DOCTOR OF PHILOSOPHY

Defining a High Throughput Sequencing identification framework for freshwater ecosystem biomonitoring

Bista, Iliana

Award date:
2016

Awarding institution:
Bangor University

Defining a High Throughput Sequencing identification framework for freshwater ecosystem biomonitoring

A thesis submitted for the degree of Doctor of Philosophy,
School of Biological Sciences,
Bangor University

Iliana - Aglaia Bista

2016

PRIFYSGOL
BANGOR
UNIVERSITY

Ysgoloriaethau Sgiliau Economi Gwybodaeth
Knowledge Economy Skills Scholarships

Declaration and Consent

Details of the Work

I hereby agree to deposit the following item in the digital repository maintained by Bangor University and/or in any other repository authorized for use by Bangor University.

Author Name: ……..

Title: ……..

Supervisor/Department: ……..

Funding body (if any): ……..

Qualification/Degree obtained: ……..

This item is a product of my own research endeavours and is covered by the agreement below in which the item is referred to as "the Work". It is identical in content to that deposited in the Library, subject to point 4 below.

Non-exclusive Rights

Rights granted to the digital repository through this agreement are entirely non-exclusive. I am free to publish the Work in its present version or future versions elsewhere.

I agree that Bangor University may electronically store, copy or translate the Work to any approved medium or format for the purpose of future preservation and accessibility. Bangor University is not under any obligation to reproduce or display the Work in the same formats or resolutions in which it was originally deposited.

Bangor University Digital Repository

I understand that work deposited in the digital repository will be accessible to a wide variety of people and institutions, including automated agents and search engines via the World Wide Web.

I understand that once the Work is deposited, the item and its metadata may be incorporated into public access catalogues or services, national databases of electronic theses and dissertations such as the British Library's EThOS or any service provided by the National Library of Wales.

I understand that the Work may be made available via the National Library of Wales Online Electronic Theses Service under the declared terms and conditions of use (http://www.llgc.org.uk/index.php?id=4676). I agree that as part of this service the National Library of Wales may electronically store, copy or convert the Work to any approved medium or format for the purpose of future preservation and accessibility. The National Library of Wales is not under any obligation to reproduce or display the Work in the same formats or resolutions in which it was originally deposited.

Statement 1:

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree unless as agreed by the University for approved dual awards.

Signed …….. (candidate)

Date ……..

Statement 2:

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

All other sources are acknowledged by footnotes and/or a bibliography.

Signed …….. (candidate)

Date ……..

Statement 3:

I hereby give consent for my thesis, if accepted, to be available for photocopying, for inter-library loan and for electronic storage (subject to any constraints as defined in statement 4), and for the title and summary to be made available to outside organisations.

Signed …….. (candidate)

Date ……..

NB: Candidates on whose behalf a bar on access has been approved by the Academic Registry should use the following version of **Statement 3:**

Statement 3 (bar):

I hereby give consent for my thesis, if accepted, to be available for photocopying, for inter-library loans and for electronic storage (subject to any constraints as defined in statement 4), after expiry of a bar on access.

Signed …….. (candidate)

Date ……..

<u>Statement 4</u>:

Choose **<u>one</u>** of the following options

a)	I agree to deposit an electronic copy of my thesis (the Work) in the Bangor University (BU) Institutional Digital Repository, the British Library ETHOS system, and/or in any other repository authorized for use by Bangor University and where necessary have gained the required permissions for the use of third party material.	
b)	I agree to deposit an electronic copy of my thesis (the Work) in the Bangor University (BU) Institutional Digital Repository, the British Library ETHOS system, and/or in any other repository authorized for use by Bangor University when the approved **bar on access** has been lifted.	
c)	I agree to submit my thesis (the Work) electronically via Bangor University's e-submission system, however I **opt-out** of the electronic deposit to the Bangor University (BU) Institutional Digital Repository, the British Library ETHOS system, and/or in any other repository authorized for use by Bangor University, due to lack of permissions for use of third party material.	

Options B should only be used if a bar on access has been approved by the University.

In addition to the above I also agree to the following:

1. That I am the author or have the authority of the author(s) to make this agreement and do hereby give Bangor University the right to make available the Work in the way described above.

2. That the electronic copy of the Work deposited in the digital repository and covered by this agreement, is identical in content to the paper copy of the Work deposited in the Bangor University Library, subject to point 4 below.

3. That I have exercised reasonable care to ensure that the Work is original and, to the best of my knowledge, does not breach any laws – including those relating to defamation, libel and copyright.

4. That I have, in instances where the intellectual property of other authors or copyright holders is included in the Work, and where appropriate, gained explicit permission for the inclusion of that material in the Work, and in the electronic form of the Work as accessed through the open access digital repository, *or* that I have identified and removed that material for which adequate and appropriate permission has not been obtained and which will be inaccessible via the digital repository.

5. That Bangor University does not hold any obligation to take legal action on behalf of the Depositor, or other rights holders, in the event of a breach of intellectual property rights, or any other right, in the material deposited.

6. That I will indemnify and keep indemnified Bangor University and the National Library of Wales from and against any loss, liability, claim or damage, including without limitation any related legal fees and court costs (on a full indemnity bases), related to any breach by myself of any term of this agreement.

Signature: …… Date : ……..

This PhD thesis is dedicated to my
parents Christos & Efi Bista

Αυτή η διδακτορική διατριβή είναι αφιερωμένη στους
γονείς μου Χρήστο & Έφη Μπίστα

Ἰθάκη

Σὰ βγεῖς στὸν πηγαιμὸ γιὰ τὴν Ἰθάκη,
νὰ εὔχεσαι νἆναι μακρὺς ὁ δρόμος,
γεμάτος περιπέτειες, γεμάτος γνώσεις.

Τοὺς Λαιστρυγόνας καὶ τοὺς Κύκλωπας,
τὸν θυμωμένο Ποσειδῶνα μὴ φοβᾶσαι,
τέτοια στὸν δρόμο σου ποτέ σου δὲν θὰ βρεῖς,
ἂν μέν᾽ ἡ σκέψις σου ὑψηλή, ἂν ἐκλεκτὴ
συγκίνησις τὸ πνεῦμα καὶ τὸ σῶμα σου ἀγγίζει.

Τοὺς Λαιστρυγόνας καὶ τοὺς Κύκλωπας,
τὸν ἄγριο Ποσειδώνα δὲν θὰ συναντήσεις,
ἂν δὲν τοὺς κουβανεῖς μὲς στὴν ψυχή σου,
ἂν ἡ ψυχή σου δὲν τοὺς στήνει ἐμπρός σου.

Νὰ εὔχεσαι νά ΄ναι μακρὺς ὁ δρόμος.
Πολλὰ τὰ καλοκαιρινὰ πρωϊὰ νὰ εἶναι
ποὺ μὲ τί εὐχαρίστηση, μὲ τί χαρὰ
θὰ μπαίνεις σὲ λιμένας πρωτοειδωμένους·

νὰ σταματήσεις σ᾽ ἐμπορεῖα Φοινικικά,
καὶ τὲς καλὲς πραγμάτειες ν᾽ ἀποκτήσεις,
σεντέφια καὶ κοράλλια, κεχριμπάρια κ᾽ ἔβενους,
καὶ ἡδονικὰ μυρωδικὰ κάθε λογῆς,
ὅσο μπορεῖς πιὸ ἄφθονα ἡδονικὰ μυρωδικά.

Σὲ πόλεις Αἰγυπτιακὲς πολλὲς νὰ πᾷς,
νὰ μάθεις καὶ νὰ μάθεις ἀπ᾽ τοὺς σπουδασμένους.
Πάντα στὸ νοῦ σου νἄχῃς τὴν Ἰθάκη.
Τὸ φθάσιμον ἐκεῖ εἶν᾽ ὁ προορισμός σου.

Ἀλλὰ μὴ βιάζῃς τὸ ταξείδι διόλου.
Καλλίτερα χρόνια πολλὰ νὰ διαρκέσει.
Καὶ γέρος πιὰ ν᾽ ἀράξῃς στὸ νησί,
πλούσιος μὲ ὅσα κέρδισες στὸν δρόμο,
μὴ προσδοκώντας πλούτη νὰ σὲ δώσῃ ἡ Ἰθάκη.

Ἡ Ἰθάκη σ᾽ ἔδωσε τ᾽ ὡραῖο ταξίδι.
Χωρὶς αὐτὴν δὲν θἄβγαινες στὸν δρόμο.
Ἄλλα δὲν ἔχει νὰ σὲ δώσει πιά.

Κι ἂν πτωχικὴ τὴν βρῇς, ἡ Ἰθάκη δὲν σὲ γέλασε.
Ἔτσι σοφὸς ποὺ ἔγινες, μὲ τόση πείρα,
ἤδη θὰ τὸ κατάλαβες ἡ Ἰθάκες τί σημαίνουν.

Κ.Π.Καβάφη

Ithaka

As you set out for Ithaka
hope the voyage is a long one,
full of adventure, full of discovery.
Laistrygonians and Cyclops,
angry Poseidon - don't be afraid of them;
you'll never find things like that on your way
as long as you keep your thoughts raised high,
as long as a rare excitement
stirs your spirit and your body.
Laistrygonians and Cyclops,
wild Poseidon - you won't encounter them,
unless you bring them along inside your soul,
unless your soul sets them up in front of you.
Hope the voyage is a long one.
May there be many a summer morning when
with what pleasure, what joy,
you come into harbours seen for the first time;
may you stop at Phoenician trading stations
to buy fine things,
mother of pearl and coral, amber and ebony,
sensual perfume of every kind -
as many sensual perfumes as you can;
and may you visit many Egyptian cities
to gather stores of knowledge from their scholars.
Keep Ithaka always in your mind.
Arriving there is what you are destined for.
But do not hurry the journey at all.
Better if it lasts for years,
so you are old by the time you reach the island,
wealthy with all you have gained on the way,
not expecting Ithaka to make you rich.
Ithaka gave you the marvelous journey.
Without her you would not have set out.
She has nothing left to give you now.
And if you find her poor, Ithaka won't have fooled you.
Wise as you will have become, so full of experience,
you will have understood by then what these Ithakas mean

C.P. Cavafy

Summary

Freshwater ecosystems are currently amongst the most threatened habitats due to high levels of anthropogenic stress and increasing efforts are required to monitor their status and assess aquatic biodiversity. Biomonitoring, which is the systematic measurement of the responses of aquatic biota to environmental stressors, is used to evaluate ecosystem status. Macroinvertebrates are commonly used organisms for ecosystem assessment, due to their numerous biomonitoring qualities, which qualify them as ecological indicators. Traditional taxonomy-based monitoring is labour intensive, which limits the throughput, and is often inefficient in providing species level identification, which limits the accuracy of detections. The introduction of molecular based methods for biomonitoring, especially when coupled with High Throughput Sequencing (HTS) applications, offers a step change in ecosystem monitoring.

Here I tested the utility of DNA based applications for increasing the efficiency of freshwater ecosystem biomonitoring, using benthic macroinvertebrates as a target group. For the first part of this work, I used DNA barcoding of the Cytochrome Oxidase Subunit I (COI), from individual specimens, to populate a barcode reference library for 94 species of Trichoptera, Gastropoda and Chironomidae from the UK. Then, I used High Throughput Sequencing (HTS) methods to characterise diversity from complex environmental samples. First, I used metabarcoding of aqueous environmental DNA (eDNA) and community invertebrate samples (Chironomidae pupal exuviae), collected on regular intervals throughout a year, to identify diversity levels and temporal patterns of community variation on ecosystem-wide and group specific scales. Finally, I used a structured design of mock macroinvertebrate communities, of known biomass content, to perform a comparison between PCR-based metabarcoding of the COI gene and PCR-free shotgun sequencing of mitochondrial genomes (mito-metagenomics), and evaluate their efficiency for accurate characterisation of biomass content of bulk samples. Overall, HTS has demonstrated great potential for advancing biomonitoring efforts, allowing ecosystem scale diversity detection from non-invasive types of samples, such as eDNA, whilst moving into mito-metagenomic work could improve the field even further by improving quantitative abundance results on the community composition level.

Acknowledgements

First, I would like to thank Bangor University and the Molecular Ecology and Fisheries Genetics laboratory, for giving me the opportunity to work here all these years in such a diverse and exciting research environment.

I would like to acknowledge the Knowledge Economy Skills Scholarships (KESS) for funding me and this project, and for the learning and networking opportunities provided to me. The Environment Agency UK, for funding this work, providing infrastructure, resources and an ongoing collaboration. In addition, I acknowledge the Freshwater Biological Association (FBA) (Gilson-Le Cren Memorial Award 2014) and the Natural Environment Research Council Biomolecular Analysis Facility (NERC-NBAF pilot project) for additional funding, and High Throughput Computer (HPC) Wales systems for allowing use of their systems for analysis.

This project would not have come true if it wasn't for my supervisors.

I want to thank my primary supervisor Simon Creer, who initiated this project and selected me to take it on. Thank you for your guidance and help throughout this project, for enthusiastically chasing new ideas and having long discussions on designing new experiments, for providing your knowledge on biodiversity and the new ways to explore it, which was what brought me to Bangor in the first place, for opening up new horizons for me to walk through, and for challenging me and even throwing me at the deep end of the pool, at times, to lean to swim; and swim I did.

I also thank my co-supervisor Gary Carvalho. Thank you for your endless enthusiasm, your insights and ability to see the bigger ecological picture and giving context to new ideas, for believing in me and letting me know when it was most needed, for making MEFGL the vibrant scientific and welcoming place that it is.

I thank my supervisors from the Environment Agency, Kerry Walsh and Martin Christmas. Kerry for her steady hand and efficiency when it came to summoning resources and people for sample collection and the applicability of the work and new ideas, and Martin

for his help especially at the beginning of this project, by availing resources and recruiting EA staff for sample collection.

I thank my collaborators at the Beijing Genomic Institute (BGI) for performing sequencing and assembly of mito-genomes (Chapter 4) and continuing collaboration, specifically Xin Zhou, Min Tang and Shanlin Liu. Also, I thank my collaborators at the Biodiversity Institute in Guelph, Canada for running the metabarcoding samples from Chapter 4 on their MiSeq and continuing collaboration, specifically Mehrdad Hajibabaei and Shadi Shokralla.

I want to thank Mathew Seymour for his valuable help and input with statistical modelling and Delphine Lallias for help with development of amplicon library protocols and bioinformatics analyses pipelines. Environment Agency staff from the Cornwall, Devon, East Anglia, Warrington, York, Cumbria and other offices, Scottish Environment Protection Agency (SEPA), Centre for Ecology and Hydrology (CEH) for invertebrate sample collection for the Barcode reference library and chironomid exuviae identification (Chapters 1 and 3). The Bangor EA office for helping me at the start of the project to learn invertebrate sampling, sorting and identification skills. Adrian Chalkley and David Bentley for collection and identification of specimens used in Chapter 4. Rosie Blackman for Trichoptera identification help. APEM and specifically David Bradley, Les Ruse and colleagues for collection of CPET samples for barcoding (Chapter 1 and 3), and for confirming identification of specimens collected by A. Chalkley and D. Bentley (Chapter 4). Shaun Dowman for providing the Hydroptilidae adult specimens for barcoding. John Bratton for collecting a great number of beetle species. Tristan Hatton-Ellis and Natural Resources Wales (NRW) for providing historical data for Padarn Lake. Ade Fewings for bioinformatics support related to using the HPC systems and Linux tips. Penny Dowdney for help with KESS matters and resources.

Especially I would like to thank my extended lab-family which has harboured me for years and helped me grow as a researcher and as a person, and all my friends in Bangor who came and went and especially those who stayed, and all my friends and family who never gave up on me, even if so far away.

Text removed at request of the author

Table of Contents

List of figures

List of tables

Chapter 1

General Introduction

Chapter 1: General Introduction

1.1 Freshwater ecosystem biodiversity and monitoring - an overview

The increasing threats on freshwater ecosystems in relation to anthropogenic stress are driving the loss of biodiversity and ecosystem degradation with consequences to environments and society (Cardinale *et al.* 2012; Heino 2013). To measure and negate such impacts, biomonitoring methods have been developed which measure the responses of biological communities to environmental stressors. By focusing on organisms such as macroinvertebrates, measures of alpha and beta diversity can be used to evaluate ecosystem status (Bonada *et al.* 2006; Kenney *et al.* 2009). Stakeholder decisions depend on such results and ecological status evaluations to implement policies for management and restoration of the affected ecosystems and preservation of pristine environments (Bonada *et al.* 2006; Friberg *et al.* 2011). Overall, the efficiency of management decisions relies to a large degree on the accuracy of the outcomes of biomonitoring (Kenney *et al.* 2009; Collins *et al.* 2012).

Traditionally, biomonitoring is performed through taxonomic identification of indicator organisms with particular interest on certain groups (Reynoldson & Metcalfe-Smith 1992). This approach has proved challenging because it is labour intensive, time consuming and limited to the identification of certain life stages etc. (Pilgrim *et al.* 2011). To advance the field of biomonitoring the use of DNA-based approaches has been proposed as a means of increasing throughput and accuracy (Baird & Hajibabaei 2012).

DNA barcoding, which is the sequencing of a 658bp fragment of the Cytochrome Oxidase subunit I (COI) gene, has been used for rapid species identification (Hebert *et al.* 2003a). Barcoding constitutes a continuously increasing source of DNA information, which can be harnessed to promote cataloguing of biodiversity, gain phylogenetic insights of communities and assist detection of new species (Joly *et al.* 2014). Furthermore, the wealth of information found in barcoding repositories could be used for enhancing the accuracy of other molecular approaches, such as High Throughput Sequencing (HTS) of whole communities (metabarcoding of bulk samples), environmental DNA (eDNA) and mitochondrial genome sequencing.

For metabarcoding of communities, bulk samples are extracted directly, removing the sorting and identification steps, and are then sequenced to reveal a wealth of biodiversity previously unknown (Fonseca *et al.* 2010; Leray & Knowlton 2015) or provide information on species richness and community composition useful for ecosystem management (Ji *et al.* 2013) and biomonitoring (Hajibabaei *et al.* 2011). These methods have been shown to outperform traditional surveys (Yu *et al.* 2012), though the accuracy of relative abundance estimation due to primer biases with metabarcoding has been questioned (Piñol *et al.* 2015). In turn, the field of shotgun mito-metagenomics has emerged offering up a new solution to the limitations of metabarcoding for measuring relative abundance (Zhou *et al.* 2013; Tang *et al.* 2015). In this case, the complete mitochondrial genomes can be shotgun sequenced without amplification, which not only removes potential PCR related bias but also provides more sequencing information, across the mitochondrial genome, moving away from the limitations of single marker approaches (Crampton-Platt *et al.* 2016).

Moreover, whilst the field advances, Environmental DNA (eDNA) has taken a place on the spearhead of the biomonitoring molecular revolution. Environmental DNA, which can be extracted directly from environmental samples, such as water, is increasingly being used for the detection of biodiversity (Lodge *et al.* 2012). For freshwater ecosystems the majority of current applications focus on the detection of rare or invasive species through the use of species or group specific assays (e.g. qPCR) (e.g. Ficetola *et al.* 2008; Goldberg *et al.* 2011; Minamoto *et al.* 2012; Biggs *et al.* 2015). The next step forward for this field is the use of metabarcoding of eDNA, which would allow multi-taxon and ecosystem-wide biodiversity assessment for aquatic environments (Thomsen & Willerslev 2015).

The subjects described so far will be discussed in finer detail in the course of this chapter. Furthermore, all three main types of DNA based methodologies mentioned, including DNA barcoding, metabarcoding of eDNA and bulk communities, and shotgun sequencing of mitochondrial genomes, are employed in the three subsequent experimental chapters.

1.2 Past and present bioassessment systems

For the assessment of lotic (flowing waters) and lentic (still waters) waterbodies, two main approaches are employed, either through the measurement of water physicochemical properties or through biological measurements (Kenney *et al.* 2009). The systematic study of the responses of the biological community to environmental stressors can then be used to evaluate ecosystem changes, a process known as biological monitoring or biomonitoring (Matthews *et al.* 1982). Biological organisms act as indicators of the quality of their environment, since different species are known to have particular requirements regarding oxygen and nutrient levels and varying tolerance limits to substances such as metals (Dunigan 1988). The selection of the species used for ecosystem monitoring is based on a variety of key basic criteria which grant the organisms the "indicator" status, such as the ease of collection and identification, width of habitat distribution, links to autecological data and possibility of bioaccumulation and laboratory culture (Dunigan 1988). Several taxonomic groups have been used for biomonitoring, ranging from bacteria to protozoa, algae, macroinvertebrates, macrophytes and fish (Dunigan 1988; Friberg *et al.* 2011).

One of the earliest efforts to use benthos for ecological assessments was through the Saprobien system, in which case individual scores were assigned to taxa in relation to their tolerance to organic pollution (Bonada *et al.* 2006). Alternative approaches developed later used diversity indices comprising abundance, richness and evenness of taxa in the community (Gray *et al.* 2015). Both types of methods were eventually replaced by a third system comprising both individual species characteristics as well as diversity indexes (Armitage *et al.* 1983; Reynoldson & Metcalfe-Smith 1992).

The Biological Monitoring Working Party (BMWP) was developed in 1978, and it was the first attempt to establish a river biomonitoring system using benthic macroinvertebrates that would be nationally applicable for the UK (Hawkes, A 1998). For this system, an aggregate score per site was taken (BMWP score) extracted from the sensitivity of different macroinvertebrate families to pollution, though the drawback was that taxa exposed to common levels of pollution did not always exhibit common scores due to natural species variability and different site characteristics. To overcome such problems,

the River Invertebrate Prediction and Classification System (RIVPACS) for assessment of ecological quality or rivers was developed (Logan 2001). The basic idea behind RIVPACS is the collection of knowledge regarding the fauna and environmental characteristics of reference sites, against which future monitored and possibly disturbed sites will be compared (Clarke *et al.* 2003). The system was originally developed and is currently used for rivers in England, Wales, Scotland and Northern Ireland through a suite of a total of 835 sites across the UK (Kille 2011), and is now superseded by the River Invertebrate Classification System (RICT) (Friberg *et al.* 2011).

Furthermore, in 2000, the European Union (EU) established the Directive 2000/60/EU, also known as the Water Framework Directive (WFD), which is the most extensive water related piece of legislation to date. The WFD refers to the management and protection of freshwater resources and ecosystems (Griffiths 2002; Howarth 2009; Collins *et al.* 2012), while attempting to draw a consistent framework for freshwater monitoring across 27 countries (Hatton-Ellis 2008). In the WFD, the ecological quality of waterbodies is assessed through environmental quality metrics, which are derived from taxon richness and abundance data. Subsequently, these metrics are used to compare the ecosystems against a reference condition and categorise them through an Ecological Quality Ratio (EQR) system (Hatton-Ellis 2008).

