.

Chapter 3 assessed whether publicly stored reference sequences for UK freshwater arthropods provide the necessary coverage and quality for sequences to be accurately assigned to species-level. Although coverage was good overall, it was found to be biased and lacked reference sequences originating from UK specimens. Gaps in coverage coupled with low data quality will lead to poor taxonomic resolution and potential misidentifications of freshwater arthropods.

Chapter 4 explored using DNA metabarcoding to characterise zooplankton communities, and found that thorough optimisation and validation of metabarcoding methods is essential in order to produce meaningful community data. Primers were optimised to amplify the target zooplankton taxa and provided very accurate and sensitive data on the presence of taxa in zooplankton community samples. Metabarcoding detected low abundance taxa that could be overlooked using traditional methods and provided higher

taxonomic resolution than is usually obtained through morphological identification. Optimisation of bioinformatics processing alongside knowledge of the target taxa enabled false positives and negatives to be reduced, improving the reliability of the data.

Chapter 5 optimised specific assays for analysing individual-level trophic interactions between zooplankton using community data from metabarcoding. Metabarcoding of community samples provided sequences for assay design and optimisation, data on the behaviour and habitat use of *Chaoborus* in Blelham Tarn, and enabled the potential prey taxa to be identified. Optimised specific assays for prey taxa provided an efficient and cost-effective method of obtaining individual-level interaction data across multiple samples. The individual-level data showed diet variation between different sizes of individuals, suggesting an ontogenetic shift in diet. Although the development of specific assays was challenging, this chapter showed that this approach has the potential to provide new data, on individual interactions and interaction strengths, that could provide a powerful, quantitative new metric for monitoring changes in communities.

6.2 Synthesis

DNA-based identification of predator diets has the potential to provide new and more informative data for assessing and monitoring freshwater biodiversity. However, obtaining high-quality DNA-based data is dependent on the methodological choices that are made at each stage of the process. Studies that focus on exploring, testing and validating methodologies are essential in driving improvements in the quality of DNA-based data so that the data are accurate and robust, enabling its appropriate use for the benefit of freshwater biodiversity research.

6.2.1 Reference databases

Comprehensive, high-quality reference databases are essential for all DNA-based identification of biodiversity. Studies using DNA-based identification frequently state that gaps in reference databases limit the taxonomic resolution of sequences from samples (Leray and Knowlton 2015). Reference databases

are constantly growing, but gaps are not usually being systematically filled, so coverage is mostly based around interests in specific taxa rather than ensuring whole taxonomic groups are represented. As such, the current "model" for increasing database coverage perpetuates biases that already exist. Systematic filling of the gaps in reference databases is key if DNA-based-identification is to be used more widely to assess ecosystem biodiversity and to address new questions. This would "unlock" the potential of these approaches to move beyond being simply a new method for detecting taxa that are already currently monitored using traditional methods, allowing assessment and monitoring of diverse and vital components of freshwater biodiversity that have been less well-studied in the past. In order to realise the potential of using DNA-based approaches in resolving trophic interactions, particularly for organisms with a small body size, we need to accurately detect prey species in predator diets, and so it is essential that all the potential prey are represented in reference databases.

Public reference databases currently contain large numbers of errors, and misidentified sequences. These errors and misidentifications can cause two problems for taxonomic assignment of metabarcoded sequences from community or dietary samples. Firstly, they can result in a reduction in the taxonomic resolution obtained. Errors and misidentifications can cause conflicts in taxonomic assignment, with a sequence matching reference sequences for multiple taxa. Where this happens, the sequence cannot reliably be assigned to species and must instead be assigned to the taxonomic level where there is agreement. Secondly, these errors can cause false assignment of sequences from samples to the wrong taxon completely. If the only reference sequences available are misidentifications, any new sequences will be falsely assigned to the species listed for those reference sequences. This can have serious implications for characterising the potential prey community and accurately detecting prey using metabarcoding of dietary samples.

Representation of all confamilial species that occur in the geographic region of interest is essential for species-level resolution using DNA-based identification. Intra- and inter-specific variation in a marker region can only be accurately assessed when there are multiple reference sequences for all of the closely-

related species that co-occur in a region. Where only low numbers of sequences are stored, intraspecific variation is likely to be underestimated and interspecific distances are likely to be overestimated. Where species within a family are not represented in reference databases, or only represented by very low numbers of sequences, it is impossible to know whether the marker region can provide accurate species-level resolution e.g. all species in the order Amhipoda. A clear barcoding gap might appear to exist but barcoding of other closely-related species could close this gap, limiting the level of taxonomic resolution possible with the marker. This means that taxonomic assignment of sequences from the potential prey community or the predator diet samples could be falsely assigned to closely-related species. In addition, for some species, intraspecific variation can increase substantially when specimens originate from different countries, e.g. *Isoperla grammatica* (Plecoptera), which can affect accurate taxonomic assignment. Specimen collection location metadata are often not completed when reference sequences are uploaded to databases but this information is important in understanding whether species-level resolution is possible in particular target locations.

In addition, these reference database issues make the design of specific assays for screening of dietary samples challenging. Reference sequences for potential prey taxa are needed in order to design specific assays. If the available reference sequences contain errors or include misidentified sequences, the sequences used to design assays might lack conserved regions for primer binding. Furthermore, if specific assays are designed that mistakenly include sequences from other species, screening methods would detect both species and could result in false positives in dietary analyses.

If DNA-based identification is to be used to assess and monitor freshwater biodiversity, it must provide accurate and high taxonomic resolution so that management decisions can be made with confidence, using the data. Efforts to find markers that can provide high taxonomic resolution, and build reference databases for these markers, are undermined if sequences containing errors or misidentified sequences are included in these databases. Adequate resources for this essential part of using DNA-based methods for biodiversity assessment is crucial. Ongoing curation of publicly stored reference sequences is also vital

if the full potential of DNA-based identification for biodiversity assessment and monitoring is to be realised.

6.2.2 Optimisation and validation

Obtaining high-quality community and dietary data using DNA-based identification is dependent on careful method choice and thorough optimisation and validation of the chosen methods. Where screening methods have been thoroughly researched, developed and validated, standard methods can be used repeatedly to produce accurate, trusted data that can be used to assess and monitor freshwater biodiversity in a consistent way. This kind of standardisation is challenging and expensive in both time and cost and so is usually only done for particular taxa that are of concern. However, standardising screening methods is a much more achievable task than standardising metabarcoding. The metabarcoding process includes many more methodological decisions that affect the data that are produced and the effects can vary depending on the taxonomic composition in different habitats. The 'right' decision at each stage depends on what data are needed for the research question or monitoring objective, and so standardisation across studies is more difficult.

If metabarcoding cannot be easily standardised, it is essential that methods are documented as metadata along with the resulting data and that the effects of each methodological choice are fully understood so that caveats and limitations can be taken into account when interpreting the obtained data or combining data from different studies (e.g. meta-analyses). As shown in this study, optimisation of primers, bioinformatic processing and data analysis can improve the detection of target taxa, reduce false positives and negatives, and improve relative abundance data. Validation of metabarcoding data against data obtained using different methods e.g. microscopy, can improve confidence in the metabarcoding data and provide knowledge of the limitations. In this study, data obtained using microscopy provided essential knowledge about the sizes and abundance of the target taxa in the samples that provided the context to fully understand the read abundance data. A thorough understanding of the strengths and weaknesses of metabarcoding data enables appropriate use of the data for the assessment and monitoring of freshwater biodiversity.

6.2.3 DNA-based resolution of interactions

With comprehensive, high-quality reference sequences and thorough optimisation and validation of methods, metabarcoding can provide sensitive, quick, cost-effective, species-level characterisations of potential prey communities and dietary samples. However, the variation in data caused by various methodological stages throughout the metabarcoding process, risk the occurrence of false positives and negatives. The current lack of reliable abundance information that can be inferred from the data means that using metabarcoding to monitor change in communities or diets is currently challenging and not a direct replacement for traditional methods. DNA-based detections provide surrogate measures of species richness and identity, but apparent changes in species richness could be caused by artefacts of the metabarcoding process rather than true changes in species presence.

Screening of predator diets could provide a powerful, sensitive, quantitative method for monitoring changes in communities and improve our understanding of the frequently hidden interactions in freshwater environments. Optimised and validated screening methods are highly sensitive and accurate at detecting the presence of taxa in samples. Monitoring the frequency of interactions needs only the presence of an interaction rather than abundance information, making this approach highly compatible with semi-quantitative DNA-based approaches. Selection of particular freshwater predators and development of specific assays for their potential prey, to obtain accurate individual-level diet analyses, could enable the monitoring of interaction strengths over time and space, providing much more informative data on change in ecosystem structure than can be obtained from metabarcoding of communities. This provides a huge step forward for food web studies where, in the past, interactions were sometimes inferred from existing literature and therefore assumed to be static in time and space.

Monitoring how interaction strengths change over time and space in food web modules builds on the extensive food web research that has been done on freshwater food webs. These studies reveal that ecosystem functioning and stability are emergent properties of the interactions between individuals

(Stouffer 2010; Thompson et al. 2012b; Staudinger et al. 2021). Predators are structurally important in ecosystems (Estes et al. 2011) and can be sensitive indicators of environmental change (Velarde et al. 2013), therefore monitoring their interactions over time could provide effective early-warning indications of community change.