1.3 Macroinvertebrates as bio-indicators

Freshwater benthic macroinvertebrates are organisms with a body length of more than 0.25mm (aquatic life stages), though most are longer than 2mm and some are several centimetres long, which are derived from a multitude of insect orders, as well as crustaceans, molluscs, oligochaetes and others (Kenney *et al.* 2009). Aquatic invertebrates and particularly insects, feature prominently in environmental impact assessment (Cranston 1990). That is because they fulfil many of the criteria for being good bioindicators like ubiquity; large species richness which covers the spectrum of environmental responses; possibilities for reflecting cumulative environmental impacts of stressors due to their relatively long life cycles and their sedentary nature, which allows for site-specific indications (Bonada *et al.* 2006). For the UK in particular, the taxonomy of

macroinvertebrates is relatively well studied and characterised to allow in depth use of the various groups for bioassessment (Kille 2011).

Some of the macroinvertebrate groups used more extensively for bioassessment include insects, such as members of the orders Ephemeroptera (mayflies), Plecoptera (stoneflies), Trichoptera (caddisflies), Diptera (true flies), Mollusca (snails and mussels), Crustaceans (crayfish), Annelida (aquatic worms and leeches) and others. Some of these groups receive particular attention, such as the EPT insects (Ephemeroptera, Plecoptera, Trichoptera), whose presence for example is considered to be an indication of a healthy stream (Kenney *et al.* 2009; Zhou *et al.* 2010). The Trichoptera group in particular is one of the most diverse freshwater insect groups and their genetic diversity has been increasingly more studied in recent years (e.g. Zhou *et al.* 2011a). Gastropods are also commonly used for bioassessment, but despite their high diversity (>4,000 species worldwide), they are generally not very well studied (Strong *et al.* 2008).

Moreover, the family Chironomidae (Diptera, non-biting midges), is of particular importance as they are the most abundant and most widely distributed macroinvertebrate group in freshwater ecosystems (Sharley *et al.* 2004; Armitage *et al.* 2012). Additionally they are the most species rich family, with over 612 species just in the UK (Wilson & Ruse 2005), and probably more than 10,000 species overall (Armitage *et al.* 2012). Chironomids are particularly useful for monitoring of acidification and eutrophication and the additional option of collecting shed pupal skins (Chironomid Pupal Exuviae Technique, CPET) makes them particularly useful for characterisation of still waters/lake ecosystems (Kille 2011). Taking into consideration the huge diversity within the Chironomidae, achieving taxonomic resolution to the species level using just morphological identification can be a major challenge (Kille 2011). Due to this impediment, several recent studies have used molecular analysis for chironomid species identification to overcome these problems (e.g. Sharley *et al.* 2004; Carew *et al.* 2005, 2013; Brodin *et al.* 2013).

1.4 Traditional taxonomic identification and the need for DNA based methods

Despite our understanding that the assessment of ecosystem health is of vital importance, the accurate identification of species, based on traditional identification methods through morphological characteristics has proved to be a difficult task (Hajibabaei *et al.* 2011). The level of identification required for the organisms to be identified to, in order to enable ecosystem quality assessment with an acceptable degree of certainty, is known as taxonomic sufficiency of identification (Jones 2008). For macroinvertebrates, the exact level of taxonomic resolution that is necessary for a meaningful biotic assessment has been debated (Bailey *et al.*), but it has become more or less clear that a species level identification is required because it produces more robust assessment results (Lenat & Resh 2001).

Identification of macroinvertebrates using traditional morphology based approaches is not always possible and several shortcomings of this work exist, especially when it comes to identifying specimens to the species level. In particular, identification is often possible only for some life stages for which morphological keys have been developed, or certain life stages are easier to identify than others (e.g. adults vs. larvae). Also, identification is either not possible or more difficult for one or the other sex. For example, adult female chironomids are often being neglected due to challenges associated with their identification (Ekrem *et al.* 2010). Furthermore, the level of taxonomic expertise required for species identification is often high and the process time consuming, as for chironomid larvae, which have to be mounted on microscope slides for species level identification (Ferrington etal. 1991). When multiple kick-samples have to be analysed, this process is very labour intensive and time consuming (Reynoldson & Metcalfe-Smith 1992).

Using an identification system based on coarse taxonomic levels may initially appear to have some advantages, such as speed and lower cost (Schmidt-Kloiber & Nijboer 2004). Nevertheless, when applied to bioassessment a coarse level of specimen identification may affect the results, distort the species-specific signals, and eventually hinder the detection of biological impact of stressors on the ecosystem (Schmidt-Kloiber & Nijboer 2004; Arscott *et al.* 2006; Pfrender *et al.* 2010).

Considering the need for accurate identification of species to the finest taxonomic level possible, and the difficulties that traditional taxonomy is facing (Kenney *et al.* 2009; Pfrender *et al.* 2010), a need for a more robust, accurate and high throughput method is required. The application of DNA sequencing can provide accuracy and speed in the identification process to overcome such difficulties (Hajibabaei *et al.* 2011), thereby significantly enhancing the capacity for taxonomic inventory of benthic macroinvertebrate species (Sweeney *et al.* 2011). Such proposed methods for incorporation of molecular approaches into biomonitoring include DNA Barcoding, High Throughput Sequencing of community invertebrate samples using metabarcoding or shotgun sequencing, and incorporation of environmental DNA (eDNA) assays.

1.5 DNA Barcoding

DNA Barcoding uses a short standardized 658bp fragment of the 5' region of the Cytochrome Oxidase subunit I (COI) mitochondrial gene as a means of genetically distinguishing between individuals of different species (Hebert *et al.* 2003a). Even though this has been promoted and conceptualised more recently for the purposes of DNA barcoding (Tautz *et al.* 2003; Hebert *et al.* 2003a), using DNA sequence divergence for species discrimination is not an entirely new concept (Avise 2004; Moritz & Cicero 2004; Ward *et al.* 2005). The novelty in this case though, resides with the standardization of the method and the increase of scale and accessibility, both in approach and baseline reference data (Moritz & Cicero 2004).

The standardized COI barcoding fragment has been adopted as the focal locus for DNA barcoding as it has been suggested to "possess a greater range of phylogenetic signal than any other mitochondrial gene" (Hebert *et al.* 2003a). Furthermore, universal primers have been designed, which can amplify this fragment from a variety of metazoan phyla (Folmer *et al.* 1994). Except for the typical barcoding use of the COI, it could also be used as a proxy for the nucleotide composition of the mitochondrial genome, as it has been suggested that this region mimics the nucleotide composition of the entire mtDNA (Min & Hickey 2007; Clare *et al.* 2008; Costa & Carvalho 2010). Nevertheless, there is still an ongoing debate as to whether one gene is suitable for the identification of all species

(Moritz & Cicero 2004; Deagle *et al.* 2014). Additionally, other barcoding genes are considered more suitable for some groups of organisms, like the 16S ribosomal RNA for bacteria (Lenobah *et al.* 2014), the internal transcribed region (ITS) for fungi (Schoch *et al.* 2012), the maturase K (*matK*) plastid gene and the ribulose-bisphosphate carboxylase (*RbcL*) for plants (Hollingsworth *et al.* 2009).

Regardless of its utility for species discovery, DNA barcoding has many advantages for species identification (Hebert *et al.* 2003b), such as resolving cases of taxonomic identification when cryptic species are present (Hebert *et al.* 2004b). Additionally, it can help link taxonomic knowledge from different life stages to create a complete profile of the species and it can be used for the identification of life stages (e.g. larval) that are very difficult or cannot be identified currently, and for very small or damaged specimens (Carew *et al.* 2005; Taylor & Harris 2012). For aquatic invertebrates for example, taxonomic identification is only possible for males and some late instars, but the coupling of barcoding with traditional taxonomy provides a robust framework for biological identification (Zhou *et al.* 2007, 2009; DeWalt 2011).

Along with the standard 658bp fragment of the COI, shorter fragments of the same region have also been used, known as “mini barcodes” (Hajibabaei *et al.* 2006b). These shorter fragments could be useful for highly degraded samples, such as old museum specimens, specimens preserved in non DNA friendly means (e.g. formalin), or processed biological material like food products (Meusnier *et al.* 2008; Baird & Sweeney 2011). Alternative universal metazoan COI primers have also been designed more recently, targeting a 313bp region of the COI, using a newly designed forward primer combined with the Folmer reverse primer (Leray *et al.* 2013).

Furthermore, one of the key concepts of DNA Barcoding relies on the sequence divergence between species (interspecific) being higher than within species (intraspecific), a concept also known as the Barcoding Gap (Meyer & Paulay 2005) (Figure 1.1). The presence of the barcoding gap was initially confirmed by studying bird species (Hebert *et al.* 2004b), where it was found that between species sequence divergence exceeded the within species divergence by far, which was also the case from the study of 207 Australian fish species (Ward *et al.* 2005). Similar findings have also been reported from invertebrate

studies, which also found distinct levels of divergence for butterflies, springtails and spiders (Hogg & Hebert 2004; Hebert *et al.* 2004a; Barrett & Hebert 2005; Hajibabaei *et al.* 2006a).

Conversely, other studies have demonstrated exceptions in this case (Ward *et al.* 2005), or have even questioned the existence of a verified barcoding gap across all organisms, while it has been suggested that the detection of a barcoding gap might be an artefact resulting from insufficient sampling (Moritz & Cicero 2004; Meyer & Paulay 2005; Wiemers & Fiedler 2007). A possible approach for evaluating the presence of a barcoding gap in obtained results, could be by focusing on comparisons of sister species, as was done on a N. American bird dataset by Johnson & Cicero (2004). Furthermore, to enhance species identification through DNA barcodes when intra and inter specific distances overlap, advanced computational methods could be employed. Some of these approaches include Bayesian Model Comparison (BMC) (Meier *et al.* 2006), Bootstrap NJ (Munch *et al.* 2008b), Bayesian (Munch *et al.* 2008a) as well as Minimum Distance (MD) plus fuzzy species set methods (Zhang *et al.* 2012). Alternatively, or in combination, additional markers systems can be employed.

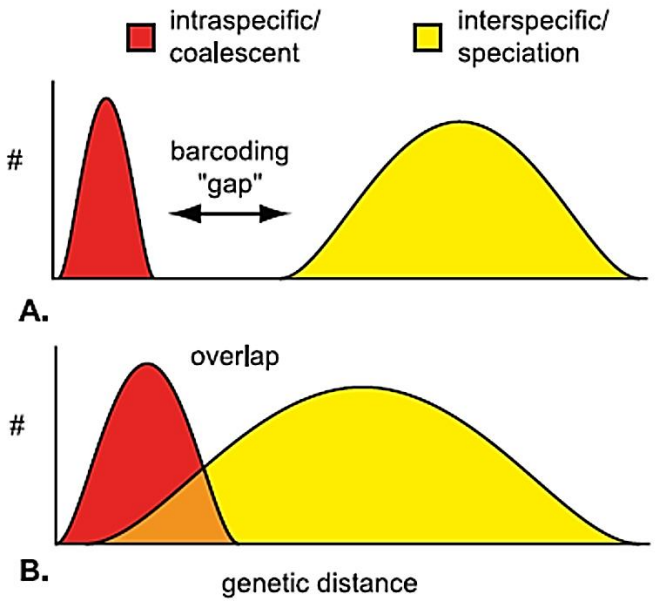


Figure 1.1: Schematic representation of the Barcoding Gap (BG) concept.
A.) the ideal case for the BG where the intra and inter species levels of divergence are clearly separated, and B.) case of overlapping divergence levels, Meyer & Paulay (2005).

The above controversy points out the importance of differentiating between the use of DNA barcodes as a diagnostic tool or as a means of discovering new species. In the first case, DNA Barcoding is used to distinguish between already identified species, while in the second case DNA information is used for species delineation through the evolutionary species concept; the second case falls into the DNA taxonomy category (Vogler & Monaghan 2007). It is the adoption of DNA taxonomy as a species discovery tool that is mainly considered to pose a threat to taxonomists (DeSalle 2007). Even though DNA Barcoding has also been advocated as a species discovery tool, it seems more likely that the combined analysis of morphological and molecular data will provide the best solution into what is currently called "integrative taxonomy" (Will *et al.* 2005; Teletchea 2010).

The diagnostic ability of DNA barcoding for species identification can also be affected by the presence of nuclear mitochondrial pseudogenes (NUMTs) (Moulton *et al.* 2010). NUMTs are non-functional copies of mitochondrial genes that have been integrated, through various mechanisms, into the nuclear genome and can be amplified along with the actual mtDNA genes during PCR (Song *et al.* 2008). Since they were first reported in 1967 (du Buy & Riley 1967), several studies have addressed the mechanisms of their formation, their presence across taxa and their possible function and evolution (Bensasson *et al.* 2001). Co-amplification of NUMTs with the orthologous mitochondrial gene, when conserved universal primers are used, challenges DNA barcoding and can lead to overestimation of the number of species present, though it is possible that their unusual mode of molecular evolution might make their detection possible (Moulton *et al.* 2010).

To date, DNA Barcoding has been applied in many groups of organisms, while the international effort of collecting DNA barcodes of species is coordinated through the Consortium for the Barcode of Life Initiative (CBOL) (http://www.barcodeoflife.org). The great interest surrounding DNA barcoding has led to the allocation of millions of dollars in research programs for its application and the establishment of the Consortium for the Barcode of Life (CBOL) initiative. For the coordination of global efforts, the International Barcode of Life was launched in October 2010 (iBOL) (Vernooy *et al.* 2010). Until now,

174,572 animal species have been barcoded, with 5,227,350 barcode sequences in total (http://www.boldsystems.org/) (November 2016).

When applied in biomonitoring related organisms such as macroinvertebrates, DNA barcoding has the potential to increase the accuracy of benthic macroinvertebrate taxonomic identification, as well as increase the level of information available for the calculation of water-quality metrics, such as species richness (up to 50%) (Baird & Sweeney 2011). An increasing number of DNA Barcoding studies have targeted freshwater benthic fauna (e.g. studies for Ephemeroptera (Webb *et al.* 2012), Trichoptera (Zhou *et al.* 2011), Chironomidae (Pfenninger *et al.* 2007; Kim *et al.* 2012) and freshwater mussels (Boyer *et al.* 2011)). Furthermore, the efficiency of DNA barcoding of benthic assemblages for monitoring purposes has been tested (e.g. Baird & Sweeney 2011; Pilgrim *et al.* 2011; Brodin *et al.* 2013). As an example, Sweeney *et al.* (2011) demonstrated that barcoding can successfully identify invertebrate species with a 2 to 4% genetic divergence and in that case the taxonomic inventory of the studied sites was increased by 70% from the barcoding data, compared to expert genus and species morphological identification.

1.6 High Throughput Sequencing and freshwater biomonitoring

Regarding traditional practises, currently used frameworks need to be updated, not only to move away from past practises, which have been disproved on occasion, but also to incorporate new technologies (Friberg *et al.* 2011). Even though the use of DNA barcoding has been a great advantage for studying freshwater invertebrates and could aid in accurate identification of specimens, the use of barcoding itself is still insufficient for applied ecosystem monitoring and currently not cost effective (Valentini *et al.* 2009). The most recently implemented High Throughput Sequencing (HTS) technologies promise to achieve large scale monitoring, faster and more accurately than traditional methods and overcome current constraints (Pfrender *et al.* 2010).

Baird & Hajibabaei (2012) reviewed the present situation in biomonitoring and the passage towards the new era by taking advantage of new sequencing technologies. Current assessment systems (named here Biomonitoring 1.0) are based on morphology, and ecosystem status outcomes are defined by restricted binary type evaluations of the

impacted/not-impacted type. In order to achieve a change from this current situation, a step increase in throughput and information content should be achieved as well, which will provide sufficient information to evaluate the impact of individual stressors (Baird & Hajibabaei 2012). This new more informative approach, named Biomonitoring 2.0, will use HTS to extract detailed species composition data from bulk environmental samples, which combined with associated metadata and accumulated ecological knowledge will provide more accurate ecosystem status diagnoses.

The method most commonly used for sequencing bulk/environmental samples is known as metabarcoding (Yoccoz 2012). For metabarcoding, DNA is extracted from bulk environmental samples, without separation of the contained organisms, which is then amplified with an appropriate barcode marker and sequenced on a high throughput sequencing platform (Yu *et al.* 2012), most frequently Illumina MiSeq. To distinguish between samples extracted from bulk communities and those extracted from trace amounts of DNA (e.g. water samples), we use the term community DNA for the former and environmental DNA (eDNA) for the latter. Even though, in both cases the samples are mixed and contain information from multiple organisms, the term community DNA also suggests the presence of tissue in the sample (e.g. invertebrates, benthic sediment samples, gut contents) (e.g. De Barba *et al.* 2014; Gibson *et al.* 2014), while the term eDNA refers to trace amounts of DNA (from cells, mitochondria, or free extracellular molecules) (Creer *et al.* 2016; Barnes & Turner 2016).

In one of the first applications of HTS for benthic macroinvertebrate samples, Hajibabaei *et al.* (2011) compared sequenced bulk samples from urban and conservation sites and found that an accurate representation of species diversity could be reached with this methodology. Since then a plethora of studies has published similar findings from environmental sample community analysis in freshwater ecosystems targeting macroinvertebrates (e.g. Gibson *et al.* 2014, 2015; Shokralla *et al.* 2015). Further from freshwater ecological monitoring applications, HTS is also extensively utilized for biodiversity monitoring aimed for conservation purposes (Schnell *et al.* 2012; Ji *et al.* 2013), and for detection of new diversity in poorly explored ecosystems (Leray & Knowlton 2015; Sinniger *et al.* 2016). Example studies also include detection of

biodiversity from past ecosystems (Willerslev *et al.* 2007; Jørgensen *et al.* 2012), for diet analysis and food web reconstruction (De Barba *et al.* 2014; Salinas-Ramos *et al.* 2015), as well as for the association of diversity with ecological function, community structure and other ecological applications (Creer *et al.* 2010; Hajibabaei *et al.* 2011; Lallias *et al.* 2015).

Nevertheless, shortcomings exist for the metabarcoding applications, which are limited by PCR biases and artefacts of the amplification process (e.g. chimeras Fonseca *et al.* 2012), the dependence on taxonomic reference libraries (Taberlet *et al.* 2012), and cases of environmental or laboratory contamination (Murray *et al.* 2015). The selection of markers for metabarcoding is also a controversial subject with supporters and opponents of the various existing markers (Deagle *et al.* 2014; Zhan *et al.* 2014). Nevertheless, the most commonly used marker for metazoan diversity remains the COI (Yoccoz 2012), though studies are also utilizing 16S (Clarke *et al.* 2014), 18S (Lallias *et al.* 2015; Sinniger *et al.* 2016), and 12S (De Barba *et al.* 2014; Miya *et al.* 2015) for metabarcoding.

To overcome biases of metabarcoding related to PCR amplification, mito-metagenomic methods have also been promoted. Mitochondrial metagenomics or mito-metagenomics is the shotgun sequencing of mitochondrial genomes from bulk samples, followed by in silico assembly of the genome sequences (Crampton-Platt *et al.* 2016). Zhou *et al.* (2013) tested the use of shotgun sequencing of mitochondrial genomes from macroinvertebrate communities and suggested that this approach could accurately present the diversity in bulk samples, whilst also providing an accurate representation of the relative abundance of species. Since then, additional studies have used mito-metagenomics for phylogenetic analysis of bulk communities (e.g. beetles) (Gillett *et al.* 2014; Linard *et al.* 2015) and biodiversity monitoring (e.g. bees) (Tang *et al.* 2014, 2015). These and similar studies demonstrate the potential for macrobial metagenome sequencing to assemble correctly mitochondrial genomes from complex samples comprising hundreds of specimens, in many cases from closely related species (Tang *et al.* 2015).

A potential difficultly when applying mito-metagenomics is that only a small fraction of the total data is assigned to mitochondrial reads, while the majority of reads is taken up by the nuclear DNA (>99% of the data) (Zhou *et al.* 2013; Liu *et al.* 2016). In order to enhance the presence of the mitochondria data in the samples, mitochondrial enrichment

through differential centrifugation has been used, but the increase in the proportion of reads assigned to mitochondria was not substantial (Zhou *et al.* 2013). More recently, Liu *et al.* (2016), tested the use of capture probes, which increased the contribution of the mitochondrial DNA in the total reads by 100-fold, presenting a more viable option for effective mitochondrial enrichment. Overall, mito-metagenomics could improve the field of ecological monitoring even further and advance HTS applications by removing PCR related biases and artefacts and improve utilization of multi-locus methodologies based on complete mitochondrial genomes (Liu *et al.* 2016; Crampton-Platt *et al.* 2016).

1.7 Using environmental DNA (eDNA) for monitoring and conservation

Environmental DNA (eDNA) is DNA extracted directly from environmental samples, without prior isolation of a particular organism (Lodge *et al.* 2012). Although the term is originally derived from microbiological studies (Ogram *et al.* 1987), eDNA has been sequenced from a multitude of different types of environments. These include terrestrial and aquatic sediments (e.g. Pawlowski *et al.* 2011; Andersen *et al.* 2012), ice cores (Willerslev *et al.* 2007), freshwater (e.g. Jerde *et al.* 2011; Goldberg *et al.* 2011; Dejean *et al.* 2012), and seawater (e.g. Foote *et al.* 2012; Thomsen *et al.* 2012a; Kelly *et al.* 2014).

A particular interest has been developed for aquatic ecosystems, which triggered the emergence of a new field of aquatic monitoring which uses eDNA to target macrobial organisms, such as animals and plants, which are in many cases important conservation species (invasive or endangered) (Turner *et al.* 2014). The onset of the freshwater macrobial eDNA monitoring field was made by Ficetola *et al.* (2008) who performed an exploratory study for the detection of an amphibian species from water samples and suggested that “the environment can retain the molecular imprint of inhabiting species”, which could prove useful for biodiversity assessment. Soon after that more studies emerged which applied eDNA as a means of diversity detection for monitoring and conservation purposes (e.g. Dejean *et al.* 2011; Darling & Mahon 2011; Thomsen *et al.* 2012b; Minamoto *et al.* 2012).

Whilst the number of studies applying eDNA methodologies has been rapidly increasing, several aspects of the “ecology” of eDNA have not yet been described sufficiently. Some

of these characteristics include, the origin of eDNA, the time that eDNA remains detectable after it is released in the environment, the state of the molecules and its transportation capability (Barnes & Turner 2016) (Figure 1.2).

The physical identity of eDNA varies, as well as the origin. It is generally suggested that eDNA includes both intracellular and extracellular forms, which co-exist in the environmental sample (Creer *et al.* 2016; Barnes & Turner 2016), though Turner *et al.* (2014) suggest that eDNA is predominantly found inside cells or mitochondria. Identifying the nature of eDNA would assist in also defining its degradation and settling rates, which can influence the fate of eDNA (see below) (Turner *et al.* 2014). Sources of eDNA include excretions and reproductive fluids (urine, faeces, sperm), shed skin cells and decomposing matter (Barnes & Turner 2016). The amount of eDNA that is released could also vary depending on the biomass of the organisms, life stage, variations in their metabolic rate or ambient temperature (Maruyama *et al.* 2014; Lacoursière-Roussel *et al.* 2016).

Even after the removal of the organism from the environment (transfer, death, adult emergence), eDNA still remains detectable for a period of time, which is known as eDNA persistence time (Dejean *et al.* 2011). The persistence time can vary largely depending on three main types of parameters, the characteristics of the DNA molecules, the biotic conditions, and the abiotic conditions (Barnes & Turner 2016). Environmental DNA persistence time has been studied in different experimental settings, including aquaria (Dejean *et al.* 2011; Goldberg *et al.* 2013), mesocosms (Thomsen *et al.* 2012a), but not many studies have investigated the fate of eDNA in the wild (Barnes *et al.* 2014). Estimations of persistence time vary greatly between studies with some studies suggesting fast degradation rates whilst others advocate that eDNA is still detectable after several weeks. For example, studies have suggested that eDNA is detectable for 7-14 days (Thomsen *et al.* 2012b), 21 days (Goldberg *et al.* 2013), or 17-25 days (Dejean *et al.* 2011), while Strickler *et al.* (2015) report rapid degradation during the first 3-10 days, though the eDNA was still detectable after 58 days. The accuracy of estimations of the persistence time can in turn influence the reliability of the detection results, hence correct determination of persistence time for eDNA, is vital in order to assure that the overall results reflect the contemporary state of the diversity (Thomsen & Willerslev 2015).

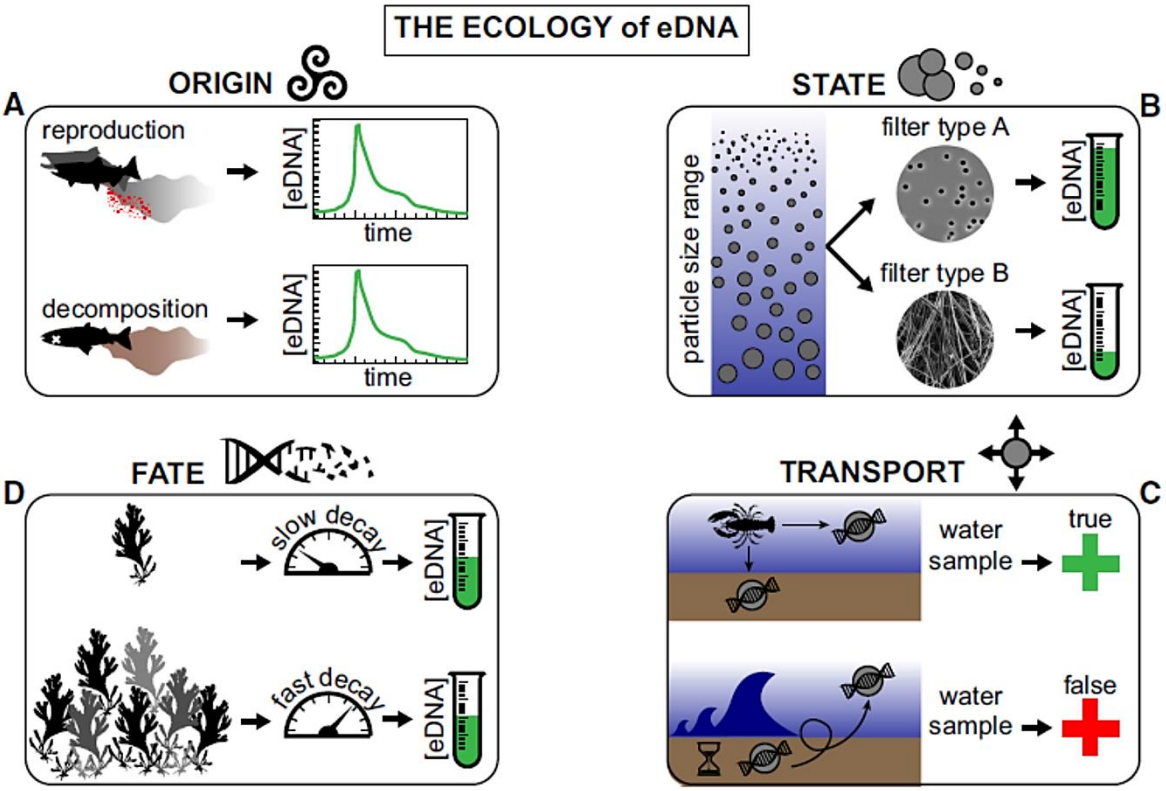


Figure 1.2: Facets of the ecology of environmental DNA (eDNA).
A. different possible origins of eDNA, B. state (particle size, cellular, extracellular), C. fate - persistence time, D. transportation distance and sediment binding, from Barnes & Turner 2016)

Generally, two types of eDNA surveys exist: from one side those which attempt to prove the presence of a particular species and on the other side those which attempt to catalogue the diversity of species in the studied area (Ficetola *et al.* 2015). The majority of taxon specific work is usually performed through PCR/qPCR detection in the form of presence absence surveys using specifically designed assays (e.g. Goldberg *et al.* 2013; Biggs *et al.* 2015; Wilcox *et al.* 2016; Padgett-Stewart *et al.* 2016). More recently, metabarcoding of eDNA has been used for community detection targeting fish and amphibians from freshwater (Evans *et al.* 2016; Valentini *et al.* 2016; Shaw *et al.* 2016; Hänfling *et al.* 2016) and marine (Port *et al.* 2016) habitats. Only two studies have been done performing metabarcoding for particular invertebrate species (Thomsen *et al.*

2012b; Deiner *et al.* 2015) though other studies have targeted particular macroinvertebrate species through PCR detection (Mächler *et al.* 2014; Deiner & Altermatt 2014). Overall, this approach once optimised provides a quick, cost-effective and standardised means of obtaining species distribution and potentially abundance data using only water samples (Thomsen et al. 2012) and could be used for population surveillance and monitoring and for multiple-species metagenetic detection (Lodge et al. 2012).

1.8 Aims and outline of the thesis

The principal aim of this thesis was to identify and explore innovative DNA based applications for advancing biomonitoring of freshwater ecosystems using macroinvertebrates as a target group. This aim was explored through three main directions and methods, including DNA Barcoding of individual specimens, metabarcoding of community DNA and eDNA samples and mitochondrial metagenomic sequencing. The thesis is divided into the following five chapters.

Chapter 1

The current state of the science of biomonitoring is introduced and an overview is given on traditional monitoring applications. Furthermore, current DNA-based developments and their possible applications in the field of ecological assessment and biodiversity monitoring are discussed, whilst identifying areas where improvement is needed.

Chapter 2

The main aim of this chapter was the construction of a DNA Barcode reference library for UK freshwater macroinvertebrate species, with a particular interest in ecological indicator species. To collect this DNA information, three main groups of macroinvertebrates were targeted, from the orders Trichoptera, Gastropoda and Diptera (Chironomidae). The levels of divergence of these species and the effectiveness of DNA Barcoding for species identification were explored and findings were placed within existing literature.

Chapter 3

The aim of this chapter was to test the use of environmental DNA metabarcoding for the characterisation of the extant diversity in a temperate lake ecosystem (Llyn Padarn, N. Wales), through an annual cycle of collected water and invertebrate samples. To allow comparison with contemporary diversity in the lake a target taxon was used (Chironomidae), which was sequenced for two COI amplicons. Overall, we looked into fine levels of biodiversity detection, including species richness and community composition variations along with implications of the persistence of eDNA related to the methodologies employed, such as fragment length, and sequencing depth.

Chapter 4

The main aim of this chapter was to compare the accuracy of metabarcoding (PCR-based) vs. shotgun mito-genomic sequencing (PCR-free) of bulk macroinvertebrate communities, with a particular interest in their efficiency for the estimation of relative species abundance. To achieve this aim, a structured design of mock macroinvertebrate communities was employed containing macroinvertebrate specimens of known biomass, which where sequenced with both approaches.

Chapter 5

This chapter presents and overall synthesis of the most important findings throughout the thesis. Furthermore, future perspectives from the application of this work and potential limitations are discussed.

References

Andersen, K., Bird, K.L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjær, K.H., Orlando, L., Gilbert, M.T.P. & Willerslev, E. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Armitage, P.D., Moss, D., Wright, J.F. & Furse, M.T. (1983). The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. *Water Research*, **17**, 333–347.

Armitage, P.D., Pinder, L.C. & Cranston, P. (2012). *The Chironomidae: biology and ecology of non-biting midges*. Chapman and Hall, London.

Arscott, D.B., Jackson, J.K. & Kratzer, E.B. (2006). Role of rarity and taxonomic resolution in a regional and spatial analysis of stream macroinvertebrates. *Journal of the North American Benthological Society*, **25**, 977–997.

Avise, J.C. (2004). *Molecular markers, natural history, and evolution*, 2nd edn. Sinauer Associates, Sunderland (Massachusetts).

Bailey, R.C., Norris, R.H. & Reynoldson, T.B. Taxonomic resolution of benthic macroinvertebrate communities in bioassessments. *J. N. Am. Benthol. Soc.*, **20**, 280–286.

Baird, D.J. & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.

Baird, D.J. & Sweeney, B.W. (2011). Applying DNA barcoding in benthology: the state of the science. *Journal of the North American Benthological Society*, **30**, 122–124.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E. & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Molecular Ecology Resources*, **14**, 306–323.

Barnes, M.A. & Turner, C.R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, **17**, 1–17.

Barnes, M.A., Turner, C.R., Jerde, C.L., Renshaw, M.A., Chadderton, W.L. & Lodge, D.M. (2014). Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science and Technology*, **48**, 1819–1827.

Barrett, R.D.H. & Hebert, P.D.N. (2005). Identifiying spiders through DNA barcodes. *Canadian Journal of Zoology*, **83**, 481–491.

Bensasson, D., Zhang, D., Hartl, D.L. & Hewitt, G.M. (2001). Mitochondrial pseudogenes : evolution ' s misplaced witnesses. *TRENDS in Ecology & Evolution*, **16**, 314–321.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R.A., Foster, J.,

Wilkinson, J.W., Arnell, A., Brotherton, P., Williams, P. & Dunn, F. (2015). Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). *Biological Conservation*, **183**, 19–28.

Bonada, N., Prat, N., Resh, V.H. & Statzner, B. (2006). DEVELOPMENTS IN AQUATIC INSECT BIOMONITORING: A Comparative Analysis of Recent Approaches. *Annual Review of Entomology*, **51**, 495–523.

Boyer, S.L., Howe, A.A., Juergens, N.W. & Hove, M.C. (2011). A DNA-barcoding approach to identifying juvenile freshwater mussels (Bivalvia:Unionidae) recovered from naturally infested fishes. *Journal of the North American Benthological Society*, **30**, 182–194.

Brodin, Y., Ejdung, G., Strandberg, J. & Lyrholm, T. (2013). Improving environmental and biodiversity monitoring in the Baltic Sea using DNA barcoding of Chironomidae (Diptera). *Molecular Ecology Resources*, **13**, 996–1004.

du Buy, H.G. & Riley, F.L. (1967). Hybridizaton between the nuclear and kinetoplasr DNA's of Leishmania enriettii and between nuclear and mitochondrial DNA's of mouse liver. *Proceedings of the National Academy of Sciences of the United States of America*, **57**, 790–7.

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., A.Wardle, D., Kinzig, A.P., Daily, G.C., Loreau, M., Grace, J.B., Larigauderie, A., Srivastava, D.S. & Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, **489**, 326–326.

Carew, M.E., Pettigrove, V. & Hoffmann, a. a. (2005). The Utility of DNA Markers in Classical Taxonomy: Using Cytochrome Oxidase I Markers to Differentiate Australian <I>Cladopelma</I> (Diptera: Chironomidae) Midges. *Annals of the Entomological Society of America*, **98**, 587–594.

Carew, M., Pettigrove, V., Metzeling, L. & Hoffmann, A. (2013). Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology*, **10**, 45.

Clare, E.L., Kerr, K.C.R., Von Königslöw, T.E., Wilson, J.J. & Hebert, P.D.N. (2008). Diagnosing mitochondrial DNA diversity: Applications of a sentinel gene approach. *Journal of Molecular Evolution*, **66**, 362–367.

Clarke, L.J., Soubrier, J., Weyrich, L.S. & Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, **14**, 1160–1170.

Clarke, R.T., Wright, J.F. & Furse, M.T. (2003). RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecological Modelling*, **160**, 219–233.

Collins, A., Ohandja, D.G., Hoare, D. & Voulvoulis, N. (2012). Implementing the Water Framework Directive: A transition from established monitoring networks in England and Wales. *Environmental Science and Policy*, **17**, 49–61.

Costa, F.O. & Carvalho, G.R. (2010). New insights into molecular evolution: Prospects from the barcode of life initiative (BOLI). *Theory in Biosciences*, **129**, 149–157.

Crampton-Platt, A., Yu, D.W., Zhou, X. & Vogler, A.P. (2016). Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience*, **5**, 15.

Cranston, P.S. (1990). Biomonitoring and invertebrate taxonomy. *Environmental Monitoring and Assessment*, **14**, 265–273.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W.K., Potter, C. & Bik, H.M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, **56**, 68–74.

Creer, S., Fonseca, V.G., Porazinska, D.L., Giblin-Davis, R.M., Sung, W., Power, D.M., Packer, M., Carvalho, G.R., Blaxter, M.L., Lambshead, P.J.D. & Thomas, W.K. (2010). Ultrasequencing of the meiofaunal biosphere: Practice, pitfalls and promises. *Molecular Ecology*, **19**, 4–20.

Darling, J.A. & Mahon, A.R. (2011). From molecules to management: Adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environmental Research*, **111**, 978–988.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., Taberlet, P., Taberlet, P., Coissac, E., Hajibabaei, M., Rieseberg, L., Yu, D., Ji, Y., Emerson, B., Wang, X., Ye, C., Yang, C., Ding, Z., Ficetola, G., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessiere, J., Taberlet, P., Pompanon, F., Geller, J., Meyer, C., Parker, M., Hawk, H., Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glockner, F., Bru, D., Martin-Laurent, F., Philippot, L., Schloss, P., Gevers, D., Westcott, S., Clarke, L., Soubrier, J., Weyrich, L., Cooper, A., Ji, Y., Barba, M. De, Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., Taberlet, P., Leray, M., Yang, J., Meyer, C., Mills, S., Agudelo, N., Ranwez, V., Boehm, J., Machida, R., Little, D., Deagle, B., Kirkwood, R., Jarman, S., Zhou, X., Shokralla, S., Gibson, J., Nikbakht, H., Janzen, D., Hallwachs, W. & Hajibabaei, M. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology letters*, **10**, 1789–1793.

Deiner, K. & Altermatt, F. (2014). Transport distance of invertebrate environmental DNA in a natural river. *PLoS ONE*, **9**, e88786.

Deiner, K., Walser, J.C., Mächler, E. & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, **183**, 53–63.

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P. & Miaud, C. (2011). Persistence of environmental DNA in freshwater ecosystems. *PLoS ONE*, **6**, e23398.

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E. & Miaud, C. (2012). Improved detection of an alien invasive species through environmental DNA barcoding: The example of the American bullfrog Lithobates catesbeianus. *Journal of Applied Ecology*, **49**, 953–959.

DeSalle, R. (2007). Phenetic and DNA taxonomy; a comment on Waugh. *BioEssays*, **29**, 1289–1290.

DeWalt, R.E. (2011). DNA barcoding: a taxonomic point of view. *Journal of the North American Benthological Society*, **30**, 174–181.

Dunigan, E.P. (1988). *Biological Indicators of Freshwater Pollution and Environmental Management*. Springer Netherlands, Dordrecht.

Ekrem, T., Stur, E. & Hebert, P.D.N. (2010). Females do count: Documenting chironomidae (Diptera) species diversity using DNA barcoding. *Organisms Diversity and Evolution*, **10**, 397–408.

Evans, N.T., Olds, B.P., Renshaw, M.A., Turner, C.R., Li, Y., Jerde, C.L., Mahon, A.R., Pfrender, M.E., Lamberti, G.A. & Lodge, D.M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, **16**, 29–41.

Ferrington JLC, Blackwood MA, Wright CA, Crisp NH, Kavanaugh JL, Sc.F. (1991). A protocol for using surface-floating pupal exuviae of Chironomidae for rapid bio assessment of changing water quality. *Sediment and stream water quality in a changing environment: trends and explanation.* (ed P.N.& W. DE), pp. 181–190. International Association of Hydrological Sciences Publication.

Ficetola, G.F.G.F., Miaud, C., Pompanon, F. & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, **4**, 423–425.

Ficetola, G.F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C.M., Boyer, F., Pompanon, F., Rayé, G. & Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, **15**, 543–556.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

Fonseca, V.G., Carvalho, G.R., Sung, W., Johnson, H.F., Power, D.M., Neill, S.P., Packer, M., Blaxter, M.L., Lambshead, P.J.D., Thomas, W.K. & Creer, S. (2010). Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature communications*, **1**, 98.

Fonseca, V.G., Nichols, B., Lallias, D., Quince, C., Carvalho, G.R., Power, D.M. & Creer, S. (2012). Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. *Nucleic Acids Research*, **40**, e66–e66.

Foote, A.D., Thomsen, P.F., Sveegaard, S., Wahlberg, M., Kielgast, J., Kyhn, L.A., Salling, A.B., Galatius, A., Orlando, L. & Gilbert, M.T.P. (2012). Investigating the Potential Use of Environmental DNA (eDNA) for Genetic Monitoring of Marine Mammals. *PLoS ONE*, **7**, e41781.

Friberg, N., Bonada, N., Bradley, D.C., Dunbar, M.J., Edwards, F.K., Grey, J., Hayes, R.B.,

Hildrew, A.G., Lamouroux, N., Trimmer, M. & Woodward, G. (2011). Biomonitoring of Human Impacts in Freshwater Ecosystems. The Good, the Bad and the Ugly. *Advances in Ecological Research*, pp. 1–68.

Gibson, J.F., Shokralla, S., Curry, C., Baird, D.J., Monk, W.A., King, I. & Hajibabaei, M. (2015). Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS ONE*, **10**, 1–15.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 8007–12.

Gillett, C.P.D.T., Crampton-Platt, A., Timmermans, M.J.T.N., Jordal, B.H., Emerson, B.C. & Vogler, A.P. (2014). Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, **31**, 2223–2237.

Goldberg, C.S., Pilliod, D.S., Arkle, R.S. & Waits, L.P. (2011). Molecular detection of vertebrates in stream water: A demonstration using rocky mountain tailed frogs and Idaho giant salamanders (B. Gratwicke, Ed.). *PLoS ONE*, **6**, e22746.

Goldberg, C.S., Sepulveda, A., Ray, A., Baumgardt, J. & Waits, L.P. (2013). Environmental DNA as a new method for early detection of New Zealand mudsnails (*Potamopyrgus antipodarum*). *Freshwater Science*, **32**, 792–800.

Gray, C., Bista, I., Creer, S., Demars, B.O.L., Falciani, F., Don, T.M., Sun, X. & Woodward, G. (2015). Freshwater conservation and biomonitoring of structure and function: Genes to ecosystems. *Aquatic Functional Biodiversity: An Ecological and Evolutionary Perspective* (eds A. Belgrano, G. Woodward & U. Jacob), pp. 241–271. Elsevier.

Griffiths, M. (2002). The European Water Framework Directive: An Approach to Integrated River Basin Management. *European Water Management Online*, 1–15.

Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W. & Hebert, P.D.N. (2006a). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 968–971.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C. & Baird, D.J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, **6**, e17497.

Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J., Whitfield, J.B. & Hebert, P.D.N. (2006b). A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, **6**, 959–964.

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

Hatton-Ellis, T. (2008). The Hitchhiker's guide to the Water Framework Directive. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **18**, 111–116.

Hawkes, A, H. (1998). Origin and development of the biological monitoring working party score system. *Water Research*, **32**, 964–968.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & deWaard, J.R. (2003a). Biological identifications through DNA barcodes. *Proceedings. Biological sciences / The Royal Society*, **270**, 313–21.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H. & Hallwachs, W. (2004a). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14812–14817.

Hebert, P.D.N., Ratnasingham, S. & Waard, J. (2003b). Barcoding animal life : cytochrome c oxidase subunit 1 divergences among closely related species Barcoding animal life : cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B*, **270**, S96–S99.

Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S. & Francis, C.M. (2004b). Identification of birds through DNA barcodes. *PLoS Biology*, **2**, e312.

Heino, J. (2013). The importance of metacommunity ecology for environmental assessment research in the freshwater realm. *Biological Reviews*, **88**, 166–178.

Hogg, I.D. & Hebert, P.D.N. (2004). Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Canadian Journal of Zoology*, **82**, 749–754.

Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L., Fazekas, A.J., Graham, S.W., James, K.E., Kim, K.-J., Kress, W.J., Schneider, H., van AlphenStahl, J., Barrett, S.C.H., van den Berg, C., Bogarin, D., Burgess, K.S., Cameron, K.M., Carine, M., ChacÃ³n, J., Clark, A., Clarkson, J.J., Conrad, F., Devey, D.S., Ford, C.S., Hedderson, T.A.J., Hollingsworth, M.L., Husband, B.C., Kelly, L.J., Kesanakurti, P.R., Kim, J.S., Kim, Y.-D., Lahaye, R., Lee, H.-L., Long, D.G., MadriÃ±Ã¡n, S., Maurin, O., Meusnier, I., Newmaster, S.G., Park, C.-W., Percy, D.M., Petersen, G., Richardson, J.E., Salazar, G.A., Savolainen, V., Seberg, O., Wilkinson, M.J., Yi, D.-K. & Little, D.P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, **106**.

Howarth, W. (2009). Aspirations and realities under the water framework directive: Proceduralisation, participation and practicalities. *Journal of Environmental Law*, **21**, 391–417.

Jerde, C.L., Mahon, A.R., Chadderton, W.L. & Lodge, D.M. (2011). 'Sight-unseen' detection of rare aquatic species using environmental DNA. *Conservation Letters*, **4**, 150–157.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P., Edwards, F.A., Larsen, T.H., Hsu, W.W., Benedick, S., Hamer, K.C., Wilcove, D.S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B.C. & Yu, D.W. (2013).

Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Johnson, N.K. & Cicero, C. (2004). New mitochondrial DNA data affirm the importance of Pleistocene speciation in North American birds. *Evolution; international journal of organic evolution*, **58**, 1122–30.