6.3 Future directions

DNA-based identification of predator diets already offers many benefits to freshwater biodiversity assessment and monitoring, but it is essential that data quality is improved if this approach is to be adopted more widely. Reference databases are key to obtaining accurate and high taxonomic resolution data. If data obtained from DNA-based identification are to be used to inform the management and conservation of freshwater ecosystems, taxonomic assignment cannot rely on incomplete and inaccurate reference databases. It is vital that reference sequences are shared publicly in order to build more complete databases, but errors and misidentifications are inevitable and curation of reference databases needs to be a much higher priority than it is currently.

Optimisation and validation of methods are essential for high-quality data. Small changes in metabarcoding methodology can have large effects on the resulting data and it is essential that these effects are understood when interpreting biodiversity data produced using metabarcoding. There is no one 'best' approach to take, but if DNA-based methods are to be used to assess and monitor potential prey communities and dietary samples, some standardisation for taxonomic groups would enable development efforts to be focused, limitations to be understood, and results to be more comparable.

It is essential that future research is not exclusively focused on using DNA-based identification to enhance or replace existing methods for assessing and monitoring freshwater biodiversity e.g. species richness of vertebrates. Decisions on which taxa and which methods to use to monitor freshwater biodiversity in the past were limited by the methodology available at the time. DNA-based identification provides the opportunity to address some of these

limitations, gaining new insights and data that can help us to better understand, monitor and protect freshwater biodiversity.

6.4 Conclusions

DNA-based resolution of freshwater predator diets is already beginning to improve our ability to carry out assessment and monitoring of freshwater species. Studies are using DNA-based identification methods to gain a better understanding of the ecology of specific taxa, especially for species of conservation concern or non-native species. The most informative and accurate data come from studies that have optimised the methods for the target taxa. I have found strong biases in the application of DNA-based methods to resolving trophic interactions, but that these methods have great potential for resolving food web interactions for difficult-to-observe species, like smaller-bodied, aquatic organisms that are difficult to investigate by other means

Using DNA-based identification to resolve predator diets has the potential to provide large benefits to the assessment and monitoring of freshwater biodiversity but is highly dependent on thorough development. My research shows that prior knowledge of the potential prey community is essential in order to make informed decisions about the methods for resolving predator diets and that metabarcoding of community samples can provide important data about potential prey. In addition, the quality and accuracy of the results depend on the coverage and quality of reference databases and the optimisation and validation of the chosen methodology.

The current potential of DNA-based approaches is hindered by biases and gaps in reference databases. I have produced a systematic framework for prioritising current and future sequencing and curation needs, to address these issues. Optimising and validating DNA-based methods to provide the most accurate data can be challenging and time consuming but once completed, the methods can be used repeatedly to quickly generate large amounts of high-quality data. It is important that efforts in optimisation and validation are publicly available so that they can be built upon and contribute towards future standardisation for the assessment and monitoring of specific groups.

The screening of predator diets offers the potential to provide a powerful and sensitive new method to monitor changes in freshwater communities. Predators are structurally important and can be sensitive indicators of environmental change. When individual predators shift their behaviour due to changes in their environment, it can cause changes in the frequency of interactions with different prey taxa. This frequency of interaction is an important measure of interaction strength between species. Screening methods are highly sensitive at detecting the presence of taxa in samples and cost-effective once developed so are very well suited to analysing large numbers of predator dietary samples. The DNA-based methods developed in this research are sensitive enough to detect small changes in the feeding behaviour of a predator over time and during its lifetime. This research shows that the combination of community metabarcoding of prey taxa and dietary screening of individual predators can provide important ecological insights about how predator-prey interactions change over time that would be extremely challenging without DNA-based methodology.

References

Abell, R. 2002. Conservation Biology for the Biodiversity Crisis: a Freshwater Follow-up. *Conservation Biology* 16(5), pp. 1435–1437. doi: 10.1046/j.1523-1739.2002.01532.x

Alberdi, A., Aizpurua, O., Gilbert, M.T.P. and Bohmann, K. 2018. Scrutinizing key steps for reliable metabarcoding of environmental samples. Mahon, A. ed. *Methods in Ecology and Evolution* 9(1), pp. 134–147. doi: 10.1111/2041-210X.12849

Antich, A., Palacin, C., Wangensteen, O.S. and Turon, X. 2021. To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics* 22(1), pp. 177. doi: 10.1186/s12859-021-04115-6

Arroyave, J. and Stiassny, M.L.J. 2014. DNA barcoding reveals novel insights into pterygophagy and prey selection in distichodontid fishes (Characiformes: Distichodontidae). *Ecology and Evolution* 4(23), pp. 4534–4542. doi: 10.1002/ece3.1321

Athanasio, C.G., Chipman, J.K., Viant, M.R. and Mirbahai, L. 2016. Optimisation of DNA extraction from the crustacean *Daphnia*. *PeerJ* 4, pp. e2004. doi: 10.7717/peerj.2004

Balian, E.V., Segers, H., Lévèque, C. and Martens, K. 2008. The Freshwater Animal Diversity Assessment: an overview of the results. *Hydrobiologia* 595(1), pp. 627–637. doi: 10.1007/s10750-007-9246-3

Bartley, T.J. et al. 2019. Food web rewiring in a changing world. *Nature Ecology & Evolution* 3(3), pp. 345–354. doi: 10.1038/s41559-018-0772-3

Bartley, T.J., Braid, H.E., McCann, K.S., Lester, N.P., Shuter, B.J., Shuter, B.J. and Hanner, R.H. 2015. DNA barcoding increases resolution and changes structure in Canadian boreal shield lake food webs. *DNA Barcodes* 3(1), pp. 30-43.

Beckerman, A., Petchey, O.L. and Morin, P.J. 2010. Adaptive foragers and community ecology: linking individuals to communities and ecosystems. *Functional Ecology* 24(1), pp. 1–6. doi: 10.1111/j.1365-2435.2009.01673.x

Belle, C.C., Stoeckle, B.C. and Geist, J. 2019a. Taxonomic and geographical representation of freshwater environmental DNA research in aquatic conservation. *Aquatic Conservation: Marine and Freshwater Ecosystems* 29(11), pp. 1996–2009. doi: 10.1002/aqc.3208

Belle, C.C., Stoeckle, B.C. and Geist, J. 2019b. Taxonomic and geographical representation of freshwater environmental DNA research in aquatic conservation. *Aquatic Conservation: Marine and Freshwater Ecosystems* 29(11), pp. 1996–2009. doi: 10.1002/aqc.3208

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. 2012. GenBank. *Nucleic Acids Research* 41(D1), pp. D36–D42. doi: 10.1093/nar/gks1195

Bergsten, J. et al. 2012. The Effect of Geographical Scale of Sampling on DNA Barcoding. *Systematic Biology* 61(5), pp. 851–869. doi: 10.1093/sysbio/sys037

Berlow, E.L. et al. 2004. Interaction strengths in food webs: issues and opportunities. *Journal of Animal Ecology* 73(3), pp. 585–598. doi: 10.1111/j.0021-8790.2004.00833.x

Biffi, M. et al. 2017a. Comparison of diet and prey selectivity of the Pyrenean desman and the Eurasian water shrew using next-generation sequencing methods. *Mammalian Biology* 87, pp. 176–184. doi: 10.1016/j.mambio.2017.09.001

Biffi, M. et al. 2017b. Novel insights into the diet of the Pyrenean desman (*Galemys pyrenaicus*) using next-generation sequencing molecular analyses. *Journal of Mammalogy* 98(5), pp. 1497-1507.

Biggs, J. et al. 2015. Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus*). *Biological Conservation* 183, pp. 19–28. doi: 10.1016/j.biocon.2014.11.029

Blackman, R. et al. 2019. Advancing the use of molecular methods for routine freshwater macroinvertebrate biomonitoring – the need for calibration experiments. *Metabarcoding and Metagenomics* 3, pp. e34735. doi: 10.3897/mbmg.3.34735

Boileau, N. et al. 2015. A complex mode of aggressive mimicry in a scale-eating cichlid fish. *Biology Letters* 11(9), pp. 20150521. doi: 10.1098/rsbl.2015.0521

Bonin, A., Guerrieri, A. and Ficetola, F. 2021. *Optimal sequence similarity thresholds for clustering of molecular operational taxonomic units in DNA metabarcoding studies*. Preprints.

Boukal, D.S. 2014. Trait- and size-based descriptions of trophic links in freshwater food webs: current status and perspectives. *Journal of Limnology* 73(s1).

Bradford, T.M., Humphreys, W.F., Austin, A.D. and Cooper, S.J.B. 2014. Identification of trophic niches of subterranean diving beetles in a calcrete aquifer by DNA and stable isotope analyses. *Marine and Freshwater Research* 65(2), pp. 95. doi: 10.1071/MF12356

Brooks, J.L. and Dodson, S.I. 1965. Predation, Body Size, and Composition of Plankton: The effect of a marine planktivore on lake plankton illustrates theory of size, competition, and predation. *Science* 150(3692), pp. 28–35. doi: 10.1126/science.150.3692.28

Buchner, D. and Leese, F. 2020. BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems. *Metabarcoding and Metagenomics* 4, pp. e53535. doi: 10.3897/mbmg.4.53535

Bunch, A.J., Carlson, K.B., Hoogakker, F.J., Plough, L.V. and Evans, H.K. 2021. Atlantic sturgeon (*Acipenser oxyrinchus oxyrinchus* Mitchill, 1815) early life stage consumption evidenced by high-throughput DNA sequencing. *Journal of Applied Ichthyology* 37(1), pp. 12–19. doi: 10.1111/jai.14153

Čandek, K. and Kuntner, M. 2015. DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources* 15(2), pp. 268–277. doi: 10.1111/1755-0998.12304

Carreon-Martinez, L., Johnson, T.B., Ludsin, S.A. and Heath, D.D. 2011. Utilization of stomach content DNA to determine diet diversity in piscivorous fishes. *Journal of Fish Biology* 78(4), pp. 1170–1182. doi: 10.1111/j.1095-8649.2011.02925.x

Carreon-Martinez, L.B., Wellband, K.W., Johnson, T.B., Ludsin, S.A. and Heath, D.D. 2014. Novel molecular approach demonstrates that turbid river plumes reduce predation mortality on larval fish. *Molecular Ecology* 23(21), pp. 5366–5377. doi: 10.1111/mec.12927

Ceballos, G., Ehrlich, P.R., Barnosky, A.D., García, A., Pringle, R.M. and Palmer, T.M. 2015. Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances* 1(5), pp. e1400253. doi: 10.1126/sciadv.1400253

Ceballos, G., Ehrlich, P.R. and Dirzo, R. 2017. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences* 114(30), pp. E6089–E6096. doi: 10.1073/pnas.1704949114

Chain, F.J.J., Brown, E.A., MacIsaac, H.J. and Cristescu, M.E. 2016. Metabarcoding reveals strong spatial structure and temporal turnover of zooplankton communities among marine and freshwater ports. *Diversity and Distributions* 22(5), pp. 493–504. doi: 10.1111/ddi.12427

Chan, K.S., Tan, J., Goh, W.L., and Earl of Cranbrook 2019. Diet profiling of house-farm swiftlets (Aves, Apodidae, Aerodramus sp.) in three landscapes in Perak, Malaysia, using high-throughput sequencing. *Tropical Ecology* 60(3), pp. 379–388. doi: 10.1007/s42965-019-00040-1

Cheng, Y.-C. and Lin, C.-P. 2016. Dietary Niche Partitioning of *Euphaea formosa* and *Matrona cyanoptera* (Odonata: Zygoptera) on the Basis of DNA Barcoding of Larval Feces. *Journal of Insect Science* 16(1), pp. 73. doi: 10.1093/jisesa/iew060

Christjani, M. and Von Elert, E. 2015. Prey-induced vertical migration in *Chaoborus* larvae under different predator and light regimes. *Journal of Plankton Research* 37(1), pp. 48–55. doi: 10.1093/plankt/fbu086

Clare, E.L. 2014. Molecular detection of trophic interactions: emerging trends, distinct advantages, significant considerations and conservation applications. *Evolutionary Applications* 7(9), pp. 1144–1157. doi: 10.1111/eva.12225

Clare, E.L., Barber, B.R., Sweeney, B.W., Hebert, P.D.N. and Fenton, M.B. 2011. Eating local: influences of habitat on the diet of little brown bats (*Myotis lucifugus*). *Molecular Ecology* 20(8), pp. 1772–1780. doi: 10.1111/j.1365-294X.2011.05040.x

Clare, E.L., Fraser, E.E., Braid, H.E., Fenton, M.B. and Hebert, P.D.N. 2009. Species on the menu of a generalist predator, the eastern red bat (*Lasiurus borealis*): using a molecular approach to detect arthropod prey. *Molecular Ecology* 18(11), pp. 2532–2542. doi: 10.1111/j.1365-294X.2009.04184.x

Clarke, L.J., Beard, J.M., Swadling, K.M. and Deagle, B.E. 2017. Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution* 7(3), pp. 873–883. doi: 10.1002/ece3.2667

Clarke, L.J., Soubrier, J., Weyrich, L.S. and Cooper, A. 2014. Environmental metabarcodes for insects: *in silico* PCR reveals potential for taxonomic bias. *Molecular Ecology Resources* 14(6), pp. 1160–1170. doi: 10.1111/1755-0998.12265

Coissac, E., Riaz, T. and Puillandre, N. 2013. Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology* 21(8), pp. 1834–1847. doi: 10.1111/j.1365-294X.2012.05550.x

Collins, R.A. and Cruickshank, R.H. 2013. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, pp. 969-975. doi: 10.1111/1755-0998.12046

Corse, E. et al. 2017. A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies. *Molecular Ecology Resources* 17(6), pp. e146–e159. doi: 10.1111/1755-0998.12703

Corse, E., Costedoat, C., Chappaz, R., Pech, N., Martin, J. and Gilles, A. 2010. A PCR-based method for diet analysis in freshwater organisms using 18S rDNA barcoding on faeces. *Molecular Ecology Resources* 10(1), pp. 96–108. doi: 10.1111/j.1755-0998.2009.02795.x

Cristescu, M.E. and Hebert, P.D.N. 2018. Uses and Misuses of Environmental DNA in Biodiversity Science and Conservation. *Annual Review of Ecology, Evolution, and Systematics* 49(1), pp. 209–230. doi: 10.1146/annurev-ecolsys-110617-062306

Darwall, W. et al. 2018. The *Alliance for Freshwater Life* : A global call to unite efforts for freshwater biodiversity science and conservation. *Aquatic Conservation: Marine and Freshwater Ecosystems* 28(4), pp. 1015–1022. doi: 10.1002/aqc.2958

Darwall, W.R.T. et al. 2011. Implications of bias in conservation research and investment for freshwater species: Conservation and freshwater species. *Conservation Letters* 4(6), pp. 474–482. doi: 10.1111/j.1755-263X.2011.00202.x

de Vargas, C. et al. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348(6237), pp. 1261605. doi: 10.1126/science.1261605

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. and Taberlet, P. 2014. DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match. *Biology Letters* 10(9), pp. 20140562. doi: 10.1098/rsbl.2014.0562

Defra, 2018. *25 Year Environment Plan*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/693158/25-year-environment-plan.pdf

Deiner, K. and Altermatt, F. 2014. Transport Distance of Invertebrate Environmental DNA in a Natural River. *PLoS ONE* 9(2), pp. e88786. doi: 10.1371/journal.pone.0088786

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P. and Miaud, C. 2011. Persistence of Environmental DNA in Freshwater Ecosystems. Gilbert, J. A. ed. *PLoS ONE* 6(8), pp. e23398. doi: 10.1371/journal.pone.0023398

Di Marco, M. et al. 2017. Changing trends and persisting biases in three decades of conservation science. *Global Ecology and Conservation* 10, pp. 32–42. doi: 10.1016/j.gecco.2017.01.008

Di Muri, C. et al. 2020. Read counts from environmental DNA (eDNA) metabarcoding reflect fish abundance and biomass in drained ponds. *Metabarcoding and Metagenomics* 4, pp. e56959. doi: 10.3897/mbmg.4.56959

Dieffenbach, C.W., Lowe, T.M. and Dveksler, G.S. 1993. General concepts for PCR primer design. *Genome Research* 3(3), pp. S30–S37. doi: 10.1101/gr.3.3.S30

Ducotterd, C., Crovadore, J., Lefort, F., Rubin, J. and Ursenbacher, S. 2021. A powerful long metabarcoding method for the determination of complex diets from faecal analysis of the European pond turtle (*Emys orbicularis*, L. 1758). *Molecular Ecology Resources* 21(2), pp. 433–447. doi: 10.1111/1755-0998.13277

Dudgeon, D. et al. 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews* 81(02), pp. 163. doi: 10.1017/S1464793105006950

Elbrecht, V. et al. 2016. Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ* 4, pp. e1966. doi: 10.7717/peerj.1966

Elbrecht, V., Hebert, P.D.N. and Steinke, D. 2018. Slippage of degenerate primers can cause variation in amplicon length. *Scientific Reports* 8(1), pp. 10999. doi: 10.1038/s41598-018-29364-z

Elbrecht, V. and Leese, F. 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PLOS ONE* 10(7), pp. e0130324. doi: 10.1371/journal.pone.0130324

Elbrecht, V. and Leese, F. 2017. Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science* 5, pp. 11.