Joly, S., Davies, T.J., Archambault, A., Bruneau, A., Derry, A., Kembel, S.W., Peres-Neto, P., Vamosi, J. & Wheeler, T.A. (2014). Ecology in the age of DNA barcoding: The resource, the promise and the challenges ahead. *Molecular Ecology Resources*, **14**, 221–232.

Jones, F.C. (2008). Taxonomic sufficiency: The influence of taxonomic resolution on freshwater bioassessments using benthic macroinvertebrates. *Environmental Reviews*, **16**, 45–69.

Jørgensen, T., Haile, J., Möller, P., Andreev, A., Boessenkool, S., Rasmussen, M., Kienast, F., Coissac, E., Taberlet, P., Brochmann, C., Bigelow, N.H., Andersen, K., Orlando, L., Gilbert, M.T.P. & Willerslev, E. (2012). A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Molecular Ecology*, **21**, 1989–2003.

Kelly, R.P., Port, J.A., Yamahara, K.M. & Crowder, L.B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS ONE*, **9**, e86175.

Kenney, M.A., Sutton-Grier, A.E., Smith, R.F. & Gresens, S.E. (2009). Benthic macroinvertebrates as indicators of water quality: The intersection of science and policy. *Terrestrial Arthropod Reviews*, **2**, 99–128.

Kille, P. (2011). *A review of molecular techniques for ecological monitoring*. Environment Agency, Bristol, UK.

Kim, S., Song, K.H., Ree, H. Il & Kim, W. (2012). A DNA barcode library for Korean Chironomidae (Insecta: Diptera) and indexes for defining barcode gap. *Molecules and Cells*, **33**, 9–17.

Lacoursière-Roussel, A., Rosabal, M. & Bernatchez, L. (2016). Estimating fish abundance and biomass from eDNA concentrations: variability among capture methods and environmental conditions. *Molecular Ecology Resources*, **16**, 1401–1414.

Lallias, D., Hiddink, J.G., Fonseca, V.G., Gaspar, J.M., Sung, W., Neill, S.P., Barnes, N., Ferrero, T., Hall, N., Lambshead, P.J.D., Packer, M., Thomas, W.K. & Creer, S. (2015). Environmental metabarcoding reveals heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine ecosystems. *The ISME journal*, **9**, 1208–21.

Lenat, D.R. & Resh, V.H. (2001). Taxonomy and stream ecology - The benefits of genus- and species-level identifications. *J. N. Am. Benthol. Soc.*, **20**, 287–298.

Lenobah, D.E., Dileep, A., Chandrasekhar, K., Sreevani, S. & Kumari, J.P. (2014). DNA barcoding on bacteria: A Review. *Advances in Biology*, **2014**, 1–10.

Leray, M. & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National*

Academy of Sciences, **2014**, 201424997.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. & Machida, R.J. (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**, 34.

Linard, B., Crampton-Platt, A., Gillett, C.P.D.T., Timmermans, M.J.T.N. & Vogler, A.P. (2015). Metagenome skimming of insect specimen pools: Potential for comparative genomics. *Genome Biology and Evolution*, **7**, 1474–1489.

Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., Zhang, H., Misof, B., Kjer, K.M., Tang, M., Niehuis, O., Jiang, H. & Zhou, X. (2016). Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, **16**, 470–479.

Lodge, D.M., Turner, C.R., Jerde, C.L., Barnes, M.A., Chadderton, L., Egan, S.P., Feder, J.L., Mahon, A.R. & Pfrender, M.E. (2012). Conservation in a cup of water: Estimating biodiversity and population abundance from environmental DNA. *Molecular Ecology*, **21**, 2555–2558.

Logan, P. (2001). Ecological quality assessment of rivers and integrated catchment management in England and Wales. *J. Limnol.*, **60**, 25–32.

Mächler, E., Deiner, K., Steinmann, P. & Altermatt, F. (2014). Utility of environmental DNA for monitoring rare and indicator macroinvertebrate species. *Freshwater Science*, **33**, 1174–1183.

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M. & Minamoto, T. (2014). The release rate of environmental DNA from juvenile and adult fish. *PLoS ONE*, **9**, e114639.

Matthews, R.A., Buikema, A.L., Cairns, J. & Rodgers, J.H. (1982). Biological monitoring. Part IIA-receiving system functional methods, relationships and indices. *Water Research*, **16**, 129–139.

Meier, R., Shiyang, K., Vaidya, G. & Ng, P.K.L. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic biology*, **55**, 715–728.

Meusnier, I., Singer, G.A.C., Landry, J.-F., Hickey, D.A., Hebert, P.D.N. & Hajibabaei, M. (2008). A Universal DNA Mini-barcode for Biodiversity Analysis. *BMC Genomics*, **9**, 214.

Meyer, C.P. & Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology*, **3**, 1–10.

Min, X.J. & Hickey, D.A. (2007). DNA barcodes provide a quick preview of mitochondrial genome composition. *PLoS ONE*, **2**, e325.

Minamoto, T., Yamanaka, H., Takahara, T., Honjo, M.N. & Kawabata, Z. (2012). Surveillance of fish species composition using environmental DNA. *Limnology*, **13**,

193–197.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J.Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M. & Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science*, **2**, 150088.

Moritz, C. & Cicero, C. (2004). DNA barcoding: Promise and pitfalls. *PLoS Biology*, **2**, e354.

Moulton, M.J., Song, H. & Whiting, M.F. (2010). Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: A case study from Orthoptera (Arthropoda: Insecta). *Molecular Ecology Resources*, **10**, 615–627.

Munch, K., Boomsma, W., Huelsenbeck, J.P., Willerslev, E. & Nielsen, R. (2008a). Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, **57**, 750–757.

Munch, K., Boomsma, W., Willerslev, E. & Nielsen, R. (2008b). Fast phylogenetic DNA barcoding. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **363**, 3997–4002.

Murray, D.C., Coghlan, M.L. & Bunce, M. (2015). From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. *PLoS ONE*, **10**, e0124671.

Ogram, A., Sayler, G.S. & Barkay, T. (1987). The extraction and purification of microbial DNA from sediments. *Journal of Microbiological Methods*, **7**, 57–66.

Padgett-Stewart, T.M., Wilcox, T.M., Carim, K.J., McKelvey, K.S., Young, M.K. & Schwartz, M.K. (2016). An eDNA assay for river otter detection: a tool for surveying a semi-aquatic mammal. *Conservation Genetics Resources*, **8**, 5–7.

Pawlowski, J., Christen, R., Lecroq, B., Bachar, D., Shahbazkia, H.R., Amaral-Zettler, L. & Guillou, L. (2011). Eukaryotic richness in the abyss: Insights from pyrotag sequencing. *PLoS ONE*, **6**, e18169.

Pfenninger, M., Nowak, C., Kley, C., Steinke, D. & Streit, B. (2007). Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic Chironomus (Diptera) species. *Molecular Ecology*, **16**, 1957–1968.

Pfrender, M., Hawkins, C., Bagley, M., Courtney, G., Creutzburg, B., Epler, J., Fend, S., Ferrington, L., Hartzell, P., Jackson, S., Larsen, D., Lvesque, C.A., Morse, J., Petersen, M., Ruiter, D., Schindel, D. & Whiting, M. (2010). Assessing Macroinvertebrate Biodiversity in Freshwater Ecosystems: Advances and Challenges in DNA-based Approaches The Quarterly Review of Biology. *Source: The Quarterly Review of Biology*, **85**, 319–340.

Pilgrim, E.M., Jackson, S. a., Swenson, S., Turcsanyi, I., Friedman, E., Weigt, L. & Bagley, M.J. (2011). Incorporation of DNA barcoding into a large-scale biomonitoring program: opportunities and pitfalls. *Journal of the North American Benthological Society*, **30**, 217–231.

Piñol, J., Mir, G., Gomez-Polo, P. & 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**, 819–830.

Port, J.A., O'Donnell, J.L., Romero-Maraccini, O.C., Leary, P.R., Litvin, S.Y., Nickols, K.J., Yamahara, K.M. & Kelly, R.P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, **25**, 527–541.

Reynoldson, T.B. & Metcalfe-Smith, J.L. (1992). An overview of the assessment of aquatic ecosystem health using benthic invertebrates. *Journal of Aquatic Ecosystem Health*, **1**, 295–308.

Salinas-Ramos, V.B., Herrera Montalvo, L.G., León-Regagnon, V., Arrizabalaga-Escudero, A. & Clare, E.L. (2015). Dietary overlap and seasonality in three species of mormoopid bats from a tropical dry forest. *Molecular Ecology*, **24**, 5296–5307.

Schmidt-Kloiber, A. & Nijboer, R.C. (2004). The effect of taxonomic resolution on the assessment of ecological water quality classes. *Hydrobiologia*, **516**, 269–283.

Schnell, I.B., Thomsen, P.F., Wilkinson, N., Rasmussen, M., Jensen, L.R.D., Willerslev, E., Bertelsen, M.F. & Gilbert, M.T.P. (2012). Erratum: Screening mammal biodiversity using dna from leeches (Current Biology (2012) 22 (R262-R263)). *Current Biology*, **22**, 1980.

Schoch, C.L., Seifert, K. a., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C. a., Chen, W., Consortium, F.B., Bolchacova, E., Voigt, K., Crous, P.W., Miller, a. N., Wingfield, M.J., Aime, M.C., An, K.-D., Bai, F.-Y., Barreto, R.W., Begerow, D., Bergeron, M.-J., Blackwell, M., Boekhout, T., Bogale, M., Boonyuen, N., Burgaz, a. R., Buyck, B., Cai, L., Cai, Q., Cardinali, G., Chaverri, P., Coppins, B.J., Crespo, A., Cubas, P., Cummings, C., Damm, U., De Beer, Z.W., de Hoog, G.S., Del-Prado, R., Dentinger, B., Dieguez-Uribeondo, J., Divakar, P.K., Douglas, B., Duenas, M., Duong, T. a., Eberhardt, U., Edwards, J.E., Elshahed, M.S., Fliegerova, K., Furtado, M., Garcia, M. a., Ge, Z.-W., Griffith, G.W., Griffiths, K., Groenewald, J.Z., Groenewald, M., Grube, M., Gryzenhout, M., Guo, L.-D., Hagen, F., Hambleton, S., Hamelin, R.C., Hansen, K., Harrold, P., Heller, G., Herrera, C., Hirayama, K., Hirooka, Y., Ho, H.-M., Hoffmann, K., Hofstetter, V., Hognabba, F., Hollingsworth, P.M., Hong, S.-B.S.-B.S.-B.S.-B., Hosaka, K., Houbraken, J., Hughes, K., Huhtinen, S., Hyde, K.D., James, T., Johnson, E.M., Johnson, J.E., Johnston, P.R., Jones, E.B.G., Kelly, L.J., Kirk, P.M., Knapp, D.G., Koljalg, U., Kovacs, G.M., Kurtzman, C.P., Landvik, S., Leavitt, S.D., Liggenstoffer, a. S., Liimatainen, K., Lombard, L., Luangsa-ard, J.J., Lumbsch, H.T., Maganti, H., Maharachchikumbura, S.S.N., Martin, M.P., May, T.W., McTaggart, a. R., Methven, a. S., Meyer, W., Moncalvo, J.-M., Mongkolsamrit, S., Nagy, L.G., Nilsson, R.H., Niskanen, T., Nyilasi, I., Okada, G., Okane, I., Olariaga, I., Otte, J., Papp, T., Park, D., Petkovits, T., Pino-Bodas, R., Quaedvlieg, W., Raja, H. a., Redecker, D., Rintoul, T.L., Ruibal, C., Sarmiento-Ramirez, J.M., Schmitt, I., Schussler, A., Shearer, C., Sotome, K., Stefani, F.O.P., Stenroos, S., Stielow, B., Stockinger, H., Suetrong, S., Suh, S.-O., Sung, G.-H., Suzuki, M., Tanaka, K., Tedersoo, L., Telleria, M.T., Tretter, E., Untereiner, W. a., Urbina, H., Vagvolgyi, C., Vialle, A., Vu, T.D., Walther, G., Wang, Q.-M., Wang, Y., Weir, B.S.,

Weiss, M., White, M.M., Xu, J., Yahr, R., Yang, Z.L., Yurkov, A., Zamora, J.-C., Zhang, N., Zhuang, W.-Y.W.-Y. & Schindel, D. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 1–6.

Sharley, D.J., Pettigrove, V. & Parsons, Y.M. (2004). Molecular identification of Chironomus spp. (Diptera) for biomonitoring of aquatic ecosystems. *Australian Journal of Entomology*, **43**, 359–365.

Shaw, J.L.A., Clarke, L.J., Wedderburn, S.D., Barnes, T.C., Weyrich, L.S. & Cooper, A. (2016). Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biological Conservation*, **197**, 131–138.

Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., Golding, G.B. & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports*, **5**, 9687.

Sinniger, F., Pawlowski, J., Harii, S., Gooday, A.J., Yamamoto, H., Chevaldonné, P., Cedhagen, T., Carvalho, G. & Creer, S. (2016). Worldwide analysis of sedimentary DNA reveals major gaps in taxonomic knowledge of deep-sea benthos. *Frontiers in Marine Science*, **3**, 92.

Song, H., Buhay, J.E., Whiting, M.F. & Crandall, K.A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences*, **105**, 13486–13491.

Strickler, K.M., Fremier, A.K. & Goldberg, C.S. (2015). Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biological Conservation*, **183**, 85–92.

Strong, E.E., Gargominy, O., Ponder, W.F. & Bouchet, P. (2008). Global diversity of gastropods (Gastropoda; Mollusca) in freshwater. *Hydrobiologia*, **595**, 149–166.

Sweeney, B.W., Battle, J.M., Jackson, J.K. & Dapkey, T. (2011). Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society*, **30**, 195–216.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

Tang, M., Hardman, C.J., Ji, Y., Meng, G., Liu, S., Tan, M., Yang, S., Moss, E.D., Wang, J., Yang, C., Bruce, C., Nevard, T., Potts, S.G., Zhou, X. & Yu, D.W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, **6**, 1034–1043.

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A. & Zhou, X. (2014). Multiplex sequencing of pooled mitochondrial genomes - A crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, **42**.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. & Vogler, A.P. (2003). A plea for DNA taxonomy. *Trends in Ecology and Evolution*, **18**, 70–74.

Taylor, H.R. & Harris, W.E. (2012). An emergent science on the brink of irrelevance: A review of the past 8years of DNA barcoding. *Molecular Ecology Resources*, **12**, 377–388.

Teletchea, F. (2010). After 7 years and 1000 citations: comparative assessment of the DNA barcoding and the DNA taxonomy proposals for taxonomists and non-taxonomists. *Mitochondrial DNA*, **21**, 206–226.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Møller, P.R., Rasmussen, M. & Willerslev, E. (2012a). Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples. *PLoS ONE*, **7**, e41732.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Wiuf, C., Rasmunsen, M., Gilbert, M.T.P., Orlando, L. & Willerslev, E. (2012b). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, **21**, 2565–2573.

Thomsen, P.F. & Willerslev, E. (2015). Environmental DNA - An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, **183**, 4–18.

Turner, C.R., Barnes, M.A., Xu, C.C.Y., Jones, S.E., Jerde, C.L. & Lodge, D.M. (2014). Particle size distribution and optimal capture of aqueous macrobial eDNA. *Methods in Ecology and Evolution*, **5**, 676–684.

Valentini, A., Pompanon, F. & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology and Evolution*, **24**, 110–117.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P.F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G.H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.M., Peroux, T., Crivelli, A.J., Olivier, A., Acqueberge, M., Le Brun, M., M??ller, P.R., Willerslev, E. & Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, **25**, 929–942.

Vernooy, R., Haribabu, E., Muller, M.R., Vogel, J.H., Hebert, P.D.N., Schindel, D.E., Shimura, J. & Singer, G.A.C. (2010). Barcoding life to conserve biological diversity: Beyond the taxonomic imperative. *PLoS Biology*, **8**, e1000417.

Vogler, A.P. & Monaghan, M.T. (2007). Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research*, **45**, 1–10.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R. & Hebert, P.D.N. (2005). DNA barcoding Australia's fish species. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **360**, 1847–1857.

Webb, J.M., Jacobus, L.M., Funk, D.H., Zhou, X., Kondratieff, B., Geraci, C.J., DeWalt, R.E., Baird, D.J., Richard, B., Phillips, I. & Hebert, P.D.N. (2012). A DNA barcode library for North American ephemeroptera: Progress and prospects. *PLoS ONE*, **7**, e38063.

Wiemers, M. & Fiedler, K. (2007). Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in zoology*, **4**, 8.

Wilcox, T.M., McKelvey, K.S., Young, M.K., Sepulveda, A.J., Shepard, B.B., Jane, S.F., Whiteley, A.R., Lowe, W.H. & Schwartz, M.K. (2016). Understanding environmental DNA detection probabilities: A case study using a stream-dwelling char Salvelinus fontinalis. *Biological Conservation*, **194**, 209–216.

Will, K.W., Mishler, B.D. & Wheeler, Q.D. (2005). The Perils of DNA Barcoding and the Need for Integrative Taxonomy. *Systematic Biology*, **54**, 844–851.

Willerslev, E., Cappellini, E., Boomsma, W., Nielsen, R., Hebsgaard, M.B., Brand, T.B., Hofreiter, M., Bunce, M., Poinar, H.N., Dahl-Jensen, D., Johnsen, S., Steffensen, J.P., Bennike, O., Schwenninger, J.-L., Nathan, R., Armitage, S., de Hoog, C.-J., Alfimov, V., Christl, M., Beer, J., Muscheler, R., Barker, J., Sharp, M., Penkman, K.E.H., Haile, J., Taberlet, P., Gilbert, M.T.P., Casoli, A., Campani, E. & Collins, M.J. (2007). Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science*, **317**, 111–114.

Wilson, R. & Ruse, L. (2005). *A guide to the identification of genera of chironomid pupal exuviae occurring in Britain and Ireland*. Freshwater Biological Association Special Publication no. 13.

Yoccoz, N.G. (2012). The future of environmental DNA in ecology. *Molecular Ecology*, **21**, 2031–2038.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhan, A., Bailey, S.A., Heath, D.D. & Macisaac, H.J. (2014). Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities. *Molecular Ecology Resources*, **14**, 1049–1059.

Zhang, A.B., Muster, C., Liang, H.B., Zhu, C.D., Crozier, R., Wan, P., Feng, J. & Ward, R.D. (2012). A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Molecular Ecology*, **21**, 1848–1863.

Zhou, X., Adamowicz, S.J., Jacobus, L.M., Dewalt, R.E. & Hebert, P.D. (2009). Towards a comprehensive barcode library for arctic life - Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. *Frontiers in zoology*, **6**, 30.

Zhou, X., Jacobus, L.M., DeWalt, R.E., Adamowicz, S.J. & Hebert, P.D.N. (2010). Ephemeroptera, Plecoptera, and Trichoptera fauna of Churchill (Manitoba, Canada): insights into biodiversity patterns from DNA barcoding. *Journal of the North American Benthological Society*, **29**, 814–837.

Zhou, X., Kjer, K.M. & Morse, J.C. (2007). Associating larvae and adults of Chinese Hydropsychidae caddisflies (Insecta:Trichoptera) using DNA sequences. *Journal of the North American Benthological Society*, **26**, 719–742.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J. & Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 4.

Zhou, X., Robinson, J.L., Geraci, C.J., Parker, C.R., Flint, O.S., Etnier, D.A., Ruiter, D., DeWalt, R.E., Jacobus, L.M. & Hebert, P.D.N. (2011). Accelerated construction of a regional DNA-barcode reference library: caddisflies (Trichoptera) in the Great Smoky Mountains National Park. *Journal of the North American Benthological Society*, **30**, 131–162.

Chapter 2

A DNA Barcoding library for UK freshwater macroinvertebrates

Chapter 2: A barcode reference library for UK macroinvertebrates

2.1 Abstract

The freshwater macroinvertebrates belonging to the Trichoptera, Gastropoda and Chironomidae are prominent indicator groups used for the biomonitoring of freshwater ecosystems. Despite the role of such groups in ecosystem assessments, routine morphological identification remains time-consuming, especially problematic for life history stages, and dependent upon high levels of taxonomic expertise. To expedite application of routine molecular taxonomic approaches we generated Cytochrome subunit Oxidase (COI) DNA barcodes for numerous representatives of each group. In total, 94 species were sequenced, including 55 Trichoptera, 17 Gastropoda and 22 Chironomidae species. We found that DNA barcoding can be used successfully for species identification of target species and a distinct barcoding gap was found for all groups analysed. More extensive sampling is needed to verify findings across broader taxonomic groupings. Low levels of misidentification were detected for Trichoptera and Gastropoda (5.4% and 5.5% respectively), with more increased levels for chironomids (8%). Nevertheless, elevated misidentification within the Chironomidae might be related to the presence of a species complex (*C. plumosus*). Finally, we found that the use of chironomid pupal exuviae for standard, chain termination DNA barcoding might be challenging due to the low amounts of DNA present in the exuviae. Overall, this work aimed to establish a barcode reference library for macroinvertebrate indicator species to facilitate future biomonitoring efforts.

2.2 Introduction

2.2.1 Biomonitoring of aquatic ecosystems- Limitations of traditional approaches

Biomonitoring, or bioassessment is the use of the composition of biological communities as an indicator of condition or stress of the ecosystem (Stein *et al.* 2014), and is widely applied for aquatic ecosystem monitoring across a range of taxa. Macroinvertebrates are amongst the most commonly used and most informative organisms for applied biomonitoring (Cranston 1990). The extended group of macroinvertebrates comprises a large variety of organisms, grouped by sizes larger than 0.25mm (Dunigan 1988), to differentiate for example from smaller organisms known as meiofauna (Creer *et al.* 2010) or larger organisms (macrofauna). The majority of freshwater macroinvertebrate groups consist of insects, crustaceans, gastropods and oligochaetes (Kenney *et al.* 2009).

Traditional biomonitoring requires taxonomic identification of specimens, a process that is labour intensive, and time consuming. Further constraints of current taxonomic work with macroinvertebrates include difficulties of identifying specimens to the species level due to the occurrence of immature life stages, size differences, sexual polymorphism, specimen condition etc. (Pilgrim *et al.* 2011). Also, the presence of cryptic species or incomplete taxonomic keys can hinder taxonomists' work (Sweeney *et al.* 2011). In many cases, specimens can only be identified to a coarse level which might pose problems with ecological assessments as the sensitivity to ecological stressors can vary for species of the same genus or family (Lenat & Resh 2001; Pilgrim *et al.* 2011; Sweeney *et al.* 2011).