Elbrecht, V., Vamos, E.E., Meissner, K., Aroviita, J. and Leese, F. 2017. Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution* 8(10), pp. 1265–1275. doi: 10.1111/2041-210X.12789

Estes, J.A. et al. 2011. Trophic Downgrading of Planet Earth. *Science* 333(6040), pp. 301–306. doi: 10.1126/science.1205106

European Commission, 2016. *Introduction to the new EU Water Framework Directive*. Available at: https://ec.europa.eu/environment/water/water-framework/info/intro_en.htm

Evans, D.M., Kitson, J.J.N., Lunt, D.H., Straw, N.A. and Pocock, M.J.O. 2016. Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. *Functional Ecology* 30(12), pp. 1904–1916. doi: 10.1111/1365-2435.12659

Fernando, C.H. 1994. Zooplankton, fish and fisheries in tropical freshwaters. In *Studies on the Ecology of Tropical Zooplankton*, pp. 105-123. Springer, Dordrecht.

Fontes, J.T., Vieira, P.E., Ekrem, T., Soares, P. and Costa, F.O. 2021. BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries. *Molecular Ecology Resources* 21(2), pp. 573–583. doi: 10.1111/1755-0998.13262

Folmer, O., Hoeh, W.R., Black, M.B. and Vrijenhoek, R.C., 1994. Conserved primers for PCR amplification of mitochondrial DNA from different invertebrate phyla. *Molecular Marine Biology and Biotechnology* 3(5), pp. 294-299.

Gamboa, M., Kimbirauskas, R.K., Merritt, R.W. and Monaghan, M.T. 2012. A Molecular Approach to Identifying the Natural Prey of the African Creeping Water Bug *Naucoris*, A Potential Reservoir of *Mycobacterium ulcerans*. *Journal of Insect Science* 12(2), pp. 1–10. doi: 10.1673/031.012.0201

GBIF Secretariat, 2020. *GBIF Backbone Taxonomy*. Available at: https://doi.org/10.15468/39omei Accessed via https://www.gbif.org/species/5284517

Gerwing, T.G., Kim, J.-H., Hamilton, D.J., Barbeau, M.A. and Addison, J.A. 2016. Diet reconstruction using next-generation sequencing increases the known ecosystem usage by a shorebird. *The Auk* 133(2), pp. 168–177. doi: 10.1642/AUK-15-176.1

Gillet, F., Tiouchichine, M.-L., Galan, M., Blanc, F., Némoz, M., Aulagnier, S. and Michaux, J.R. 2015. A new method to identify the endangered Pyrenean desman (*Galemys pyrenaicus*) and to study its diet, using next generation sequencing from faeces. *Mammalian Biology* 80(6), pp. 505–509. doi: 10.1016/j.mambio.2015.08.002

Goldstein, P.Z., Desalle, R., Amato, G. and Vogler, A.P. 2000. Conservation Genetics at the Species Boundary. *Conservation Biology* 14(1), pp. 120–131. doi: 10.1046/j.1523-1739.2000.98122.x

Grooten, M., Almond, R.E.A. and WWF (Organization) eds. 2018. *Living planet report 2018: aiming higher*. Gland, Switzerland: WWF--World Wide Fund for Nature.

Guillerault, N., Bouletreau, S., Iribar, A., Valentini, A. and Santoul, F. 2017. Application of DNA metabarcoding on faeces to identify European catfish *Silurus glanis* diet. *Journal of Fish Biology* 90(5), pp. 2214–2219. doi: 10.1111/jfb.13294

Gunn, I.D.M. et al. 2018. UK Checklist of freshwater species. NERC Environmental Information Data Centre. (Dataset). https://doi.org/10.5285/57653719-434b-4b11-9f0d-3bd76054d8bd

Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W. and Hebert, P.D.N. 2006. DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences* 103(4), pp. 968–971. doi: 10.1073/pnas.0510466103

Hardy, C.M., Adams, M., Jerry, D.R., Court, L.N., Morgan, M.J. and Hartley, D.M. 2011. DNA barcoding to support conservation: species identification, genetic structure and biogeography of fishes in the Murray - Darling River Basin, Australia. *Marine and Freshwater Research* 62(8), pp. 887. doi: 10.1071/MF11027

Harper, L.R. et al. 2020. Assessing the impact of the threatened crucian carp (*Carassius carassius*) on pond invertebrate diversity: A comparison of conventional and molecular tools. *Molecular Ecology* 30(13), pp. 3252-3269. doi: 10.1111/mec.15670

Hawlitschek, O., Fernández-González, A., Balmori-de la Puente, A. and Castresana, J. 2018. A pipeline for metabarcoding and diet analysis from fecal samples developed for a small semi-aquatic mammal. *PLOS ONE* 13(8), pp. e0201763. doi: 10.1371/journal.pone.0201763

Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. 2003a. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270(1512), pp. 313–321. doi: 10.1098/rspb.2002.2218

Hebert, P.D.N., Ratnasingham, S. and de Waard, J.R. 2003b. Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270, pp. S96-S99.

Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S. and Francis, C.M. 2004. Identification of Birds through DNA Barcodes. *PLoS Biology* 2(10), pp. e312. doi: 10.1371/journal.pbio.0020312

Hopkins, G.W. and Freckleton, R.P. 2002. Declines in the numbers of amateur and professional taxonomists: implications for conservation. *Animal Conservation* 5(3), pp. 245–249. doi: 10.1017/S1367943002002299

Huemer, P., Mutanen, M., Sefc, K.M. and Hebert, P.D., 2014. Testing DNA barcode performance in 1000 species of European Lepidoptera: large geographic distances have small genetic impacts. *Plos one*, *9*(12), doi: 10.1371/journal.pone.0115774

International Barcode of Life Consortium, 2022. International Barcode of Life Project (iBOL). Available at: https://ibol.org/

Jäger, I.S., Hölker, F., Flöder, S. and Walz, N. 2011. Impact of *Chaoborus flavicans* - Predation on the Zooplankton in a Mesotrophic Lake - a Three Year Study. *International Review of Hydrobiology* 96(2), pp. 191–208. doi: 10.1002/iroh.201011253

Janzen, D.H., 1974. The deflowering of central America. Natural History 83, pp. 48-53.

Jeppesen, E. et al. 2011. Zooplankton as indicators in lakes: a scientific-based plea for including zooplankton in the ecological quality assessment of lakes according to the European Water Framework Directive (WFD). *Hydrobiologia* 676(1), pp. 279–297. doi: 10.1007/s10750-011-0831-0

JNCC, 2020. *Conservation Designations for UK Taxa*. Available at: https://hub.jncc.gov.uk/assets/478f7160-967b-4366-acdf-8941fd33850b

Jo, H. et al. 2016. Discovering hidden biodiversity: the use of complementary monitoring of fish diet based on DNA barcoding in freshwater ecosystems. *Ecology and Evolution* 6(1), pp. 219–232. doi: 10.1002/ece3.1825

Jo, H. et al. 2019. Responses of fish assemblage structure to large-scale weir construction in riverine ecosystems. *Science of The Total Environment* 657, pp. 1334–1342. doi: 10.1016/j.scitotenv.2018.11.446

Jo, H., Gim, J., Jeong, K., Kim, H. and Joo, G. 2014. Application of DNA barcoding for identification of freshwater carnivorous fish diets: Is number of prey items dependent on size class for *Micropterus salmoides*? *Ecology and Evolution* 4(2), pp. 219–229. doi: 10.1002/ece3.921

Katoh, K. and Standley, D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30(4), pp. 772–780. doi: 10.1093/molbev/mst010

Kelling, C.J., Isermann, D.A., Sloss, B.L. and Turnquist, K.N. 2016. Diet Overlap and Predation Between Largemouth Bass and Walleye in Wisconsin Lakes Using DNA Barcoding to Improve Taxonomic Resolution. *North American Journal of Fisheries Management* 36(3), pp. 621–629. doi: 10.1080/02755947.2016.1146179

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16(2), pp. 111–120. doi: 10.1007/BF01731581

King, R.A., Read, D.S., Traugott, M. and Symondson, W.O.C. 2008. Invited review: Molecular analysis of predation: a review of best practice for DNA-based approaches. *Molecular Ecology* 17(4), pp. 947–963. doi: 10.1111/j.1365-294X.2007.03613.x

Koroiva, R. and Kvist, S. 2018. Estimating the barcoding gap in a global dataset of *cox1* sequences for Odonata: close, but no cigar. *Mitochondrial DNA Part A* 29(5), pp. 765–771. doi: 10.1080/24701394.2017.1357709

Kovac, R., Boileau, N., Muschick, M. and Salzburger, W. 2019. The diverse prey spectrum of the Tanganyikan scale-eater *Perissodus microlepis* (Boulenger, 1898). *Hydrobiologia* 832(1), pp. 85–92. doi: 10.1007/s10750-018-3714-9

Kumari, P., Dong, K., Eo, K.Y., Lee, W.-S., Kimura, J. and Yamamoto, N. 2019. DNA metabarcoding-based diet survey for the Eurasian otter (*Lutra lutra*): Development of a Eurasian otter-specific blocking oligonucleotide for 12S rRNA gene sequencing for vertebrates. *PLOS ONE* 14(12), pp. e0226253. doi: 10.1371/journal.pone.0226253

Lawson Handley, L. 2015. How will the 'molecular revolution' contribute to biological recording? *Biological Journal of the Linnean Society* 115(3), pp. 750–766. doi: 10.1111/bij.12516

Lawson Handley, L. et al. 2019. Temporal and spatial variation in distribution of fish environmental DNA in England's largest lake. *Environmental DNA* 1(1), pp. 26–39. doi: 10.1002/edn3.5

Leese, F., Sander, M., Buchner, D., Elbrecht, V., Haase, P. and Zizka, V.M.A. 2021. Improved freshwater macroinvertebrate detection from environmental DNA through minimized nontarget amplification. *Environmental DNA* 3(1), pp. 261–276. doi: 10.1002/edn3.177

Leray, M. et al. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology* 10(1), pp. 34. doi: 10.1186/1742-9994-10-34

Leray, M. and Knowlton, N. 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences* 112(7), pp. 2076–2081. doi: 10.1073/pnas.1424997112

Leray, M. and Knowlton, N. 2017. Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ* 5, pp. e3006. doi: 10.7717/peerj.3006

Lindeque, P.K., Parry, H.E., Harmer, R.A., Somerfield, P.J. and Atkinson, A. 2013. Next Generation Sequencing Reveals the Hidden Diversity of Zooplankton Assemblages. *PLoS ONE* 8(11), pp. e81327. doi: 10.1371/journal.pone.0081327

Lu, Y.-T., Liu, M.-Y., He, Y. and Liao, T.-Y. 2016. *Smilosicyopus leprurus* (Teleostei: Gobiidae) is a Fin-eater. *Zoological Studies* 55, pp. 7.