For the management of water bodies throughout the European Union (EU), the Water Framework Directive (WFD) has been established, to provide a regulatory framework for management and conservation of aquatic ecosystems (Collins *et al.* 2012). Assessment of water bodies through the WFD is performed by comparison of the present water body condition against an expected "reference condition" (Schmidt-Kloiber & Nijboer 2004). The development of multimetric indices for this system was carried out based on species or near-species level data, but as mentioned above, the use of species identified specimens is not always possible, and so genus or family level information is often used. Nevertheless,

when genus or family level information is used in place of species, ecological site classification was found to differ in 50% and 40% of the cases respectively (Schmidt-Kloiber & Nijboer 2004).

The implications of incorrect classification of sites could be both economical (efforts to improve quality when actually unnecessary) or environmental (failing to take measures when necessary) (Schmidt-Kloiber & Nijboer 2004). Such costs suggest that difficulties with taxonomic identification from traditional methods could very much affect the outcomes of monitoring efforts; hence achieving more accuracy to the species level by the use of molecular approaches could greatly benefit assessment efforts. Some problems can be addressed by applying molecular analysis to macroinvertebrate species identification, such as DNA Barcoding (Pilgrim *et al.* 2011).

2.2.2 DNA Barcoding and invertebrate identification

DNA barcoding was first proposed by Hebert *et al.* in (2003), as a method of molecular identification of species, using a standardised mitochondrial marker, which is part of the Cytochrome Subunit Oxidase I gene (COI), henceforth known as the COI barcoding region. Since the initial development of DNA Barcoding, the field has matured to occupy the gap between traditional taxonomy and molecular systematics (Hubert & Hanner 2015). Furthermore, the expansion of the DNA Barcoding community and effort has been exponential, despite associated controversies (Costa & Carvalho 2007). Under the umbrella of the International Barcode of Life (iBOL) (Ratnasingham & Hebert 2007), the Barcode of Life Database (BOLD) systems v.4 has been fully developed, and harbouring almost 5 million barcodes to date (June 2016). One of the basic assumptions of DNA barcoding is the existence of a "barcoding gap", which is based on the assumption that the levels of intraspecific diversity are lower than the interspecific (the difference between the two constitutes the gap) (Meyer & Paulay 2005). When this assumption is correct, species delimitation through DNA barcoding is efficient (Puillandre *et al.* 2012), but this theory has also been heavily criticised (Wiemers & Fiedler 2007). To address concerns about the effectiveness of a barcoding gap based species delimitation, additional measures have been promoted such as the use of ranking systems (Costa *et al.* 2012) or threshold based analysis (Meyer & Paulay 2005).

2.2.3 Taxa used in this study

Here, we employed DNA Barcoding for sequencing of specimens from three groups of macroinvertebrates representative of freshwater biomonitoring efforts: Trichoptera, Gastropoda and Chironomidae.

2.2.3.a **Trichoptera.**

The Trichoptera order (caddis flies) comprises 45 families and about 13,000 described species (Morse 1997) (http://trichopterabol.org/). They are essential components of freshwater ecosystems and excellent bioindicators due to their high diversity and their larvae's sensitivity to pollution (Kjer et al. 2001). The increased interest in the order of Trichoptera has given rise to the Trichoptera Barcode of Life Project, which was launched in 2007, aiming to provide a comprehensive DNA Barcode library for all known caddis fly species and up to now, more than 2,779 species have been sequenced (http://trichopterabol.org/) (March 2016). The estimated number of Trichoptera species for Britain is 197 (Wallace 1991; Wiberg-Larsen 2008).

2.2.3.b Gastropoda.

Freshwater molluscs, including gastropods and bivalves, are commonly used for biomonitoring due to their high abundance, ease of collection and ease of identification (for gastropods) (Elder & Collins 1991). Currently, 4,000 species of gastropods have been described, with the highest species diversity derived from small streams, springs and groundwater systems (Strong et al. 2008). The described number of aquatic Gastropoda found in Britain is 48 (including 2 marine Pulmonates) (Anderson 2005). Despite their great importance for freshwater ecosystems, our knowledge of gastropod systematics is limited, with the majority of taxa still being unknown (Strong et al. 2008).

2.2.3.c Chironomidae.

The family Chironomidae (non-biting midges) is one of the most species rich families of aquatic invertebrates with >10,000 species (1,200 in Europe and more than 600 in the UK) (Armitage *et al.* 2012). Chironomids are very important indicators of acidification and eutrophication, especially for lake ecosystems (Ruse 2010, 2011). Despite their huge importance for biomonitoring, they tend to be overlooked during biological assessments

due to their difficulty of identification even for experienced taxonomists, and the inability to identify females or certain life stages (Ekrem *et al.* 2010; Brodin *et al.* 2013). Using the Chironomid Pupal Exuviae Technique (CPET), which involves collection of the shed exuviae of the pupae, for chironomid identification (Wilson & Ruse 2005), provides many advantages for applied biomonitoring (Wilson & Ruse 2005; Raunio *et al.* 2011). Furthermore, the CPET technique has been developed in detail for UK lake ecosystems with extensive lists of species and their ecological attributes (Ruse 2013).

2.2.4 Connecting DNA barcoding and ecological applications

Some of the possible benefits of applied DNA barcoding include, but are not limited to, discovery of cryptic diversity and its relations with species ecological interactions, phylogenetic insights into the functional structure of communities, as well as the study of intraspecific diversity and wealth of available metadata (Joly *et al.* 2014). Therefore, ecology can greatly benefit from the development and evolutionary information content currently available by the DNA barcoding movement (Joly *et al.* 2014). Moreover, the effort related to DNA barcoding is associated with both sequencing of difficult to identify taxa, as well as targeting those which are important ecological indicators (Pilgrim *et al.* 2011).

The main purpose of the present work was to collect and sequence a range of macroinvertebrate species in order to establish a Barcode Reference Database with members of the Trichoptera, Gastropoda and Chironomidae (Diptera) groups for the UK. Members of these groups are important indicators for aquatic monitoring and in many cases very difficult to identify to the species level (e.g. Chironomidae) (Ruse 2011; Zhou *et al.* 2011). In addition, we aimed to evaluate the performance of the barcoding method for species delimitation and estimate the levels of accuracy of taxonomic identification for these groups, as well as increase the knowledge on the levels of divergence and phylogenetic relationships of the studied taxa. Furthermore, we provide novel barcoding data, which could be valuable for downstream High Throughput Sequencing (HTS) applications such as metabarcoding of eDNA or shotgun sequencing of bulk samples, and will act to the benefit of advancing biomonitoring efforts in the UK.

2.3 Methods

2.3.1 Sample collection and processing

Initially a list of indicator species of macroinvertebrates was compiled in collaboration with Environment Agency (EA) experts to identify the most ecologically relevant indicator species from the groups of Trichoptera, Gastropoda and Chironomidae (Diptera), which should be targeted during the construction of DNA Barcode reference database for UK macroinvertebrates. Existing strategies employed by the EA at the time involved preservation in IMS (Industrial Methylated Spirit), but 100% ethanol preservation was deemed necessary for achieving highest quality DNA for barcoding and avoiding the possible detrimental effects of methanol contained in IMS on extracted DNA (Stein *et al.* 2013). Additionally, we wanted to achieve a wide geographic coverage of sampled species, (≥5 specimens per species from various locations), which would allow detection of possible intraspecific diversity in barcode species without hugely increasing the number of processed specimens.

Sample collection (fresh samples) was performed during spring and summer time from 2012 through to 2014. Larvae, adults or pupal exuviae specimens were collected depending on the group (see below). Samples were acquired from direct collection by members of the Environment Agency (EA), Scottish Environment Protection Agency (SEPA) and Natural Resources Wales (NRW), volunteer taxonomists, I.B and Les Ruse (APEM Ltd.). Additional samples were also acquired from existing collections of the Centre for Ecology and Hydrology (CEH) and private collections (Hydroptilidae adult specimens).

All specimens were preserved in absolute ethanol prior to molecular analysis. Samples received from CEH were first frozen at -20°C and then preserved in 100% ethanol. For direct collection of samples (Trichoptera and Gastropoda), a sampling kit was supplied to the teams containing 100% ethanol and clean tubes of various sizes (1.5ml, 8ml, and 50ml). Sampling was performed following a Standard Operating Procedure (SOP) to ensure replicability of methods across sampling teams. Collection of benthic samples was

performed using a standard kick-net method. For a summary table of collection areas for benthic invertebrate samples see Supplementary Table 2.1.

Collection and identification of Chironomidae samples was commissioned by the EA and performed by APEM (Les Ruse). Here samples were collected from 13 lakes in England and Wales, during October 2012 (Supplementary Table 2.2). Chironomids were collected based on the field protocol of the Chironomid Pupal Exuviae Technique (CPET) using a 250µm mesh collection net (Ruse 2010). To enhance our collection of chironomid sequences a number of unidentified chironomid exuviae collected from Llyn Padarn (N. Wales), during the period 2013-14, were also sequenced (referred to as PA specimens). For the Chironomid Pupal Exuviae Technique (CPET), the floating pupal skins (pupal exuviae) are collected from the leeward side of water bodies, such as lakes, as a safe and easy way of obtaining abundance data that are representative of at least a large part of the lake (Wilson & Ruse 2005; Ruse 2010). Identification of the pupae instead of the larvae is preferred as identification of the larvae is very challenging, with many of the species being superficially very similar (Raunio *et al.* 2011). Pupal exuviae on the other hand, exhibit characteristic forms allowing experienced taxonomists to identify them more easily (Wilson & McGill 1979) and providing more accurate species level identifications. After collection, the chironomid exuviae were preserved in absolute ethanol and identification was performed within one week of collection.

All specimens used for DNA Barcoding were photographed prior to DNA extraction. Photographs were taken using an SLR camera mounted on a standard base for larger specimens or using a dissecting microscope for smaller specimens. For documentation, each specimen was assigned a unique code (location - species code - number, e.g. ANG5-T27-1, sample collected in East Anglia area, site 5, species T27 Trichoptera *Halesus radiatus*, specimen 1). For the photography step, the specimens were positioned according to Barcode of Life Database (BOLD) requirements and instructions.

2.3.2 DNA extraction

Extraction of DNA was performed from ethanol-preserved tissue using different protocols depending on the specimen tissue type. Trichoptera specimens were extracted with a

modified salting out protocol, adapted from Sunnucks & Hales (1996). Generally, 1-3 legs of each specimen were used, depending on size, while trying to avoid abdominal tissue to minimise *Wolbachia* contamination (Smith *et al.* 2012). For Molluscan specimens a CTAB – chloroform based extraction protocol was used, utilising part of the foot muscle of the animal. Testing of various protocols proved this option most effective for molluscs, due to the presence of mucus in mollusc tissues causing inhibition of downstream PCR amplification. Finally, extraction of chironomid exuviae samples was performed using a Qiagen DNEasy Blood and Tissue extraction kit. Fine chopping of pupal exuviae, overnight incubation of specimens with 20µl Proteinase K (20mg/µl) (Sigma – Aldrich) and multiple final elution steps were used to maximise DNA yield of chironomid samples.

2.3.3 PCR amplification

Extracted DNA was amplified with Polymerase Chain Reaction (PCR) of the Cytochrome Oxidase Subunit I gene (COI). Universal primers were used for amplification (Folmer *et al.* 1994) as described previously for sequencing a 658bp fragment of the COI (Barcoding region). PCRs were performed in 25 µl reactions, each containing: 5µl GoTaq Reaction Buffer, 0.5 µl forward primer (10mM), 0.5 µl reverse primer (10mM), 0.25 µl Promega Go Taq DNA Polymerase (5U/µl), 0.5µl dNTPs mix (10mM), 1 µl Bovine Serum Albumin (BSA), 1 µl DNA template (diluted at 10ng/µl) and 16.25 µl PCR grade water. The following thermocycling conditions were used: denaturation at 94 °C for 2 min, followed by 35 cycles of: denaturation at 94 °C for 30 sec, annealing at 52 °C for 30 sec, extension at 72 °C for 1 min, followed by a final extension step 72°C for 10 min.

PCR products were visualised on a 2% agarose gel. Successfully amplified samples underwent a purification step to remove residual primers using an Exo-TSAP (Exonuclease – Thermosensitive Alkaline Phosphatase) protocol. For the Exo-TSAP protocol: 1µl of Exo-TSAP mix (0.1 µl Exonuclease, 0.1µl TSAP, 0.8 µl PCR water) was added to obtained PCR product from each sample and incubated for 15min at 37°C, 15min at 74°C and 15min at 4°C. Purified products (> 35ng/µl concentration) were sent to Macrogen, Holland for Sanger sequencing. Unidirectional sequencing was performed using the forward universal COI primer (LCO1490).

2.3.4 Data analysis

Sanger generated sequences were edited using CodonCode Aligner v.3.7.1 (CodonCode Corporation, Massachusetts). The sequences were sorted according to quality score and grouped based on taxonomically identified species to allow direct comparison of same taxa. Sequences were aligned using the software MEGA 4.0 (Tamura *et al.* 2007), using the ClustalW method (Thompson *et al.* 1994). All sequences were translated and checked for the presence of stop codons and insertions - deletions in order to detect and remove possible nuclear mitochondrial pseudogenes (NUMTs) (Bensasson *et al.* 2001). Construction of phylogenetic trees was performed with the Neighbor-Joining (NJ) (Saitou & Nei 1987) and the Maximum Likelihood (ML) (Nei & Kumar 2000) methods, with pairwise deletion and Kimura-2-Parameter (K2P) distance calculation (Kimura 1980), with 1000 bootstrap replicates. Using the K2P model allowed direct comparison of our results with similar studies.

To assign taxonomy to the non-identified specimens collected from Padarn Lake (PA), we used either the BOLD online identification system, or identification through the NJ and ML phylogenetic resemblance with other identified specimens. Taxon names in parentheses were assigned through the BOLD online identification tool (e.g. Figure 2.6). Only for hits >99% was species level identification assigned to the sequence (e.g. PA3 *Microtendipes chloris*). For lower match hits, the sequence was identified only to the genus level (Figure 6, e.g. PA 17 *Virgatanytarsus sp.*).

We tested identified species delineation based on the use of set thresholds as has been suggested by Meyer & Paulay (2005), by investigating the presence of false positive and false negative species annotations. False positives were defined as conspecifics with higher diversity than the threshold, which would be annotated as new species. False negatives were defined as heterospecific sequences with less diversity than the threshold from the nearest species, which would be attributed to the same species (Hubert & Hanner 2015). Distance calculation and testing for the existence of the barcoding gap, were conducted in package SPIDER in R (v 3.1.3). Function [dist.dna] was used to calculate a distance matrix using K2P distances with pairwise deletion, and [threshopt] was used to perform threshold optimisation analysis. Subsequently the data were tested for instances where the barcoding

gap was absent and results were plotted to present cumulative error according to set threshold and K2P distances within each group.

2. 4 Results

2.4.1 Sequencing results

Overall, DNA Barcoding resulted in successfully obtaining 217 sequences from 94 species across target groups. These include, 55 Trichoptera species with 111 barcodes (16 families and 36 genera), 17 Gastropoda species with 55 barcodes (16 families and 36 genera), and finally 22 species of Chironomidae with 35 barcodes (19 genera). Additionally, one Bivalvia (*S. corneum*), two Amphipoda (*C. pseudogracilis* and *G. pulex*), one Hemiptera (*N. glauca*), one Coleoptera (*G. marinus*) and one Isopoda (*A. aquaticus*) species were barcoded. Barcoding of these individual species was undertaken for the needs of another experiment (Chapter 4). Furthermore, invertebrate sampling efforts resulted in the collection of numerous other specimens of Trichoptera, Coleoptera and representatives of other groups (Ephemeroptera, Plecoptera, Gastropoda, and Isopoda). Sequencing of these additional specimens was not undertaken here due to time and budgetary constraints, but they will be incorporated into future projects.

Table 2.1: Summary table of calculated K2P distances.
Within species, genus and family level (where applicable) divergences are shown for Trichoptera, Gastropoda and Chironomidae. See also variation in barcode sequence length and total number of sequences per group.

Taxon	Category	No. of Groups	K2P (%)			Sequence length (bp)			No. of sequences
			Min	Mean	Max	min	mean	max	
Trichoptera	**Within species**	55	0	0.86	4.24	366	608	622	111
	Within genus	36	0	8.14	25.4				
	Within family	16	0	18.48	31.4				
Gastropoda	**Within species**	17	0	0.4	1.6	587	622	631	55
	Within genus	14	0	2.13	10.5				
	Within family	6	0	15.27	22.6				
Chironomidae	**Within species**	22	0	0.39	1.99	309	524	606	35
	Within genus	19	0	4.48	12.87				

2.4.2 Phylogenetic analysis results

2.4.2.a **Trichoptera**.

For the Trichoptera species, congeneric and con-familial species always clustered together on the NJ and ML phylogenetic tree with 100% bootstrap support (Figure 2.1a-b, Supplementary Figure 2.2a-b), suggesting that COI barcodes for Trichoptera are highly conserved at the genus and family level. The complete NJ tree for Trichoptera with collapsed information at the family level can be seen in Supplementary Figure 2.1. To provide better resolution at the specimen level the tree is split in two sub-trees (NJ: Figures 2.1a-b, ML: Supplementary Figures 2.2a-b) based on the two main subgroups found (split position is indicated with an arrow in Supplementary Figure 2.1).

At the sub-order level, groupings also follow the known phylogeny of Trichoptera as per Kjer *et al.* (2001). For the 31 morphologically identified Trichoptera species with multiple representatives, intraspecific diversity measured with the K2P model ranged between 0-4.24% (0.86% average) (Table 2.1), while zero intraspecific diversity was observed for 10 species, and 24 species were represented by a single sequence (singleton species). The highest intraspecific diversity was observed within the species *S. personatum* (4.24%) and *H. radiatus* (4.04%) (Figure 2.1a).

At the family level, the most well represented in our data was the Limnephilidae family (Figure 2.1a), with 10 genera. Within family distances range between 0- 31.4% with a mean divergence of 18.48% (Table 2.1). The highest within family diversity was found in Hydroptilidae (29.06%, four genera) and lowest in Brachycentridae, Lepidostomatidae and Odontoceridae, each represented by a single genus (Figure 2.2). At the genus level, K2P distances ranged between 0-25.4% with an average of 8.14% (Table2. 2). The highest diversity was found within the genus Hydroptila (average 24.6%), followed by Oxyethira (average 22.5%) and Aglaylea (average 20%).

Possible geographic variation was detected at the species level, with geographic structure being mainly evident for species comprising specimens from distant sampling locations. Species *D. annulatus* collected in Scotland (SCO) clustered separately from those collected in E. Anglia (ANG) (with 99% NJ and 89% ML bootstrap support, 1.2% K2P distance between

the two sub-groups) (Supplementary Figure 2.3a). Similarly for species *L. marmoratus* (with 100% NJ and 92% ML bootstrap support, 2.4% K2P between subgroups) (Supplementary Figure 2.3b). The same pattern is also found for species *S. pallipes* (100% NJ and ML support) (Goeridae) (Supplementary Figure 2.3c) with all southern collected specimens clustering separately from Scottish sample SCO9-T79 (lower between subgroups K2P distance at 0.4%). Within species variation for species *A. fuscipes* (Glossosomatidae) did not exhibit a clear geographically related pattern, as more geographically distant specimens clustered better with each other than with samples derived from more proximate localities, for both analysis (Supplementary Figure 2.3c). For species *O. albicorne* (Odontoceridae) conflicting results were obtained between the two phylogenetic approaches, with ML tree suggesting total absence of variation within the group in contrast to NJ analysis (similarly for *A. nervosa* (Limnephilidae) (Figure 2.1a, Supplementary Figure 2.1).

The highest intraspecific diversity detected for the species *H. radiatus* (4.04%) could be related to the presence of the SCO2_T27 sequence, which forms a separate cluster [bootstrap 83% NJ (Figure 2.3), 64% ML (Supplementary Figure 2.1)]. Blast search (BOLD online search engine) did not provide a definite identification, as it returned close matches with both neighbouring species (99.8 -100% *H. radiatus*, 99.7% *H. digitatus*). Calculated divergence for the outlying sequence was 4% from *H. radiatus*, and 5% from *H. digitatus*. Divergence between the two later species was 5.7%. Excluding this sequence reduced the intraspecific divergence of *H. radiatus* to 0.1%.

For the genus Hydropsyche (Hydropsychidae), misidentification was the most likely explanation for inconsistencies between nomenclature and phylogeny for species *H. pellucidula* and *H. instabilis*. Three specimens that were originally identified as *H. pellucidula* (NWC8-T32, NWC6-T32 and WAL13-T32), clustered with the *H. instabilis* species, and not with sequences COR4-T33 and NWC5-T33 of *H. pellucidula* (Figure 2.1b). Moreover, NWC8-T32, NWC6-T32 and WAL13-T32 were all identified via BOLD as *H. instabilis* (98.35 – 99.45% hits). Furthermore, for *H. instabilis*, distance calculations between the two subgroups showed a 2.2% divergence, which might be the result of geographic isolation, as specimens of the subgroups were collected in Cheshire and Wales respectively. Overall, six cases of misidentification (5.4%) were found in the Trichoptera dataset and a possible misidentification which cannot be verified by our data (specimen COR6-T77 could belong to

species *Sericostoma baeticum* based on BOLD identification). Low-level presence of the bacterial endosymbionts of the genus *Wolbachia* was also detected. In total, six specimens from species *S. personatum* returned verified *Wolbachia* sequences after sequencing with universal COI primers (specimens COR4_T77, DEV3_T77, NWC6_T77, COR1_T77, WAL13_T77, COR3_T77, 67% of analysed samples for this species). One more *S. personatum* specimen (OXF2_T77), was annotated via blast as a rotifer parasite, which was also the case for two specimens of species *Rhyacophila dorsalis* and *Hydropsyche siltalai* (specimens SCO1_T76 and SCO9_T35) (NCBI: GI:157365474).

Figure 2.1a: Neighbor - Joining phylogenetic tree of Trichoptera species.
Values on branches represent bootstrap support. Coloured boxes show Trichoptera family groupings (part1) (1000 bootstrap replications).

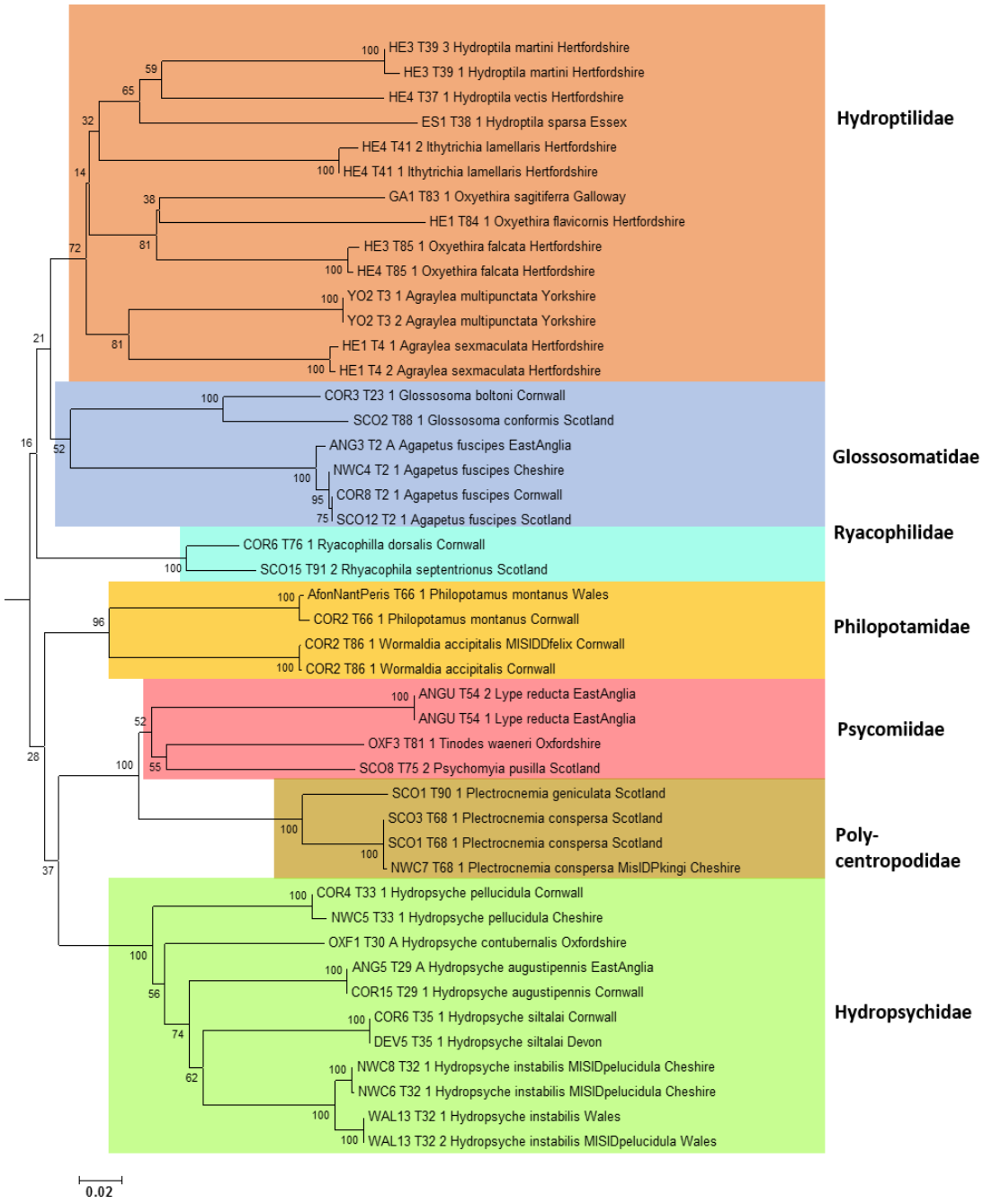


Figure 2.1b: **Neighbor - Joining phylogenetic tree of Trichoptera species.**
Values on branches represent bootstrap support. Coloured boxes show Trichoptera family groupings (part1) (1000 bootstrap replications).

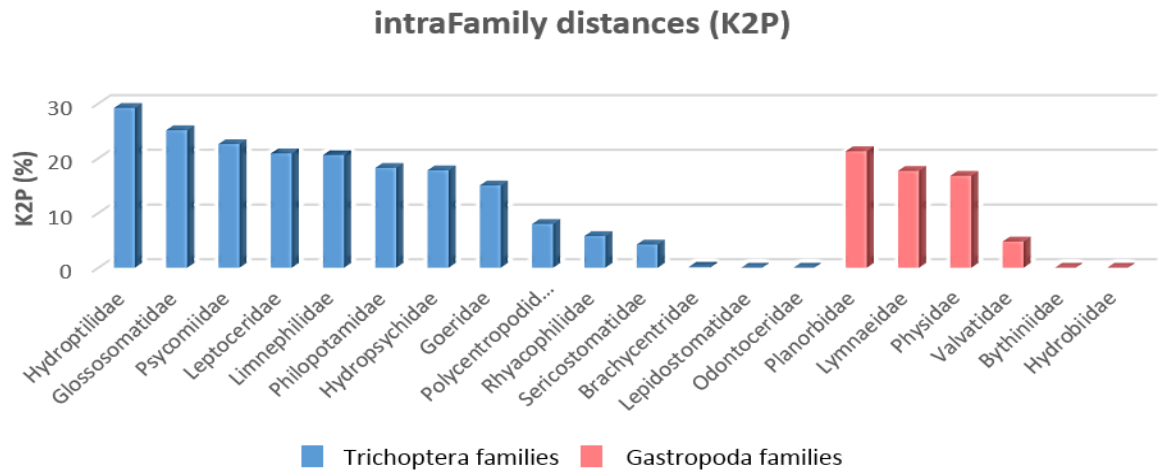


Figure 2.2: Mean within family distances calculated with the K2P parameter for Trichoptera (blue) and Gastropoda (red) families (except from Molanidae and Phryganeidae).

Table 2.2: Mean within genus K2P (%) distances for Trichoptera and Gastropoda genera. Only genera, which were represented by more than one species, are shown.

Order	Family	Genus	Mean intra-Genus K2P (%)
Trichoptera	Glossosomatidae	Glossosoma	10.5
	Hydropsychidae	Hydropsyche	17.9
	Hydroptilidae	Hydroptila	24.6
		Oxyethira	22.5
		Agraylea	20
		Ithytrichia	0.64
	Leptoceridae	Athripsodes	18.4
	Limnephilidae	Limnephilus	14.6
		Halesus	5.9
	Philopotamidae	Philopotamus	0.64
	Polycentropodidae	Plectrocnemia	7.9
Gastropoda	Planorbidae	Planorbis	9.7
	Physidae	Physella	10

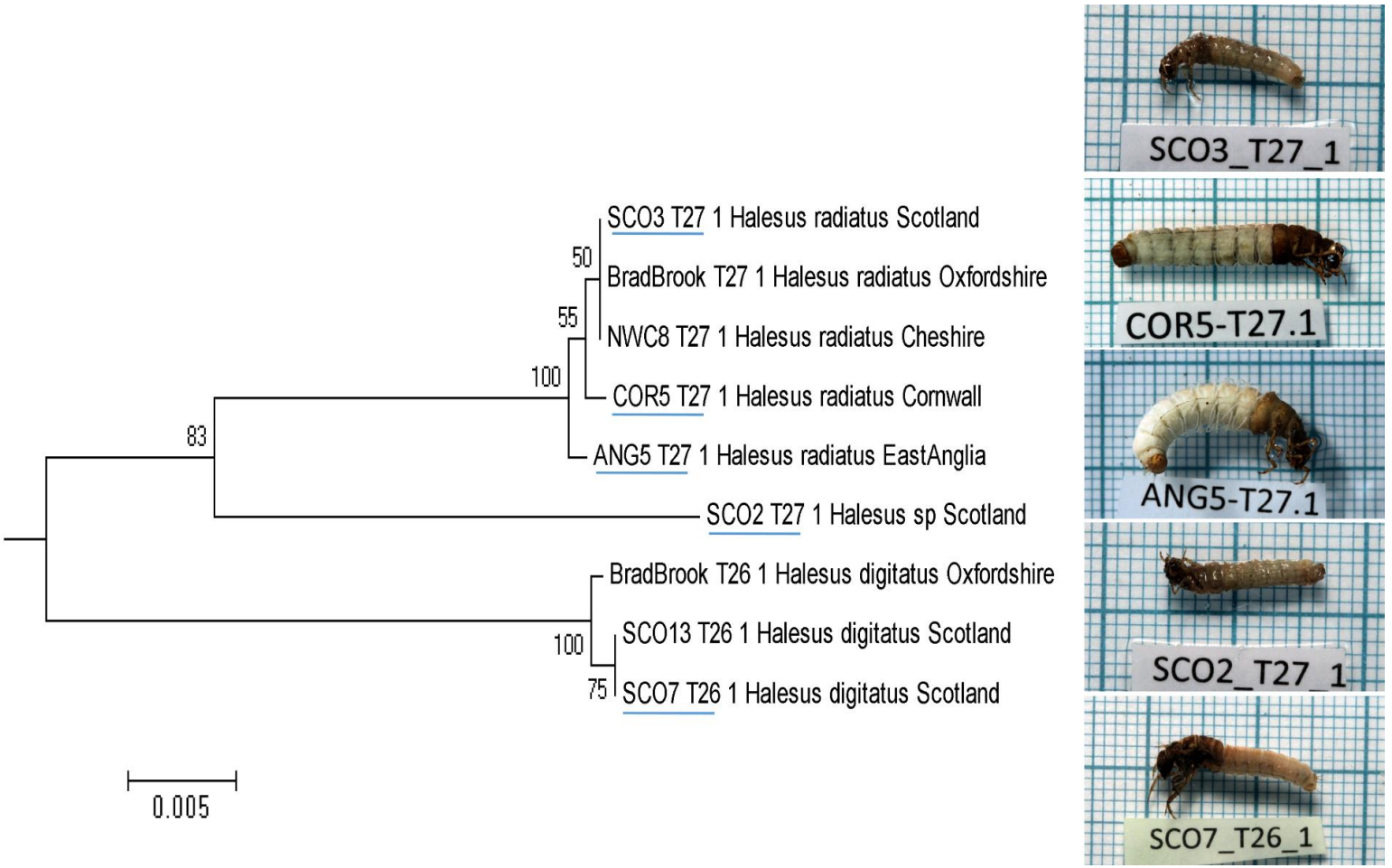


Figure 2.3: Neighbor – Joining sub-tree for the species *H. radiatus* and *H. digitatus*.
Marked *Halesus sp*. possible misidentified or cryptic species sequence SCO2-T27. On the right, images from five of the represented specimens (underlined).

2.4.2.b Gastropoda.

For the Gastropoda group, sequences were also found to form monophyletic genera (14) and family (6) groups, and all species groups with multiple representatives were supported with 100 bootstrap of the NJ and ML phylogenetic trees (Figure 2.4, Supplementary Figure 2.4). Divergence (K2P) at the family level, ranged between 0-22.6% (average 15.27%) (Figure 2.2). The Planorbidae family was the best represented in our data (Figure 2.4), with six genera, and presented the highest mean intra-family variation, in contrast to the lowest by Hydrobiidae. At the genus level, two genera show high levels of diversity: genus Physella (Physidae) and genus Planorbis (Planorbidae) with 10% and 9.7% respectively (Table 2.2). Evidence of geographic variation was found for species *A. fluviatilis* (Planorbidae) and *V. piscinallis* (Valvatidae) (Supplementary Figure 2.5a-b). For *A. fluviatilis*, three geographic subgroups (1. Cornwall/Devon (87% NJ), 2. East Anglia (97% NJ), and 3. Cheshire (100% NJ) (Supplementary Figure 2.5a) (K2P distance between subgroups: 1-3: 0.5%, 1-2:0.5% and 2-3:0.3%). For *V. piscinallis*, two deep subgroups are identified between Somerset and E. Anglia collected samples (100% bootstrap on NJ and ML trees) (Supplementary Figure 2.5b), while the intraspecific diversity for this species was high (4.8%). Intraspecific diversity for species *R. balthica* (Lymnaeidae) was 1.5%, and probable geographic variation may be present, without a clear pattern (Supplementary Figure 2.5c). A subgroup collected from Devon and Cornwall was 1.4% divergent from the other sequences. Specimens COR1-G18 and ANG2-G18 were initially taxonomically identified to genus and were then assigned species level identity through BLAST and NJ tree.

The sequences obtained from two *B. leachii* and one *B. tentaculata* specimens presented almost zero interspecific diversity, which suggests possible misidentification for at least one of these specimens (Figure 2.4). Blasting did not provide any information since there are currently no sequences or good hits available for these species. To decipher sequence identity, the DNA barcodes from these species were aligned against mitochondrial genome scaffolds obtained from another experiment (see chapter 4). Phylogenetic analysis suggests that specimen Q12_1 belongs to species *B. leachii*, even though it was originally identified as B*. tentaculata* (Figure 2.5) (Divergence: within *B. leachii*: 0.08%, between species: 1.57%). Finally, low amounts of intraspecific variation were evident for other species, probably due

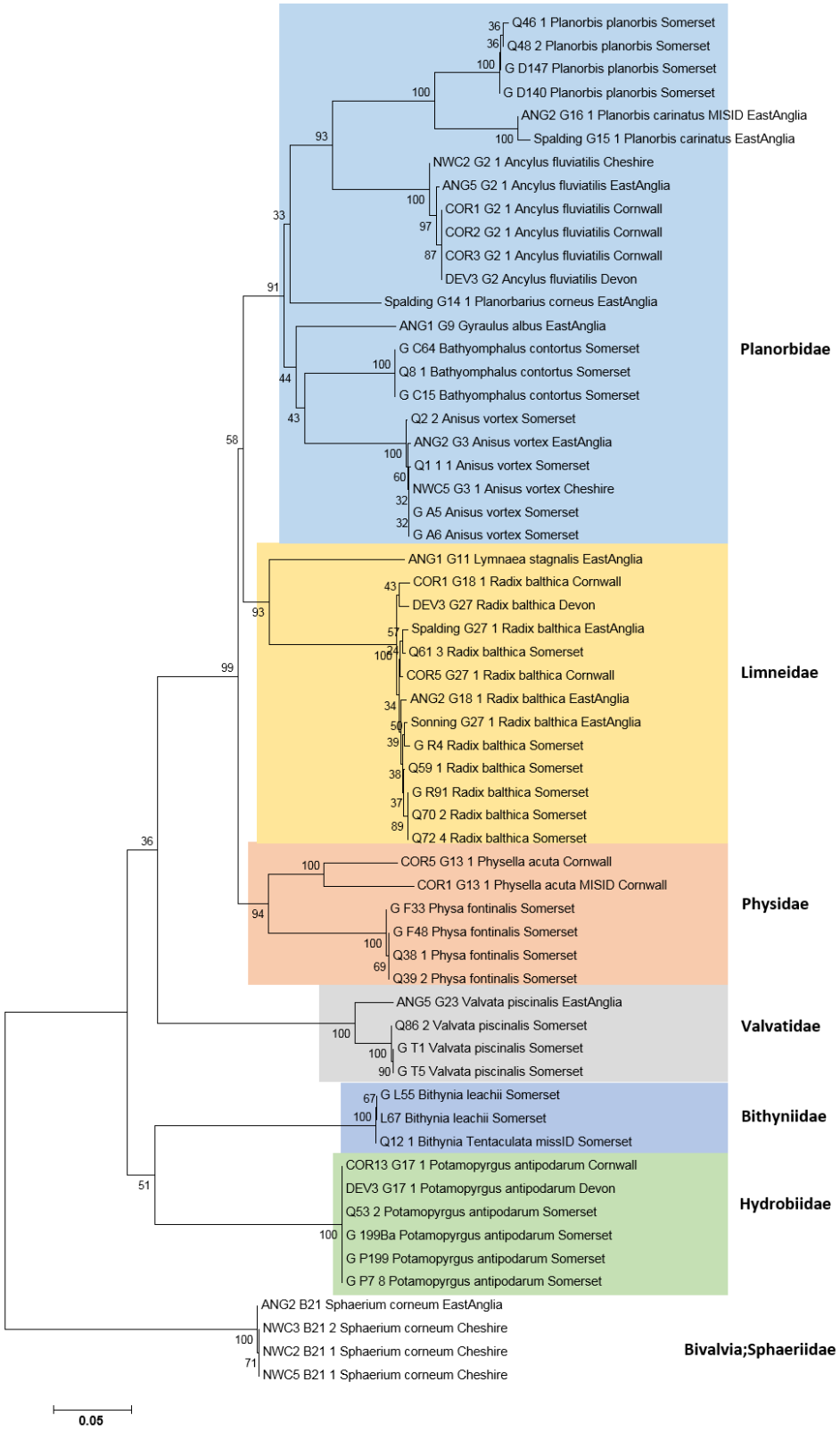


Figure 2.4: **Neighbor-Joining phylogenetic tree of Gastropoda sequences.**
Branches are numbered by bootstrap support (1000 bootstrap replications). The six Gastropoda families are highlighted using different colours. The single Bivalvia species (*S. corneum*) can be seen as outgroup.

to lack of geographic variation present amongst the specimens used (e.g. *P. antipodarum* and *P. planorbis*).

Overall, three cases of misidentification were detected in the Gastropoda data (5.5%). Sequencing of the species *B. tentaculata* also presented difficulties due to co-amplification of parasitic oligochaete species (possibly *Chaetogaster limanei*), as two other specimens originally identified as *B. tentaculata* (ANG2-G5, Q20_4), were infected with the oligochaete parasite. Furthermore, a 3-bp insertion was found to occur in all species members of the families Planorbidae, Limeidae and Physidae (Supplementary Figure 2.6). The same insertion was not found for members of the other gastropod families Valvatidae, Bithiiidae and Hydrobiidae.

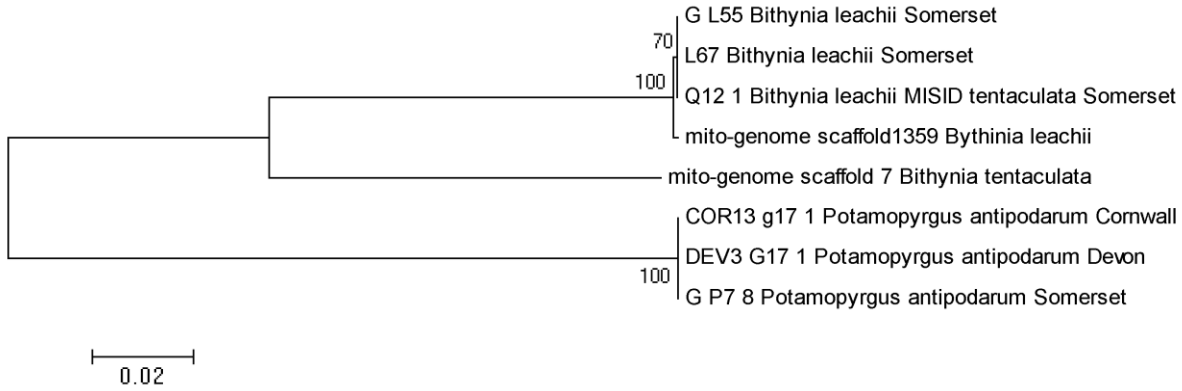


Figure 2.5: **Neighbor-Joining phylogenetic tree for Gastropoda species *B. leachii* and *B. tentaculata*.**
The tree is showing a possible misidentification case for specimen Q12_1. The barcode sequences were aligned against mitochondrial genome scaffolds obtained through shotgun sequencing (see chapter 4). Species *P.antipodarum* was used as outgroup. Values show bootstrap support.

2.4.2.c Chironomidae.

In total ~139 chironomid exuviae specimens were extracted, including 120 specimens which had previously been identified to species level, and 19 unidentified specimens collected from Padarn Lake during the period September 2013 – 2014 (PA specimens). Out of these 35 (25%) Chironomidae DNA Barcodes were finally obtained. Success rate for the identified specimens was 22%, while for the unidentified was 47%. Generally, DNA extraction resulted in relatively low quantity of DNA with concentrations <10 ng/µl and amplification using

universal COI primers was achieved in approximately 90 cases. After Sanger sequencing, the majority of the successful PCR products returned either a mixed sequence signal suggesting co-amplification of different templates, or a sequence of good quality, which in many cases did not match the chironomid target taxon. BLAST identification revealed that 35 non-chironomid sequences had been obtained, from a number of contaminant taxa, which were being preferentially amplified and sequenced over chironomid trace DNA. These taxa were identified as water mould (Saprolegniaceae; *Achlya* or Saprolegniaceae; *Saprolegnia*), bacteria (Nitrosomonas), gastropods (*Gyraulus sp.*) and annelids (*Chaetogaster*). These sequences were removed from further analysis.

Intraspecific variation calculated from species with multiple representatives ranged between 0-2% with an average of 0.4%. Eighteen cases of singleton species were excluded from divergence calculations. The highest intraspecific diversity found within species *O. consobrinus* (ORTHCON) (1.9%), collected from Windermere and Derwent Reservoir (Figure 2.6). For *O. consobrinus*, the NJ tree shows that specimen ORTHCON7 clustered closer to a sequence from a different species (MACRNEB7), than to its conspecific ORTHCON3. Nevertheless, ML analysis did not support the same clustering of these two sequences (Supplementary Figure 2.7). In addition, the genetic distance between ORTHCON7 and MACRNEB7 was high (4.6%). Using >99% BLAST and BOLD hits, 5 out of 9 unidentified PA specimens, were assigned species level identification, while genus level identification was assigned for 4 sequences (Figure 2.6, PA specimens, taxonomic names in parentheses).

Specimens CHIRP2-4-2 and CLATATR10-1 (Gr1), and CHIRTEN13-1 and CHIRTEN13-1b (Gr2) were taxonomically identified as three distinct species (*Chironomus plumosus*, *Cladotanytarsus anthracinus* and *Chironomus tentants*). However, divergence calculations showed 0% distance between CHIRP2-4-2 and CLATATR10-1 (Gr1), and 0.83% between the 2 subgroups of the branch (Gr1 and Gr2, Figure 2.6). Complementary BLAST analysis of these sequences (BOLD) showed close matches with species *Chironomus plumosus* (99.32 - 100%) and *Chironomus usenicus* (99.28%). These hits suggest possible misidentification and we would suggest that the taxon was most likely *C. plumosus* according to molecular identification, but further work/sampling would be needed to fully corroborate this assertion.

Within-genus variation (taxonomic and BOLD identified species and genera) ranged between 0-12.9%, with average intrageneric variation of 4.48% (excluding 8 cases of genera represented by single specimens). Significantly higher intra-genera variation was found for genus *Tanytarsus* (12.8%) and *Orthocladius* (11.9%) compared to other taxa. Representatives of these genera were not monophyletic on the NJ and ML trees (Figure 2.6, Supplementary Figure 2.7, symbol marked: *Tanytarsus* - tringles, *Orthocladius* - circles).

Other taxa: Additionally for the six "other" taxa sequenced (Bivalvia: *S. corneum*, Amphipoda: *C. pseudogracilis* and *G. pulex*, Hemiptera: *N. glauca*, Coleoptera: *G. marinus*, Isopoda: *A. aquaticus*), within species divergences were generally low (<0.7%) and one case of misidentification was detected for a specimen of the species *Gammarus pulex,* which was originally identified as *Crangonyx pseudogracilis*.