Luo, M., Ji, Y. and Yu, D.W. 2022. Extracting abundance information from DNA-based data. *bioRxiv*, pp. 40. Preprint. doi.org/10.1101/2022.01.06.475221

Maberly, S.C. et al. 2016. A survey of the status of the lakes of the English Lake District: The Lakes Tour 2015. NERC/Centre for Ecology & Hydrology. Available at: http://nora.nerc.ac.uk/id/eprint/513514/

Macheriotou, L. et al. 2019. Metabarcoding free-living marine nematodes using curated 18S and COI reference sequence databases for species-level taxonomic assignments. *Ecology and Evolution* 9(3), pp. 1211-1226.

McCann, K. 2007. Protecting biostructure. *Nature* 446(7131), pp. 29–29. doi: 10.1038/446029a.

McMeans, B.C. et al. 2016. The adaptive capacity of lake food webs: from individuals to ecosystems. *Ecological Monographs* 86(1), pp. 4–19. doi: 10.1890/15-0288.1

Meier, R., Zhang, G. and Ali, F. 2008. The Use of Mean Instead of Smallest Interspecific Distances Exaggerates the Size of the "Barcoding Gap" and Leads to Misidentification. *Systematic Biology* 57(5), pp. 809–813. doi: 10.1080/10635150802406343

Meyer, J.M., Leempoel, K., Losapio, G. and Hadly, E.A. 2020. Molecular Ecological Network Analyses: An Effective Conservation Tool for the Assessment of Biodiversity, Trophic Interactions, and Community Structure. *Frontiers in Ecology and Evolution* 8, pp. 588430. doi: 10.3389/fevo.2020.588430

Michel, C.J., Smith, J.M., Demetras, N.J., Huff, D.D. and Hayes, S.A., 2018. Non-native fish predator density and molecular-based diet estimates suggest differing impacts of predator species on juvenile salmon in the San Joaquin River, California. *San Francisco Estuary and Watershed Science*, *16*(4), pp. 1-19. doi: 10.15447/sfews.2018v16iss4art3

Moore, M.V. 1988. Differential use of food resources by the instars of *Chaoborus punctipennis*. *Freshwater Biology* 19(2), pp. 249–268. doi: 10.1111/j.1365-2427.1988.tb00346.x

Moran, Z., Orth, D.J., Schmitt, J.D., Hallerman, E.M. and Aguilar, R. 2016. Effectiveness of DNA barcoding for identifying piscine prey items in stomach contents of piscivorous catfishes. *Environmental Biology of Fishes* 99(1), pp. 161–167. doi: 10.1007/s10641-015-0448-7

Moss, B., 2010. *Ecology of fresh waters: a view for the twenty-first century*. 4th ed. John Wiley & Sons.

Nelson, E.J.H., Holden, J., Eves, R. and Tufts, B. 2017. Comparison of diets for Largemouth and Smallmouth Bass in Eastern Lake Ontario using DNA barcoding and stable isotope analysis. *PLOS ONE* 12(8), pp. e0181914. doi: 10.1371/journal.pone.0181914

Nielsen, J.M., Clare, E.L., Hayden, B., Brett, M.T. and Kratina, P. 2018. Diet tracing in ecology: Method comparison and selection. *Methods in Ecology and Evolution* 9(2), pp. 278–291. doi: 10.1111/2041-210X.12869

Northam, W.T., Allison, L.A. and Cristol, D.A. 2012. Using group-specific PCR to detect predation of mayflies (Ephemeroptera) by wolf spiders (Lycosidae) at a mercury-contaminated site. *Science of The Total Environment* 416, pp. 225–231. doi: 10.1016/j.scitotenv.2011.11.083

Oehm, J., Thalinger, B., Eisenkölbl, S. and Traugott, M. 2017. Diet analysis in piscivorous birds: What can the addition of molecular tools offer? *Ecology and Evolution* 7(6), pp. 1984–1995. doi: 10.1002/ece3.2790

Ohlberger, J., Thackeray, S.J., Winfield, I.J., Maberly, S.C. and Vøllestad, L.A. 2014. When phenology matters: age–size truncation alters population response to trophic mismatch. *Proceedings of the Royal Society B: Biological Sciences* 281(1793), pp. 20140938. doi: 10.1098/rspb.2014.0938

Oksanen, J. et al. 2020. Vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan

Oliveira, L.M., Knebelsberger, T., Landi, M., Soares, P., Raupach, M.J. and Costa, F.O. 2016. Assembling and auditing a comprehensive DNA barcode reference library for European marine fishes: DNA barcode library for European marine fishes. *Journal of Fish Biology* 89(6), pp. 2741–2754. doi: 10.1111/jfb.13169

Pearman, J.K., El-Sherbiny, M.M., Lanzen, A., Al-Aidaroos, A.M. and Irigoien, X. 2014. Zooplankton diversity across three Red Sea reefs using pyrosequencing. *Frontiers in Marine Science* 1, pp. 27.

Pearson, C.E., Symondson, W.O.C., Clare, E.L., Ormerod, S.J., Iparraguirre Bolaños, E. and Vaughan, I.P. 2018. The effects of pastoral intensification on the feeding interactions of generalist predators in streams. *Molecular Ecology* 27(2), pp. 590–602. doi: 10.1111/mec.14459

Pereira-da-Conceicoa, L., Elbrecht, V., Hall, A., Briscoe, A., Barber-James, H. and Price, B. 2021. Metabarcoding unsorted kick-samples facilitates macroinvertebrate-based biomonitoring with increased taxonomic resolution, while outperforming environmental DNA. *Environmental DNA* 3(2), pp. 353–371. doi: 10.1002/edn3.116

Pimm, S.L., Russell, G.J., Gittleman, J.L. and Brooks, T.M. 1995. The Future of Biodiversity. *Science* 269(5222), pp. 347–350. doi: 10.1126/science.269.5222.347

Piñol, J., San Andrés, V., Clare, E.L., Mir, G. and Symondson, W.O.C. 2014. A pragmatic approach to the analysis of diets of generalist predators: the use of next-generation sequencing with no blocking probes. *Molecular Ecology Resources* 14(1), pp. 18–26. doi: 10.1111/1755-0998.12156

Piñol, J., Mir, G., Gomez-Polo, P. and Agustí, N., 2015. Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular ecology resources* 15(4), pp. 819-830.

Piñol, J., Senar, M.A. and Symondson, W.O.C. 2019. The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology* 28(2), pp. 407–419. doi: 10.1111/mec.14776

Pompanon, F., Deagle, B.E., Symondson, W.O.C., Brown, D.S., Jarman, S.N. and Taberlet, P. 2012. Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology* 21(8), pp. 1931–1950. doi: 10.1111/j.1365-294X.2011.05403.x

Porazinska, D.L., Giblin-Davis, R.M., Sung, W. and Thomas, W.K. 2010. Linking operational clustered taxonomic units (OCTUs) from parallel ultra sequencing (PUS) to nematode species. *Zootaxa* 2427(1), pp. 55. doi: 10.11646/zootaxa.2427.1.6

Porter, T.M. and Hajibabaei, M. 2020a. METAWORKS: A flexible, scalable bioinformatic pipeline for multi-marker biodiversity assessments. *Bioinformatics*.