2.4.3 Threshold analysis and Barcoding gap calculation.

Using the package SPIDER, we estimated the optimum divergence threshold level for species discrimination per taxonomic group, based on our sequencing data. The optimum threshold level was calculated based on a combination of minimum cumulative error and minimum number of false negatives. Results suggest that the optimum threshold for Trichoptera was >1.3%, for Chironomidae 0.2-0.7% and for Mollusca 0.5% (see calculated cumulative error pre threshold, Figure 2.7). This suggests that if the generally applied by the BOLD BIN system 1% threshold would underestimate the number of Trichoptera species in our data, while conversely it would overestimate the number of species in the Chironomidae and Mollusca data. Investigation for the detection of the Barcoding gap in our data (Figure 2.8) showed that a Barcoding gap was found in all cases, with interspecific divergence exceeding intraspecific variation.

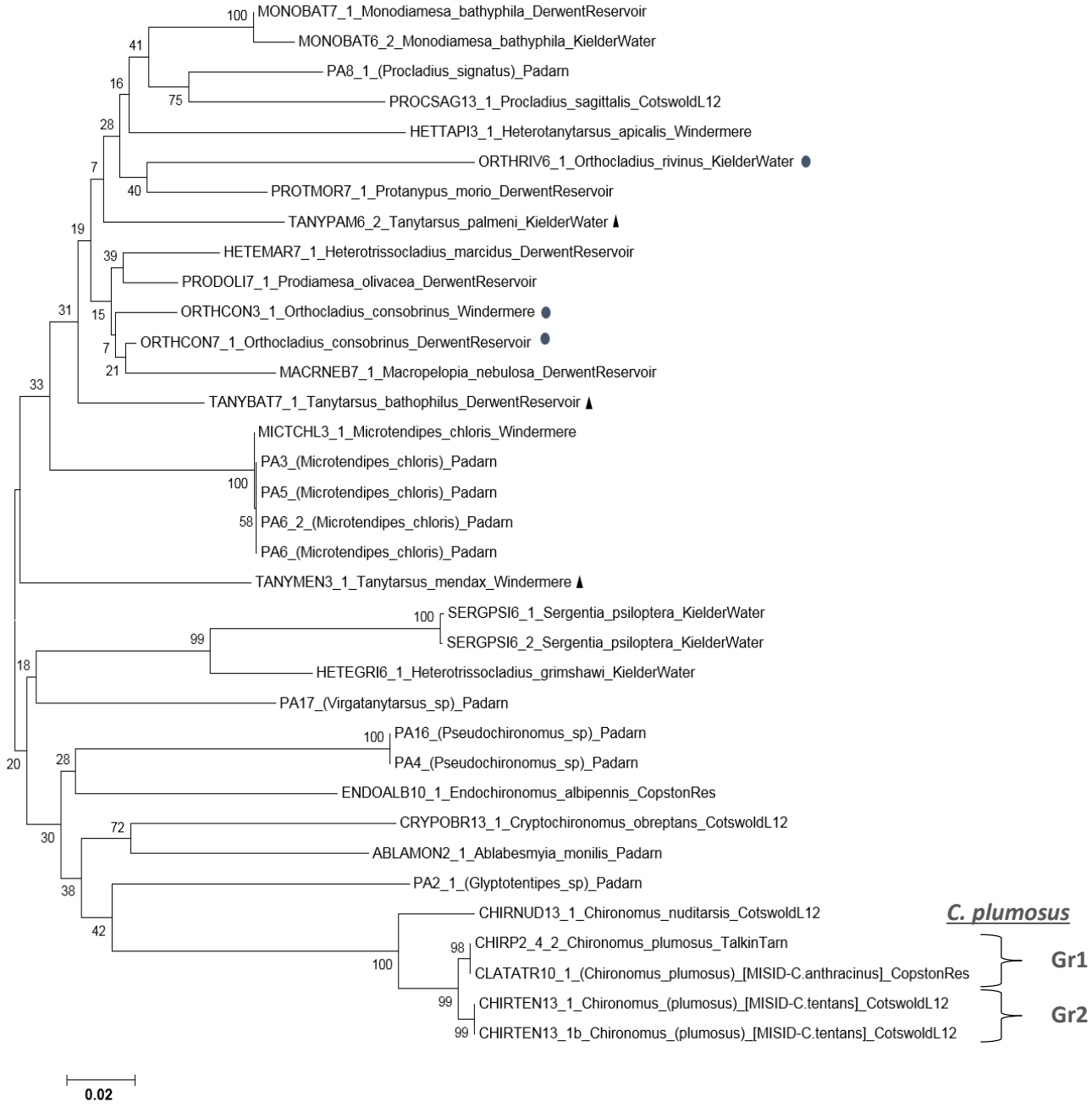


Figure 2.6: **Neighbor-Joining phylogenetic tree of 35 Chironomidae sequences.**
Twenty-two species were identified through taxonomy or BOLD identification (taxonomic name in parenthesis). Specimens used were collected from seven lakes in England and Wales Possible species complex (*C. plumosus*) is shown, with two defined subgroups: Gr1 and Gr2. (Non-monophyletic genera are marked; *Tanytarsus*: tringles, *Orthocladius*: circles) (1000 bootstrap replications).

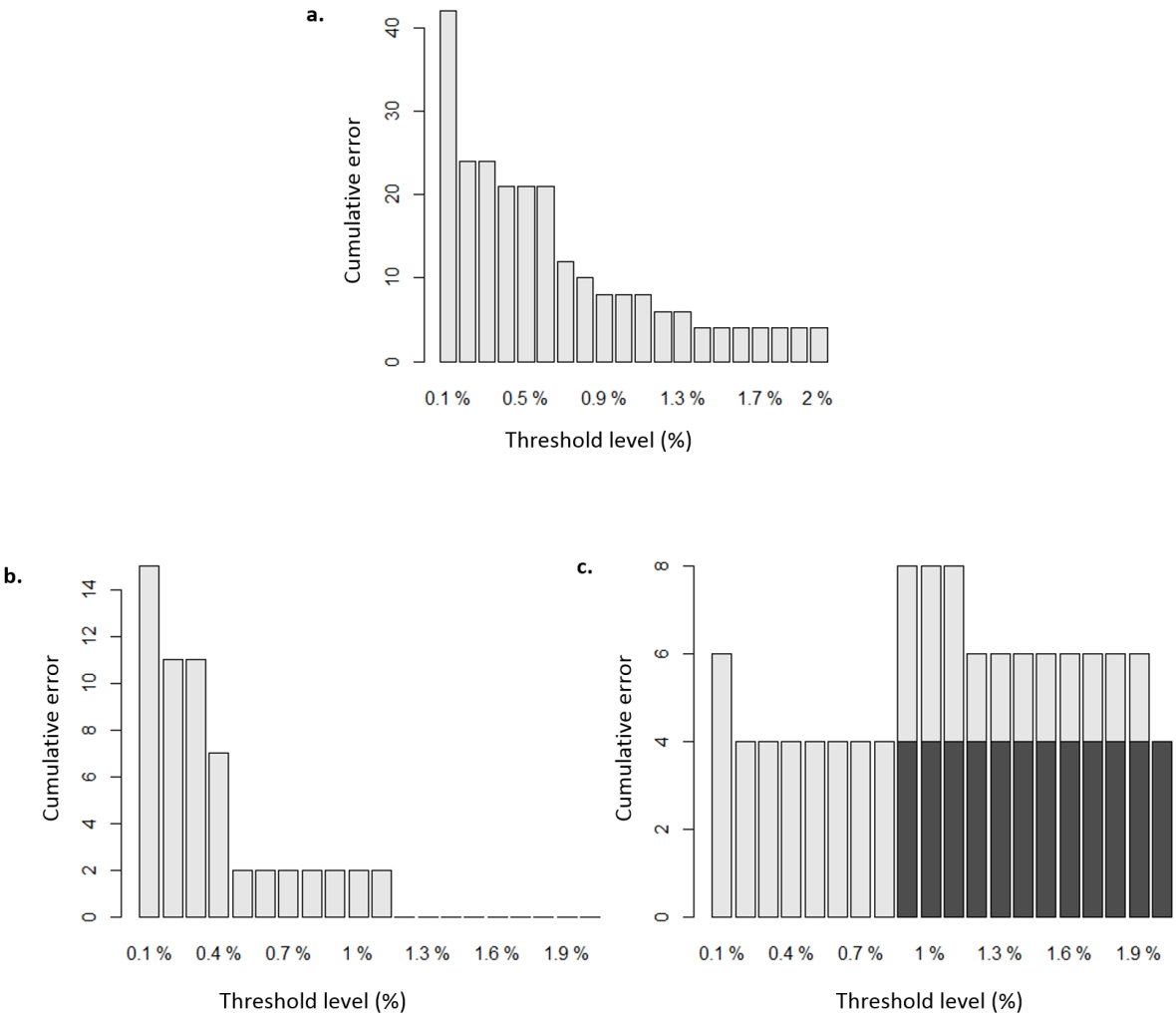


Figure 2.7: Barplot showing false positive and false negative identification of species. False positive (f.p) (grey) and false negative (f.n.) (black) identification of species for each of the main groups used for analysis depending on the set threshold used, based on cummulative error calculation. Optimum threshold for species identification per group, a) Trichoptera: >1.3% (min error = 6%), b), Chironomidae: >0.5% (min error = 2%), c) Mollusca: 0.2-0.7% (min error = 4%) (x-axis: threshold level, on the y-axis: cumulative error).

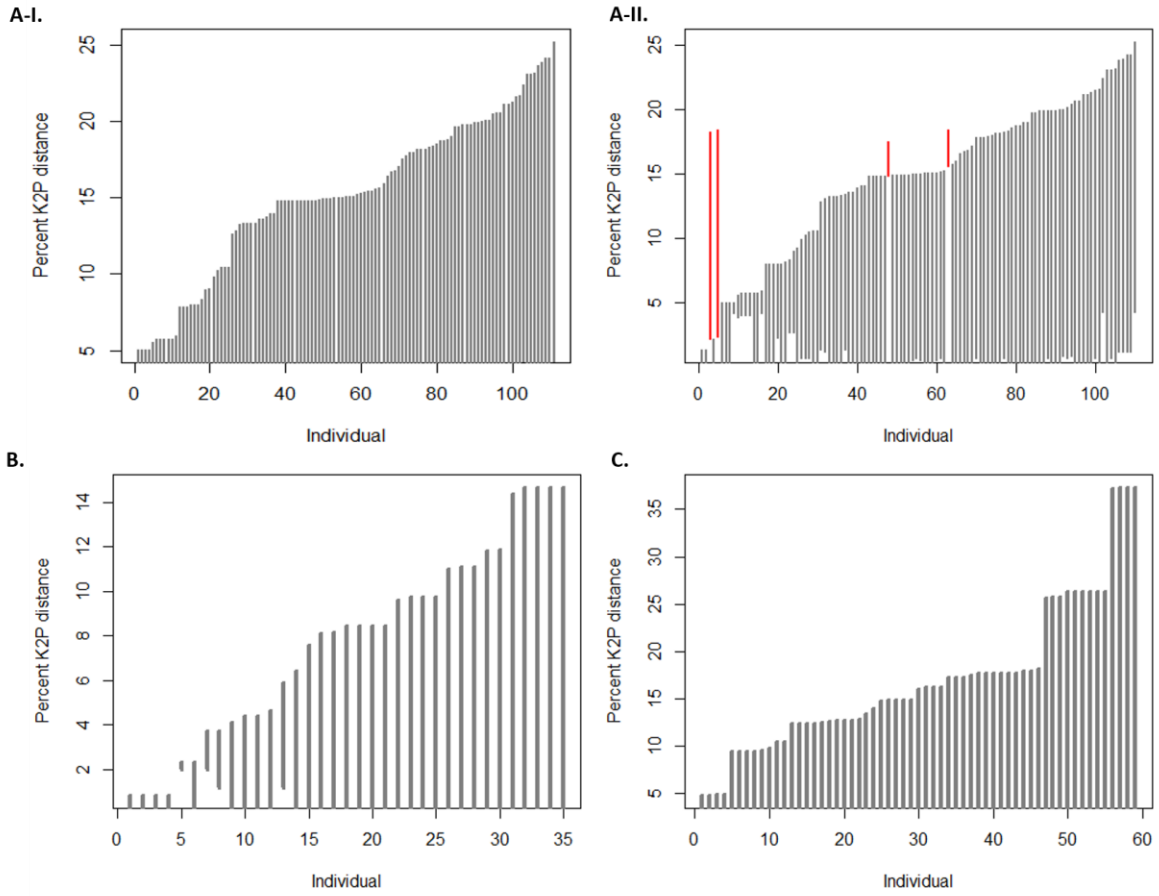


Figure 2.8: Line-plot of the calculated barcoding gap.
A-I) Trichoptera (111 sequences), A-II.) Example of the presence of reverse relationships, using hypothetical barcoding data, B.) Chironomidae (35 sequences) and C.) Mollusca (59 sequences). Grey lines represent the furthest intraspecific distance (bottom) and the closest interspecific distance (top). Red lines indicate the absence of barcoding gap. Graph A-II is used here as an example to highlight how the data can be screened for the presence of individuals for which the principle of barcoding gap does not apply.

2.5 Discussion

DNA Barcoding was efficient in providing species level identification for sequenced specimens of Trichoptera, Gastropoda and Chironomidae. In the former two, COI barcodes were also conserved at the genus and family level, with all congener and con-familial sequences forming distinct groups, while for chironomids congeneric species did not always form monophyletic groups. Amplification and sequencing of chironomid exuviae was challenging, especially for specimens that had been through the process of identification. Low levels of Wolbachia endosymbionts and rotifer parasites were detected in some Trichoptera specimens. Furthermore, DNA barcoding was able to detect low numbers of taxonomically misidentified specimens and two possible cryptic species, across all three groups, which were supported by phylogenetic analysis.

2.5.1 Investigating divergence levels within Trichoptera, Gastropoda and Chironomidae

We estimated genetic distances among mtDNA sequences, to investigate the ability of the COI barcoding region to delimit macroinvertebrate species with accuracy. Cases of high intraspecific divergence were further explored to determine whether they were related to geographic variation, cases of misidentification or possible cryptic species (Costa & Carvalho 2010). Misidentification for species with high intraspecific divergences was subsequently confirmed through NJ and ML phylogenetic analysis and blasting of sequences against the NCBI and BOLD databases.

2.5.1.a Trichoptera

Previous work using DNA barcoding for identification of Trichoptera species has verified the presence of a barcoding gap for this group, with low intraspecific and high interspecific divergence (Zhou *et al.* 2011; Ruiter *et al.* 2013). Even though the maximum intraspecific diversity found for Trichoptera in our dataset was 4.24%, the mean divergence was 0.86% (Table 2.1), which is below the 1% threshold used by BOLD or 2% used by other studies (Zhou *et al.* 2009). Additionally, interspecific diversity ranged between 5-25.4% (Figure 2.8). From threshold analysis of our data, the optimum threshold found was 1.3% (Figure

2.7), which is within the expected range for Trichoptera (2%) according to Zhou *et al.* (2009).

Elevated divergence values could theoretically be attributed to one of several causes, such as misidentification, unrecognised cryptic species or hybridization events (Wiemers & Fiedler 2007). More specifically, for the seven Trichoptera species with intraspecific divergence >1% in our data: five could be attributed to geographic variation (species *D. annulatus*, *L. marmoratus*, *M. sequax*, *A. fuscipes* and *H. instabilis*) (Supplementary Figure 2.3), one could be the result of misidentification (*S. personatum*), and one could be a possible cryptic species (*Halesus sp.* specimen SCO2-T27) (Figure 2.3).

This high level of intraspecific divergence found within the species *H. radiatus* could be related to the presence of an ambiguous specimen (SCO2_T27) (Figure 2.3). This specimen was originally identified as *H. radiatus,* but clustered separately from the other conspecific sequences (83% bootstrap NJ), between *H. radiatus* and sister species *H. digitatus*. The possibility of misidentification could not be verified through BLAST searches (NCBI and BOLD), which returned close matches of this sequence with both species (99.8 -100% *H. radiatus*, 99.7% *H. digitatus*). These findings suggest that this specimen could be an undetected cryptic species. Further studies should be required to determine the existence of a new species, which are beyond the scope of this work. Nevertheless, the usefulness of DNA barcoding for uncovering cryptic diversity has been discussed, as the COI gene could help to clarify species boundaries and serve as a starting point for discovery of new taxa (Kress *et al.* 2015).

Furthermore, our Trichoptera data presented relationships following the major phylogenetic structure at the suborder level, as described by Kjer *et al.* (2001). All the families formed clearly defined monophyletic groups on the NJ and ML trees (Figures 2.1a-b). Moreover, representative families from three major sub-order groups of Trichoptera also formed distinct groups, including Annulipalpia or retreat-maker caddisflies (4 families), Spicipalpia or cocoon-maker caddisflies (3 families) and Integripalpia or tube case-maker caddisflies (9 families) (Kjer *et al.* 2001).

Even though specimens from the same species, congeneric and con-familial species formed monophyletic groups, we cannot expect to fully resolve the depths of the Trichoptera phylogeny from only this depth of sampling and sequencing. Nevertheless, our findings support the accuracy of the barcoding method, since failure to achieve monophyly at the species level would also compromise the method itself (Meyer & Paulay 2005). More extensive studies have shown the utility of COI for Trichoptera species delimitation (e.g. Geraci *et al.* 2011; Ruiter *et al.* 2013), but we anticipate that this work will add to the ongoing efforts for collecting barcoding information for Trichoptera.

2.5.1.b Gastropoda

The COI gene has been shown to successfully identify phylogenetic relationships across a broad range of gastropod groups (Remigio & Hebert 2003). For the 17 species of Gastropoda analysed, levels of intraspecific variation were generally low, ranging between 0-1.6% (average 0.4%), and all the species with multiple representatives formed monophyletic groups (100 bootstrap support, Figure 2.5, Supplementary Figure 2.4). The high diversity initially observed in the *Valvata piscinallis* group of sequences (4.8%) (Supplementary Figure 2.5b), might suggest cryptic diversity or misidentification. Blasting of these sequences against the databases did not assist with deciphering the taxonomy for this species. Levels of variation in species *A. fluviatilis* (Supplementary Figure 2.5a), might suggest possible geographic variation between samples collected over a North – South gradient.

Patterns of geographic variation for species *R. balthica* were more ambiguous because even though a subgroup of Cornwall/Devon samples was 1.5% divergent from the rest, a Cornish sample was also present in the other subgroup (Supplementary Figure 2.5c). The species *R. balthica* is part of a morphologically cryptic species complex inhabiting mainly lentic water bodies but also slow flowing rivers and streams (Pfenninger *et al.* 2011). Mechanisms of dispersal, which could affect the levels of genetic variation of this species, mainly depend on connectivity of habitats and passive dispersal (e.g. waterfowl) (Pfenninger *et al.* 2011). The same dispersal mechanisms are also typical for many other freshwater molluscs and non-flying freshwater invertebrates (Pfenninger *et al.* 2011).

Low to zero levels of divergence were found for some other gastropod species probably due to absence of specimens from remote locations. Species *P. antipodarum*, which also showed low diversity, is a small invasive snail species, known as the New Zealand mud snail, which can reach very high population densities. An assay for early detection of this species through water extracted environmental DNA (eDNA) has been tested (Goldberg *et al.* 2013), and collection of DNA barcoding data for this and other invasive species could prove useful for accurate application of new detection methods (see also chapter 3).

Furthermore, the presence of a 3bp insertion (single codon, 5' side) was detected, for all species of the gastropod families Lymnaeidae, Physidae and Planorbidae (Supplementary Figure 2.6). These insertions could be related to the presence of pseudogenes (Bensasson *et al.* 2001), nevertheless, stop codons were not found. Similar findings have been reported by other studies, (e.g. Remigio & Hebert 2003; Layton *et al.* 2014; Borges *et al.* 2016), which also recorded several cases of 3bp deletions and insertions for various species of marine and freshwater species of molluscs. Furthermore, Remigio & Hebert (2003) also report a 12bp insertion present in species of the *Planorbis* genus, but no similar insertions were observed in the two *Planorbis* species used in this study. The presence of length variants in gastropods appears to be a common phenomenon (Hebert *et al.* 2003; Remigio & Hebert 2003). Nevertheless, their functional significance is not yet clear and more in depth analysis would be required to resolve the mechanisms behind their occurrence and associated impacts (Remigio & Hebert 2003).

On the family level, our phylogenetic analysis suggests the monophyly of families Lymnaeidae, Physidae and Planorbidae (Figure 2.4). The monophyletic origin of the same Pulmonate freshwater families has also been proposed by Remigio & Hebert (2003). More recent work by Smith et al. (2011), based on transcriptome data, confirms the monophyly of the whole Gastropoda clade; nevertheless finer level phylogenetic relationships still remain to be investigated to a large degree (Borges *et al.* 2016).

2.5.1.c Chironomids

Due to the nature of the samples collected (shed pupal skins), the chironomid DNA used for barcoding was only trace DNA left on the exoskeleton by the adult during the

emergence process (Ferrington JLC, Blackwood MA, Wright CA, Crisp NH, Kavanaugh JL 1991). The exuvial DNA can also be up to two days old, as the floating skins can be still collected from the surface of the water up to 48 hours after emergence (Ferrington JLC, Blackwood MA, Wright CA, Crisp NH, Kavanaugh JL 1991). Additionally to the presence of low amounts of DNA, the exuviae skins could also be carriers of DNA from exogenous sources such as microbial eukaryotes or other organisms co-inhabiting the same ecosystems (Dick 1970). As was expected, DNA extraction of chironomid pupal exuviae proved challenging on many occasions and the success rate for sequencing chironomid exuviae was only 25%, while another 25% of sequenced samples matched non-target taxa.

Previous attempts for DNA extraction of exuviae using salting out protocols (S.A.Miller 1988) have failed, but using a Qiagen DNeasy kit was more successful for extracting various life stages of chironomids (Krosch *et al.* 2011; Kranzfelder *et al.* 2016). In the present study, we also employed Qiagen DNeasy kit for DNA extractions, following overnight incubation of ground up specimens and using multiple elution steps to maximize yield. Nevertheless, DNA yields were generally low, and amplification was achieved in approximately 65% of the samples. Similar difficulties in sequencing chironomid exuviae have also been reported by Kranzfelder *et al.* (2015), while the success rates for attaining chironomid sequences in that case were even lower, at only 13.7%. Similarly, that study also obtained barcodes from exogenous sources such as cladocerans, water moulds, humans etc., while in our case sequences from gastropods, water moulds, annelids and bacteria were found. Handling of the specimens during the identification process (slide mounting) could further contribute to DNA degradation and contamination of exuvial samples. This is indicated by the higher success rates obtained from unidentified/unhandled specimens over identified ones, with 47% over 22% successful sequencing events respectively.