Porter, T.M. and Hajibabaei, M. 2020b. Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. *Frontiers in Ecology and Evolution* 8, pp. 248. doi: 10.3389/fevo.2020.00248

Price, B. 2018. FreshBase: *A genomic voucher collection of UK freshwater macroinvertebrates*. Available at: https://freshbase.myspecies.info/

Price, B. 2020. *The UK Barcode of Life Project*. Available at: https://ukbol.org/

Prosser, S., Martínez-Arce, A. and Elías-Gutiérrez, M. 2013. A new set of primers for COI amplification from freshwater microcrustaceans. *Molecular Ecology Resources*, pp.1151-1155. doi: 10.1111/1755-0998.12132

R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raper C. 2022. *United Kingdom Species Inventory (UKSI)*. Version 37.9. Natural History Museum. Available at: https://www.nhm.ac.uk/our-science/data/uk-species.html

Ratnasingham, S. and Hebert, P.D.N. 2007. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes* 7(3), pp. 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Ratnasingham, S. and Hebert, P.D.N. 2013. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE* 8(7), pp. e66213. doi: 10.1371/journal.pone.0066213

Rees, H.C., Bishop, K., Middleditch, D.J., Patmore, J.R.M., Maddison, B.C. and Gough, K.C. 2014. The application of eDNA for monitoring of the Great Crested Newt in the UK. *Ecology and Evolution* 4(21), pp. 4023–4032. doi: 10.1002/ece3.1272

Reid, A.J. et al. 2019. Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews* 94(3), pp. 849–873. doi: 10.1111/brv.12480

Rennstam Rubbmark, O., Sint, D., Cupic, S. and Traugott, M. 2019. When to use next generation sequencing or diagnostic PCR in diet analyses. *Molecular Ecology Resources* 19(2), pp. 388–399. doi: 10.1111/1755-0998.12974

Robinson, E., Blagoev, G., Hebert, P. and Adamowicz, S. 2009. Prospects for using DNA barcoding to identify spiders in species-rich genera. *ZooKeys* 16, pp. 27–46. doi: 10.3897/zookeys.16.239

Roslin, T. and Majaneva, S. 2016. The use of DNA barcodes in food web construction—terrestrial and aquatic ecologists unite! *Genome* 59(9), pp. 603–628. doi: 10.1139/gen-2015-0229

Roy, H.E. et al. 2014. Horizon scanning for invasive alien species with the potential to threaten biodiversity in Great Britain. *Global Change Biology* 20(12), pp. 3859–3871. doi: 10.1111/gcb.12603

Sansom, B.J. and Sassoubre, L.M. 2017. Environmental DNA (eDNA) Shedding and Decay Rates to Model Freshwater Mussel eDNA Transport in a River. *Environmental Science & Technology* 51(24), pp. 14244–14253. doi: 10.1021/acs.est.7b05199

Schmitt, J.D., Hallerman, E.M., Bunch, A., Moran, Z., Emmel, J.A. and Orth, D.J. 2017. Predation and Prey Selectivity by Non-native Catfish on Migrating Alosines in an Atlantic Slope Estuary. *Marine and Coastal Fisheries* 9(1), pp. 108–125. doi: 10.1080/19425120.2016.1271844

Schmitt, J.D., Peoples, B.K., Castello, L. and Orth, D.J. 2019. Feeding ecology of generalist consumers: a case study of invasive blue catfish *Ictalurus furcatus* in Chesapeake Bay, Virginia, USA. *Environmental Biology of Fishes* 102(3), pp. 443–465. doi: 10.1007/s10641-018-0783-6

Schnell, I.B. et al. 2012. Screening mammal biodiversity using DNA from leeches. *Current Biology* 22(8), pp. R262–R263. doi: 10.1016/j.cub.2012.02.058

Sint, D., Raso, L., Kaufmann, R. and Traugott, M. 2011. Optimizing methods for PCR-based analysis of predation. *Molecular Ecology Resources* 11(5), pp. 795–801. doi: 10.1111/j.1755-0998.2011.03018.x

Smith, M.A., Eveleigh, E.S., McCann, K.S., Merilo, M.T., McCarthy, P.C. and Van Rooyen, K.I. 2011. Barcoding a Quantified Food Web: Crypsis, Concepts, Ecology and Hypotheses. *PLoS ONE* 6(7), pp. e14424. doi: 10.1371/journal.pone.0014424

Staudinger, M.D. et al. 2021. How Does Climate Change Affect Emergent Properties of Aquatic Ecosystems? *Fisheries* 46(9), pp. 423–441. doi: 10.1002/fsh.10606

Stouffer, D.B. 2010. Scaling from individuals to networks in food webs. *Functional Ecology* 24(1), pp. 44–51. doi: 10.1111/j.1365-2435.2009.01644.x

Symondson, W.O.C. 2002. Molecular identification of prey in predator diets. *Molecular Ecology* 11(4), pp. 627–641. doi: 10.1046/j.1365-294X.2002.01471.x

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. and Willerslev, E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* 21(8), pp. 2045–2050. doi: 10.1111/j.1365-294X.2012.05470.x

Taguchi, T., Miura, Y., Krueger, D. and Sugiura, S. 2014. Utilizing stomach content and faecal DNA analysis techniques to assess the feeding behaviour of largemouth bass *Micropterus salmoides* and bluegill *Lepomis macrochirus*: stomach content and faecal dna analysis. *Journal of Fish Biology* 84(5), pp. 1271–1288. doi: 10.1111/jfb.12341.

Tang, K.W., Flury, S., Vachon, D., Ordóñez, C. and McGinnis, D.F. 2018. The phantom midge menace: Migratory *Chaoborus* larvae maintain poor ecosystem state in eutrophic inland waters. *Water Research* 139, pp. 30–37. doi: 10.1016/j.watres.2018.03.060.

Thackeray, S.J., Henrys, P.A., Feuchtmayr, H., Jones, I.D., Maberly, S.C. and Winfield, I.J., 2013. Food web de-synchronization in England's largest lake: an assessment based on multiple phenological metrics. *Global Change Biology* 19(12), pp. 3568-3580

Thalinger, B., Oehm, J., Mayr, H., Obwexer, A., Zeisler, C. and Traugott, M. 2016. Molecular prey identification in Central European piscivores. *Molecular Ecology Resources* 16(1), pp. 123–137. doi: 10.1111/1755-0998.12436

Thalinger, B., Oehm, J., Zeisler, C., Vorhauser, J. and Traugott, M. 2018. Sex-specific prey partitioning in breeding piscivorous birds examined via a novel, noninvasive approach. *Ecology and Evolution* 8(17), pp. 8985–8998. doi: 10.1002/ece3.4421

Thierry, A., Beckerman, A.P., Warren, P.H., Williams, R.J., Cole, A.J. and Petchey, O.L. 2011. Adaptive foraging and the rewiring of size-structured food webs following extinctions. *Basic and Applied Ecology* 12(7), pp. 562–570. doi: 10.1016/j.baae.2011.09.005

Thompson, R.M. et al. 2012a. Food webs: reconciling the structure and function of biodiversity. *Trends in Ecology & Evolution* 27(12), pp. 689–697. doi: 10.1016/j.tree.2012.08.005

Thompson, R.M., Dunne, J.A. and Woodward, G. 2012b. Freshwater food webs: towards a more fundamental understanding of biodiversity and community dynamics: Freshwater food webs - a review. *Freshwater Biology* 57(7), pp. 1329–1341. doi: 10.1111/j.1365-2427.2012.02808.x

Thomsen, P.F. et al. 2012. Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology* 21(11), pp. 2565–2573. doi: 10.1111/j.1365-294X.2011.05418.x

Thomsen, P.F. and Willerslev, E. 2015. Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation* 183, pp. 4–18. doi: 10.1016/j.biocon.2014.11.019

Tickner, D. et al. 2020. Bending the Curve of Global Freshwater Biodiversity Loss: An Emergency Recovery Plan. *BioScience* 70(4), pp. 330–342. doi: 10.1093/biosci/biaa002

Tréguier, A., Paillisson, J.-M., Dejean, T., Valentini, A., Schlaepfer, M.A. and Roussel, J.-M. 2014. Environmental DNA surveillance for invertebrate species: advantages and technical limitations to detect invasive crayfish *Procambarus clarkii* in freshwater ponds. *Journal of Applied Ecology* 51(4), pp. 871–879. doi: 10.1111/1365-2664.12262

Trevelline, B.K., Latta, S.C., Marshall, L.C., Nuttle, T. and Porter, B.A. 2016. Molecular analysis of nestling diet in a long-distance Neotropical migrant, the Louisiana Waterthrush (*Parkesia motacilla*). *The Auk* 133(3), pp. 415–428. doi: 10.1642/AUK-15-222.1

Trevelline, B.K., Nuttle, T., Hoenig, B.D., Brouwer, N.L., Porter, B.A. and Latta, S.C. 2018a. DNA metabarcoding of nestling feces reveals provisioning of aquatic prey and resource partitioning among Neotropical migratory songbirds in a riparian habitat. *Oecologia* 187(1), pp. 85–98. doi: 10.1007/s00442-018-4136-0

Trevelline, B.K., Nuttle, T., Porter, B.A., Brouwer, N.L., Hoenig, B.D., Steffensmeier, Z.D. and Latta, S.C. 2018b. Stream acidification and reduced aquatic prey availability are associated with dietary shifts in an obligate riparian Neotropical migratory songbird. *PeerJ* 6, pp. e5141. doi: 10.7717/peerj.5141

Tylianakis, J.M., Didham, R.K., Bascompte, J. and Wardle, D.A. 2008. Global change and species interactions in terrestrial ecosystems. *Ecology Letters* 11(12), pp. 1351–1363. doi: 10.1111/j.1461-0248.2008.01250.x

Valiente-Banuet, A. et al. 2015. Beyond species loss: the extinction of ecological interactions in a changing world. *Functional Ecology* 29(3), pp. 299-307

Velarde, E., Ezcurra, E. and Anderson, D.W. 2013. Seabird diets provide early warning of sardine fishery declines in the Gulf of California. *Scientific Reports* 3(1), pp. 1-6. doi: 10.1038/srep01332

Vesterinen, E.J., Lilley, T., Laine, V.N. and Wahlberg, N. 2013. Next Generation Sequencing of Fecal DNA Reveals the Dietary Diversity of the Widespread Insectivorous Predator Daubenton's Bat (*Myotis daubentonii*) in Southwestern Finland. *PLoS ONE* 8(11), pp. e82168. doi: 10.1371/journal.pone.0082168

Virgilio, M., Backeljau, T., Nevado, B. and Meyer, M.D. 2010. Comparative performances of DNA barcoding across insect orders. *BMC bioinformatics*, 11(1), pp.1-10.

Voss, S. and Mumm, H., 1999. Where to stay by night and day: Size-specific and seasonal differences in horizontal and vertical distribution of *Chaoborus flavicans* larvae. *Freshwater biology* 42(2), pp. 201-213.

Waraniak, J.M., Baker, E.A. and Scribner, K.T. 2018a. Molecular diet analysis reveals predator-prey community dynamics and environmental factors affecting predation of larval lake sturgeon *Acipenser fulvescens* in a natural system. *Journal of Fish Biology* 93(4), pp. 616–629. doi: 10.1111/jfb.13726

Waraniak, J.M., Blumstein, D.M. and Scribner, K.T. 2018b. Barcoding PCR primers detect larval lake sturgeon (*Acipenser fulvescens*) in diets of piscine predators. *Conservation Genetics Resources* 10(2), pp. 259–268. doi: 10.1007/s12686-017-0790-5

Waraniak, J.M., Marsh, T.L. and Scribner, K.T. 2019. 18S rRNA metabarcoding diet analysis of a predatory fish community across seasonal changes in prey availability. *Ecology and Evolution* 9(3), pp. 1410–1430. doi: 10.1002/ece3.4857

Ward, R.D., Hanner, R. and Hebert, P.D., 2009. The campaign to DNA barcode all fishes, FISH-BOL. *Journal of fish biology* 74(2), pp. 329-356.

Weigand, H. et al. 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of The Total Environment* 678, pp. 499–524. doi: 10.1016/j.scitotenv.2019.04.247

Weisser, M., Hofmann, H., Fernández, J.E. and Peeters, F. 2018. Vertical migration patterns of the different larval instars of *Chaoborus flavicans* and the influence of dissolved oxygen concentrations. *Canadian Journal of Fisheries and Aquatic Sciences* 75(7), pp. 1142–1150. doi: 10.1139/cjfas-2017-0157

Wiemers, M. and Fiedler, K. 2007. Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology* 4(1), pp. 8. doi: 10.1186/1742-9994-4-8

Winder, M. and Schindler, D.E. 2004. Climate change uncouples trophic interactions in an aquatic ecosystem. *Ecology* 85(8), pp. 2100–2106. doi: 10.1890/04-0151

Wirta, H.K., Hebert, P.D.N., Kaartinen, R., Prosser, S.W., Várkonyi, G. and Roslin, T. 2014. Complementary molecular information changes our perception of food web structure. *Proceedings of the National Academy of Sciences* 111(5), pp. 1885–1890. doi: 10.1073/pnas.1316990111

Woodward, G. 2009. Biodiversity, ecosystem functioning and food webs in fresh waters: assembling the jigsaw puzzle. *Freshwater Biology* 54(10), pp. 2171–2187. doi: 10.1111/j.1365-2427.2008.02081.x

Woodward, G. et al. 2010. Individual-Based Food Webs: species identity, body size and sampling effects. *Advances in Ecological Research* 43, pp. 211-266 Elsevier.

Woodward, G., Ebenman, B., Emmerson, M., Montoya, J., Olesen, J., Valido, A. and Warren, P. 2005. Body size in ecological networks. *Trends in Ecology & Evolution* 20(7), pp. 402–409. doi: 10.1016/j.tree.2005.04.005

Yang, J. et al. 2017a. Indigenous species barcode database improves the identification of zooplankton. *PLOS ONE* 12(10), pp. e0185697. doi: 10.1371/journal.pone.0185697

Yang, J., Zhang, X., Xie, Y., Song, C., Zhang, Y., Yu, H. and Burton, G.A. 2017b. Zooplankton Community Profiling in a Eutrophic Freshwater Ecosystem-Lake Tai Basin by DNA Metabarcoding. *Scientific Reports* 7(1), pp. 1773. doi: 10.1038/s41598-017-01808-y

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. and Ding, Z., 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), pp.613-623. doi: 10.1111/ j.2041-210X.2012.00198.xZhan, A., Bailey, S.A., Heath, D.D. and Macisaac, H.J. 2014. Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities. *Molecular Ecology Resources* 14(5), pp. 1049-1059. doi: 10.1111/1755-0998.12254

Zhang, G.K., Chain, F.J.J., Abbott, C.L. and Cristescu, M.E. 2018. Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evolutionary Applications* 11(10), pp. 1901–1914. doi: 10.1111/eva.12694

Supplementary materials

Appendix S1: Quick scoping review methods

Evidence was gathered using Web of Science and Scopus literature searches, conducted on the 20th May 2019 and updated on the 28th July 2021. The following search terms were used for each search (nested terms and wildcards (*) were used to pick up alternative terms and multiple word endings; no date restrictions were applied):

(freshwater OR lake* OR river* OR stream OR streams OR wetland* OR pond* OR "inland water*") AND (prey OR "gut content*" OR diet* OR "food web*" OR "trophic interaction*") AND ("next generation sequencing" OR ngs OR "high throughput sequencing" OR metabarcoding OR barcoding OR "multiplex PCR" OR "molecular methods" OR "molecular tools" OR "molecular diet analysis" OR "molecular prey detection" OR "prey DNA")*

The results from the two searches were combined and duplicates were removed. The results were screened to remove irrelevant studies by applying inclusion criteria in three stages (title only, abstract, full text). The following inclusion criteria were used during the screening process:

- Study organism(s): macroscopic predatory animals feeding on macroscopic animal prey (excluded microscopic predators/prey, excluded herbivory)
- Habitat(s): studies where study organisms (predator/prey) inhabit freshwater (e.g. included terrestrial predators, living around freshwaters, feeding on freshwater prey, excluded terrestrial predators living around freshwaters, feeding on terrestrial prey).
- Molecular method: analysis of DNA in dietary samples (excluded environmental DNA)
- Focus: predator-prey interactions (excluded energy transfer, carbon accounting, nutrition/metabolism)
- System: diet of animals in natural habitats (excluded experimental systems and feeding trials)

A record was made of the number of studies included at each stage (see Table S1). Included studies then were categorised on:

- Taxonomy: predator class
- Freshwater habitat type: running water, standing water, riparian habitat, mixed habitat types
- Study focus: single predator, multiple predators, food web
- Temporal resolution of study: snapshot, temporal change in diet
- Sample type: invasive (gut contents, whole organism), non-invasive (faeces, regurgitated stomach contents)
- Molecular method: barcoding (methods involving sequencing of DNA barcodes), screening (methods involving screening for presence of specific sequences)

Table S1 Search terms and numbers of records at each stage

<table>
<tr><td rowspan="2">Identification</td><td>Web of Science Search (28/7/21)
Terms:
(freshwater* OR lake* OR river* OR stream OR streams OR wetland* OR pond* OR "inland water*") AND (prey OR "gut content*" OR diet* OR "food web*" OR "trophic interaction*") AND ("next generation sequencing" OR ngs OR "high throughput sequencing" OR metabarcoding OR barcoding OR "multiplex PCR" OR "molecular methods" OR "molecular tools" OR "molecular diet analysis" OR "molecular prey detection" OR "prey DNA")

records = 443</td><td>Scopus Search (28/7/21)
Terms:
freshwater* OR lake* OR river* OR stream OR streams OR wetland* OR pond* OR "inland water*"
AND
prey OR "gut content*" OR diet* OR "food web*" OR "trophic interaction*"
AND
"next generation sequencing" OR ngs OR "high throughput sequencing" OR metabarcoding OR barcoding OR "multiplex PCR" OR "molecular methods" OR "molecular tools" OR "molecular diet analysis" OR "molecular prey detection" OR "prey DNA"

records = 271</td></tr>
<tr><td colspan="2">Combined search results
records = 714</td></tr>
<tr><td rowspan="3">Screening</td><td colspan="2">Removed duplicates
records = 523</td></tr>
<tr><td colspan="2">Screening: title review
records = 131</td></tr>
<tr><td colspan="2">Screening: abstract review
records = 63</td></tr>
<tr><td>Cited papers</td><td colspan="2">Papers added: cited in papers selected in screening stage
new records = 11
total records = 67</td></tr>
</table>

Appendix S2: Electrophoresis gel images of PCR tests

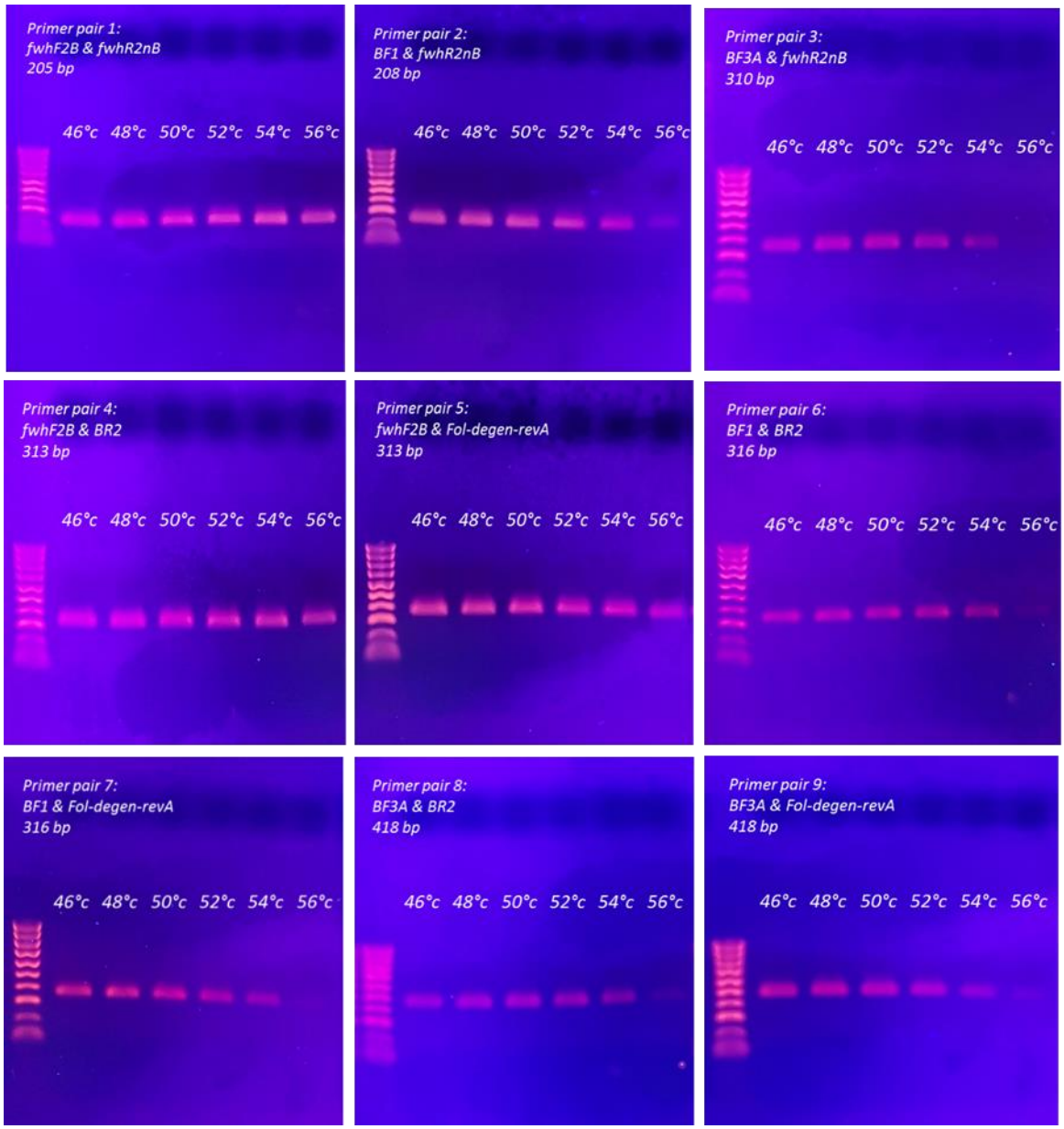

Figure S2.1 Electrophoresis gel photos showing results of gradient PCR for each of the nine primer pairs using mock community DNA template.

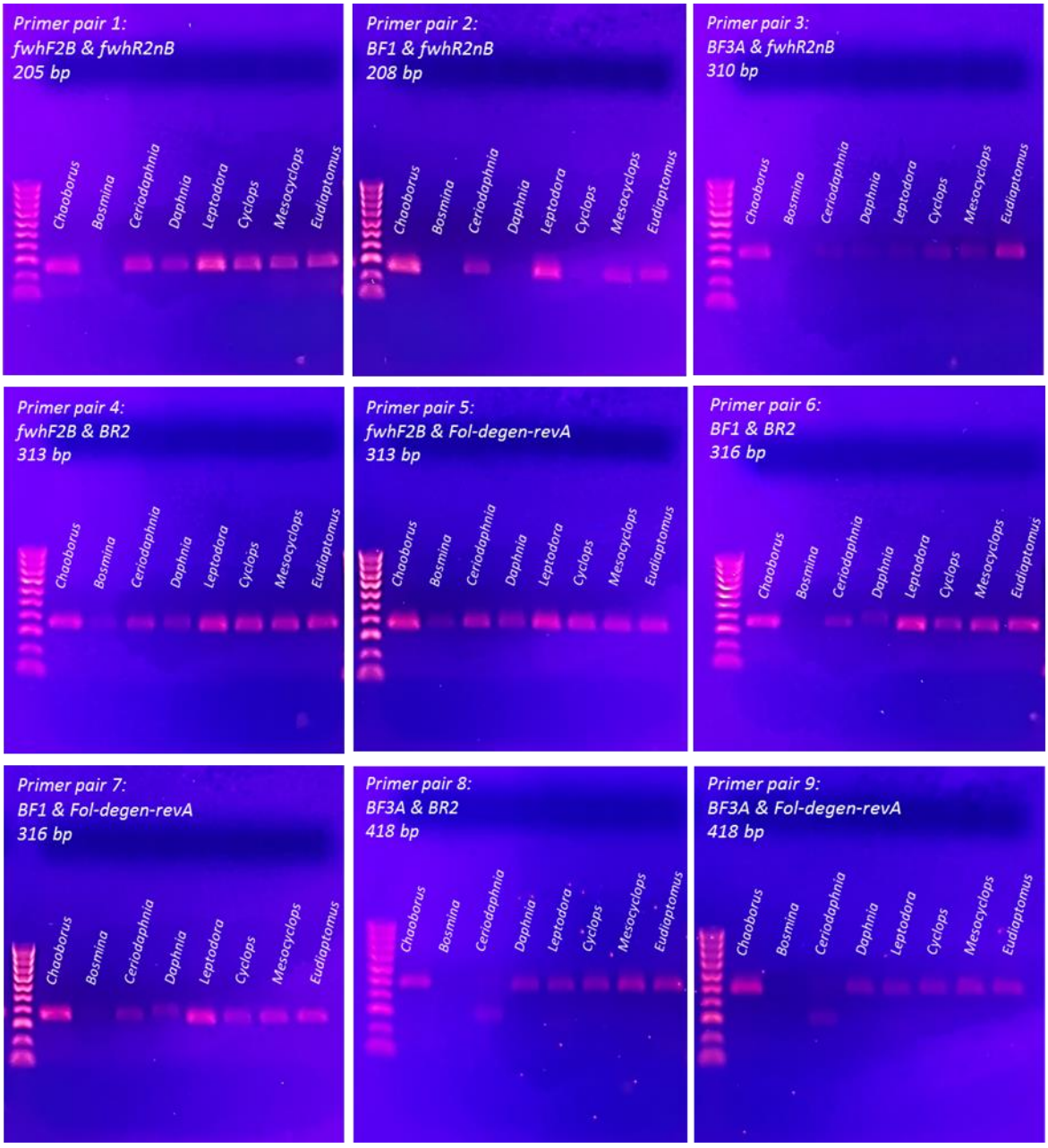

Figure S2.2 Electrophoresis gel photos showing results of single-taxon PCR for each of the nine primer pairs.

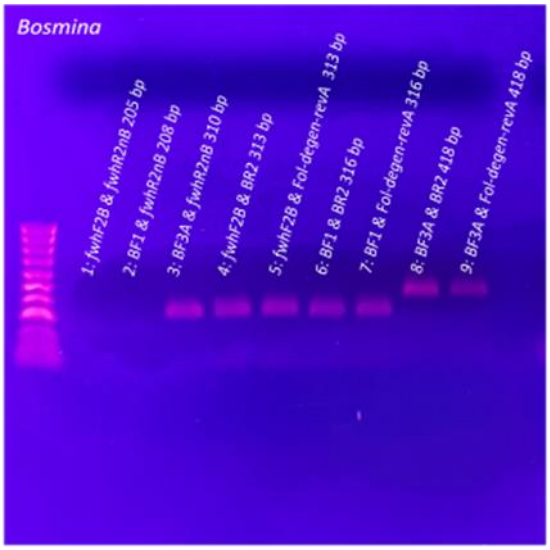

Figure S2.2 Electrophoresis gel photos showing results of repeated single-taxon PCR for Bosmina for each of the nine primer pairs.