The levels of intraspecific diversity detected in our chironomid sequences were generally low, ranging between 0-2%, (average 0.4%). Levels of intraspecific diversity for chironomids have been reported to be somewhat higher, with an expected range between 0 - 4.9% (Ekrem *et al.* 2007; Carew *et al.* 2013), or 0 - 3.15% (Brodin *et al.* 2013) and an average between 0.82 - 0.9%. Similarly, interspecific diversity was lower in our samples

with a range between 2.3 -14.6%, compared to other studies 5.1- 25.2% (Ekrem *et al.* 2007), 7 - 34.1% (Carew *et al.* 2013). A small "barcoding gap" was found (Figure 2.8b), which could be the result of incomplete sampling of lineages. The detection of a barcoding gap is not generally the case with other work on chironomid COI barcoding (Ekrem *et al.* 2007; Brodin *et al.* 2013). Nevertheless, the higher levels of interspecific divergence reported in other cases (0 - 24.38%) (Brodin *et al.* 2013), could be related to the presence of species complexes or geographic variation as these samples were collected across a wide area along the Baltic coast (Denmark - Sweden). Removal of ambiguous species in that case lowered interspecific divergence to 7-19% (Brodin *et al.* 2013).

Two of the genera sequenced (*Tanytarsus*; Chironominae and *Orthocladius*; Orthocladiinae) were not found to form monophyletic groups on NJ and ML phylogenetic trees (Figure 2.6, Supplementary Figure 2.7). The absence of monophyletic relationships for Chironomid genera has been reported in a wider context and particular cases in the subfamily Chironominae and genus Tanytarsus have been described (Ekrem *et al.* 2007; Demin *et al.* 2011).

Additionally, we encountered difficulties with resolving taxonomic identification for members of the *Chironomus plumosus* group (Figure 2.6). Results from phylogenetic and BLAST analysis indicate possible misidentification of three specimens (CLATATR10-1, CHIRTEN13-1, CHIRTEN13-1b), which could be identified as either *Chironomus plumosus* or *Chironomus usenicus*. Both these species returned very high match hits with our sequences from BOLD database. The species *C. usenicus* has been characterised in Eastern Europe (Poland, Russia) (Polukonova & Beljanina 2002), but we could not verify its presence in the UK. Absence of this species from our collection areas suggests that our specimens indeed belong to *C. plumosus*, which appears to have a much wider European distribution (Pfenninger *et al.* 2007; Gunderina *et al.* 2009; Gunderina 2010). The ambiguous hits in BOLD could also be related to possible hybridization events for this group, as it is believed that species *C. usenicus* is the result of hybridization between species *C. plumosus* and *C. behnigni* (Polukonova & Beljanina 2002).

2.5.2 Possible limitations of DNA Barcoding

2.5.2.b Endosymbionts and other parasitic infections

Bacteria of the genus *Wolbachia* are known endosymbionts, which are common in most arthropod groups (Smith *et al.* 2012). *Wolbachia* are transmitted vertically through maternal lineages, while inducing reproductive alterations such as cytoplasmic incompatibility, feminization and parthenogenesis (Dobson *et al.* 1999). As described in Hilgenboecker *et al.* (2008), up to 66% of insect species are expected to be infected with this endosymbiont. In our data, *Wolbachia* were found in the *Sericostoma personatum* species (Trichoptera), where out of nine specimens sequenced, six (66.7 %) were infected.

Wolbachia bacteria are predominately found at reproductive tissues (ovaries), but their presence has also been documented in somatic tissues such as muscles (Dobson *et al.* 1999) and also legs, which are commonly used for DNA barcoding. Regardless of their documented presence in somatic tissues, they are occurring at a lower rate than in the abdomen; therefore, muscle or clean leg tissue should be preferred for DNA extraction (Smith *et al.* 2012).

Suggestions that the presence of bacterial endosymbionts might compromise barcoding analysis do not stand, as it has been shown that sequencing of the bacteria does not represent a serious risk for barcoding surveys (Smith *et al.* 2012). Nevertheless, checking the data for possible bacterial amplification should always be performed. Differentiation of *Wolbachia* sequences should be easy as there are an average 167bp discrepancies of host to endosymbiont sequence inside the COI barcoding region (for insects) (Smith *et al.* 2012). Testing of the BOLD contents for the presence of undetected *Wolbachia* sequences showed only a 0.01% presence for Trichoptera species at the time and 0.05% for Diptera (not only Chironomidae) (Smith *et al.* 2012). The low presence of Wolbachia documented sequences in our data could be either the result of low presence of contamination or due to the precautions taken during DNA extraction, such as limiting the tissue used to legs of specimens and specifically avoiding contact with gut tissue.

Further to Wolbachia infections, a low number of Trichoptera species were found to be carriers of rotifers, which were preferentially amplified over the target species in 3 cases,

one for each for species *S. personatum, R. dorsalis* and *H. siltalai*. It has been suggested that rotifers can grow on aquatic insect larvae (Örstan, 1999), which may have caused contamination of the DNA extracts. Since rotifers are an important trophic component of freshwater ecosystems (Park & Marshall 2000), co-collection with insect larvae for simultaneous assessment could be a possibility, given that the appropriate controls are put in place to avoid cross-contamination. In addition, two gastropod specimens were carriers of the oligochaete parasite *Chaetogaster limanei*. This parasite is known to infect many types of freshwater snails, by embedding itself in the mucus of the foot or living in the mantle or the pulmonary cavity (for some species) (Hopkins *et al.* 2013). Since DNA was extracted from the foot of the snail specimens, it is possible that the parasites were co-extracted and amplified.

2.5.2.c Nuclear mitochondrial pseudogenes

Nuclear Mitochondrial pseudogenes (NUMTs) are copies of mitochondrial genes, which have been incorporated into the nuclear genome and can exist in multiple copies and varying abundance (Bensasson *et al.* 2001). Importantly for DNA barcoding studies, NUMT sequences can amplify, or co-amplify with the target mtDNA marker when universal primers are used, thereby hindering analysis (Hurst & Jiggins 2005). NUMTs have been detected in many eukaryotic clades, with different abundance, which might differ even between closely related species (Bensasson *et al.* 2001), although their overall effect on the results of DNA barcoding applications has not been extensively studied yet (Song *et al.* 2008). They are more common in arthropods than other groups and not very common in Mollusca species (Bensasson *et al.* 2001). Proposed methods for NUMT identification include use of species specific primers instead of universal, purification of mitochondria prior to DNA extraction, use of tissue with high mitochondrial numbers like muscle, cloning, and long PCR amplification (Bensasson *et al.* 2001; Song *et al.* 2008). To prevent the inclusion of pseudogenes in our final data, leg tissue was used for DNA extraction and all sequences were screened for the presence of stop-codons in MEGA5.

2.5.3 Levels of misidentification

Misidentification levels were similar for Trichoptera and Gastropoda, with 6 (5.4%) and 3 (5.5%) confirmed misidentified specimens respectively. Considering that the three ambiguous sequences found in the chironomid data (see discussion above) were misidentified *C. plumosus*, brings the level of misidentification for this group to 8.6%. Additionally, one more misidentified specimen was found from the miscellaneous taxa, belonging to the species *Gammarus pulex,* that was originally identified as *Crangonyx pseudogracilis,* which is a known freshwater invasive species in the UK (Oreska & Aldridge 2011). Deciphering misidentified specimens was easier when multiple specimens had been sequenced, allowing comparisons; or alternatively, existing records in public databases could assist. Misidentifications re-emphasises the issue of comprehensive sampling effort for the construction of accurate reference databases, since the sampling effort can affect the accuracy of the results (Meyer & Paulay 2005), as does the geographic scale over which specimens were sampled (Bergsten *et al.* 2012).

2.5.4 Benefits of using DNA barcoding in benthology

Using DNA barcoding data occupies a middle ground between molecular phylogenetic and population genetics, with the former dealing with deep relationships of taxa and the latter dealing with intra and inter population diversity. Alternatively, DNA barcoding focuses mainly on delineating species rather than investigating their relationships (Hajibabaei *et al.* 2007).

It has been suggested that incorporation of DNA based methods would decrease the costs associated with bio-assessment. Calculations of the cost per barcode vary depending on the laboratory and pipeline used, but past estimations have placed the cost per individual at 2.5 - 8 $ (Cameron *et al.* 2006; Valentini *et al.* 2009). Comparison between the costs involved in the production of individual barcodes of indicator species for biomonitoring has found that the cost of barcoding exceeds that of traditional (taxonomic) identification by 1.7- 3.4 times (Stein *et al.* 2014). Nevertheless, when taxonomic methods costs were compared against HTS methods the cost was comparable or even lower for the new

sequencing technologies (Stein *et al.* 2014). In that sense, when barcoding is linked to HTS, the collection of barcoding data could provide a valuable base for future HTS applications (Gray *et al.* 2015), since to a large degree, the correct taxonomic assignment of Operational Taxonomic Units (OTUs) largely relies on properly populated reference databases (Deagle *et al.* 2014). Further benefits from the collection of barcoding data include the use of barcode sequences for the association of life stages (e.g. adult and larvae), which could in turn be used to detect or develop diagnostic characters for species identification from difficult to identify life stages (Ruiter *et al.* 2013).

The value of DNA Barcoding for biodiversity assessment of unknown faunas has been demonstrated by the matching of taxonomically identified morphological species with DNA barcode clusters, which had been assigned by using a specified threshold (illustrated by Zhou *et al.* 2009). Overall, incorporation of DNA based identification approaches like DNA Barcoding in biomonitoring could increase its accuracy (Baird & Sweeney 2011), and promote objectivity and comparability of biodiversity assessment and community ecology studies (Pfenninger *et al.* 2007). Therefore, coupling DNA barcoding efforts with HTS for monitoring of benthic samples may provide a more cost efficient way to achieve assessment of ecosystem status and biodiversity in accordance with national and EU level legislation (Brodin *et al.* 2013), and this will be explored further in Chapters 3 and 4.

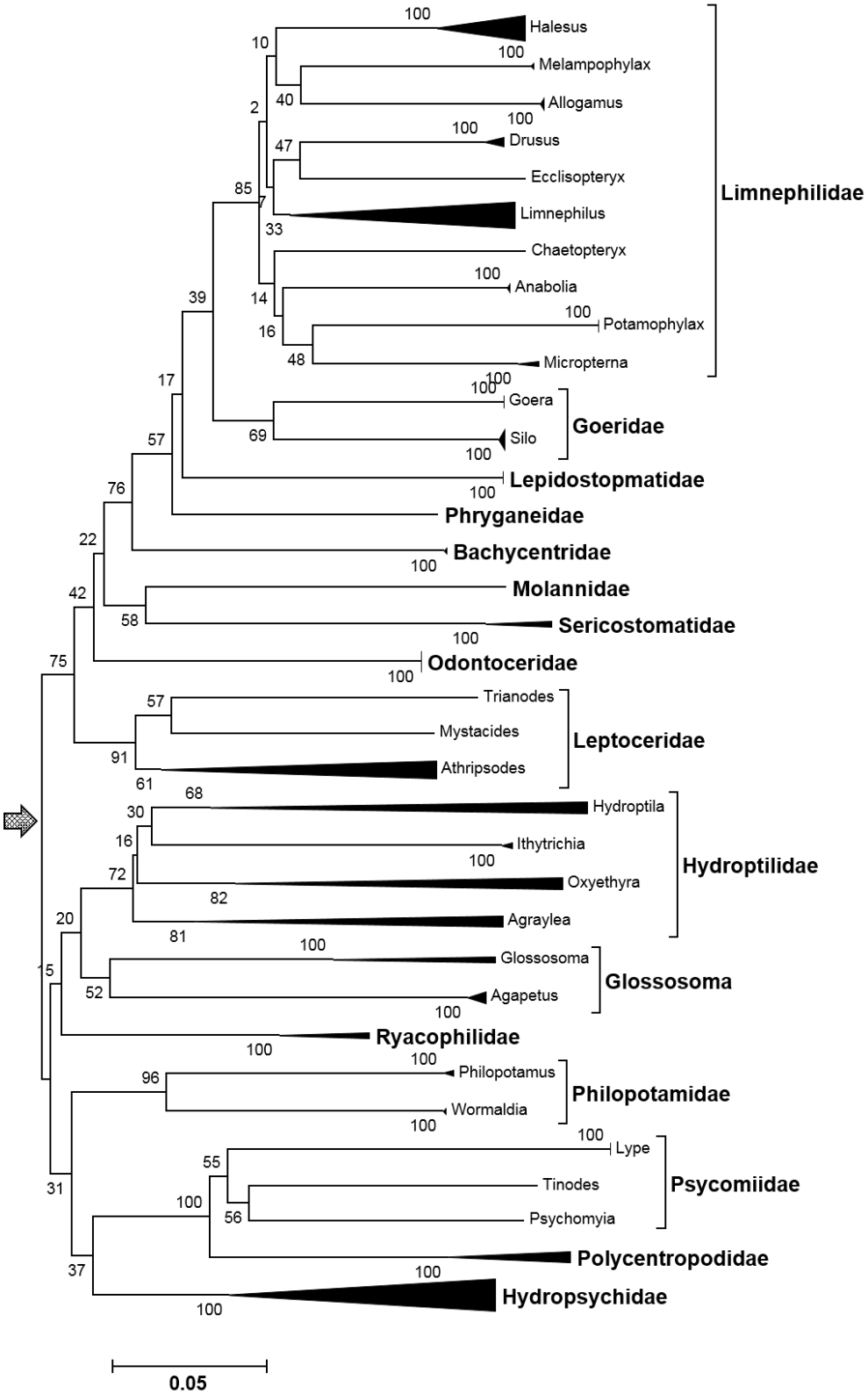
2.6 Supplementary information

Supplementary Table 2.1: List of geographical regions of invertebrate sample collection. See also area code used for cataloguing individual barcodes

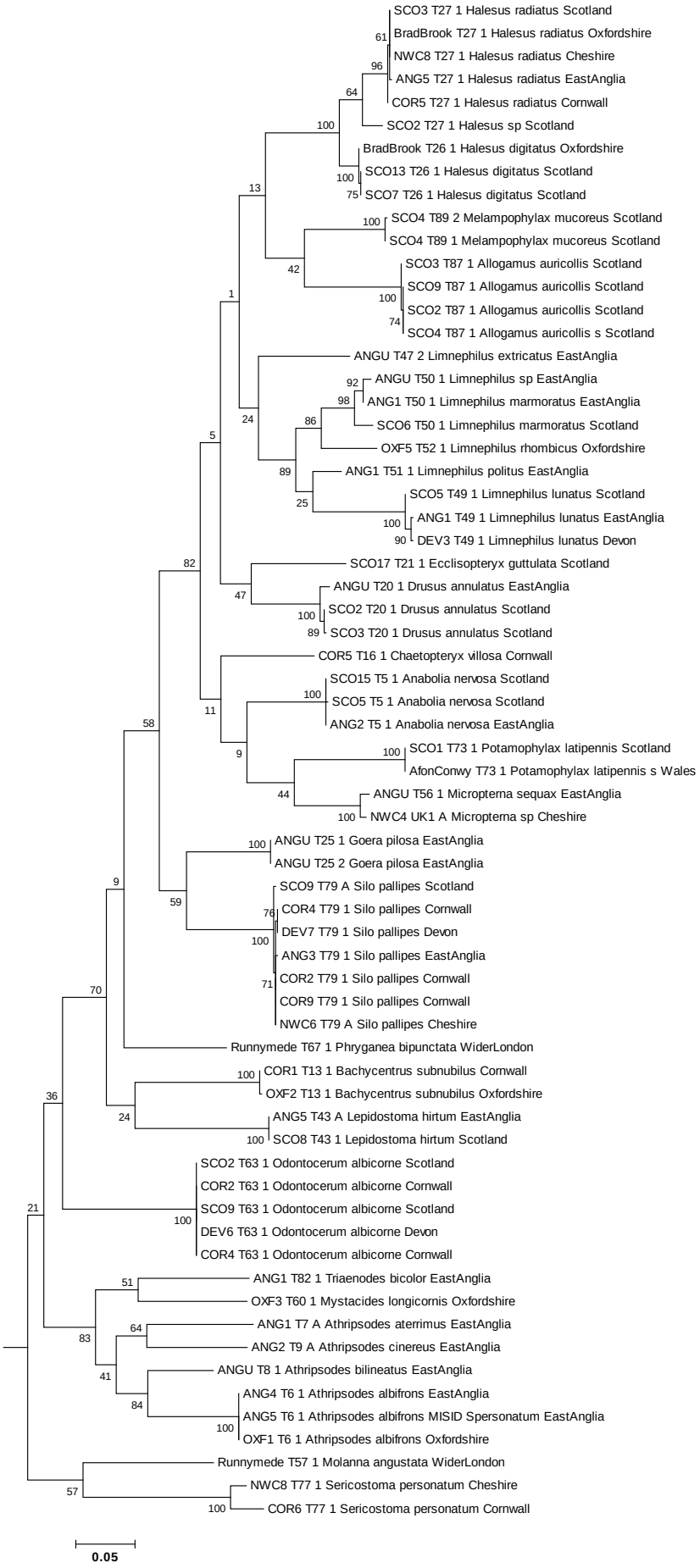
Number	Area code	Geographical Region
1	NWC	Cheshire
2	COR	Cornwall
3	DEV	Devon
4	ANG/ANGU	East Anglia
5	ES	Essex
6	GA	Galloway
7	HE	Hertfordshire
8	OXF	Oxfordshire
9	SCO	Scotland
10	Q	Somerset
11	SUF	Suffolk
12	WAL	Wales
13	Other	Wider London
14	YO	Yorkshire

Supplementary Table 2.2: List of lakes used for collection chironomid exuviae samples. All samples were collected during October 2012. Latitude and longitude information also shown.

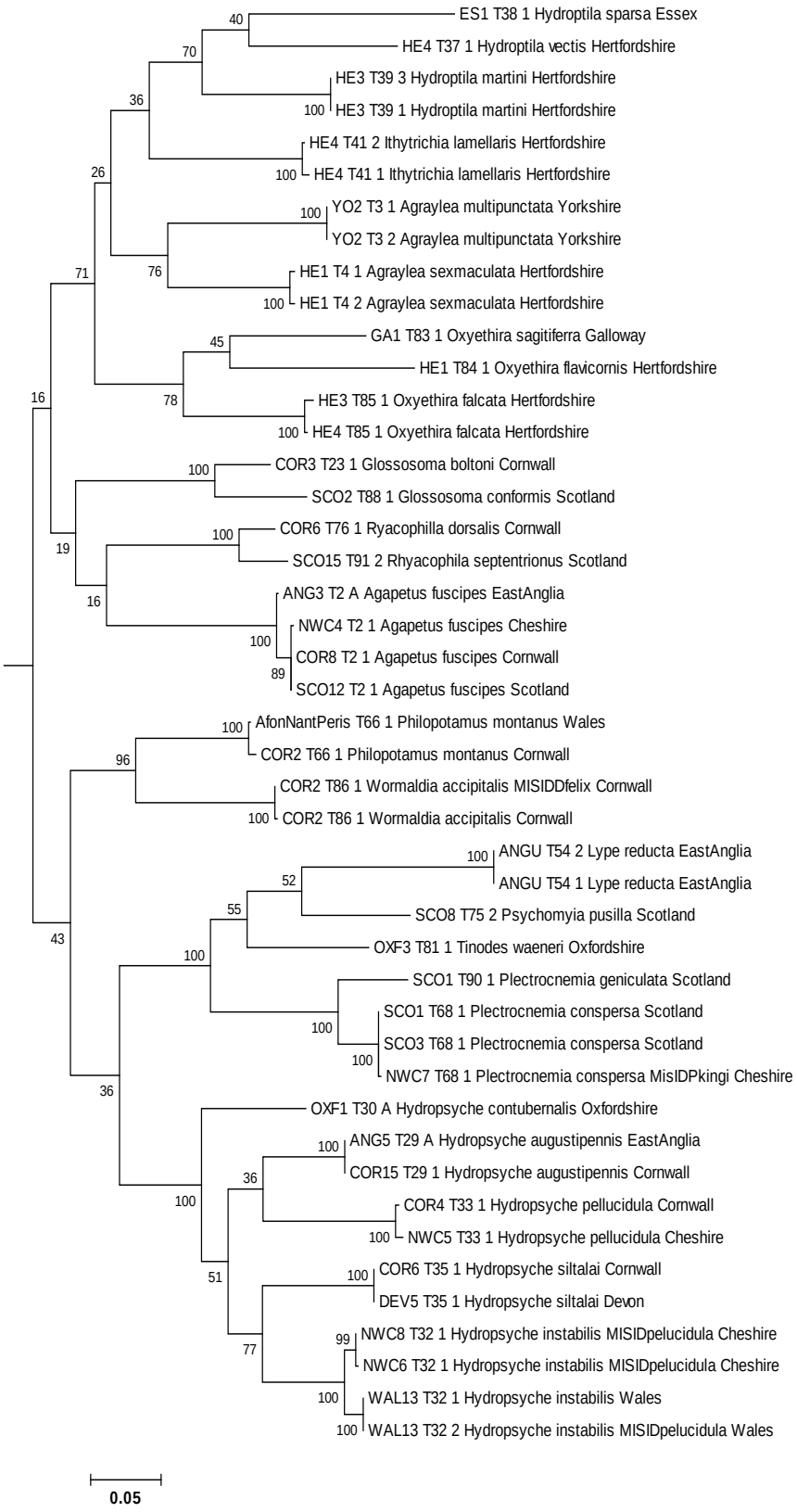
Number	Lake	County	Lat_Lon
1	White Mere	Shropshire	52.84 N 001.47 E
2	Llyn Padarn	Gwynedd	52.87 N 001.35 E
3	Windermere (South)	Cumbria	52.98 N 001.48 E
4	Talkin Tarn	Cumbria	53.04 N 001.51 E
5	Crag Lough	Northumberland	53.05 N 001.54 E
6	Kielder Water	Northumberland	53.07 N 001.53 E
7	Derwent Reservoir	Durham	53.04 N 001.58 E
8	Swinsty Reservoir	North Yorkshire	52.95 N 001.59 E
9	Carsington Water	Derbyshire	52.85 N 001.59 E
10	Cropston Reservoir	Leicestershire	52.82 N 001.64 E
11	Sowley Pond	Hampshire	52.63 N 001.60 E
12	Chew Valley Lake	Avon	52.69 N 001.48 E
13	Cotswold Water Park Lake 12	Wiltshire	52.72 N 001.56 E



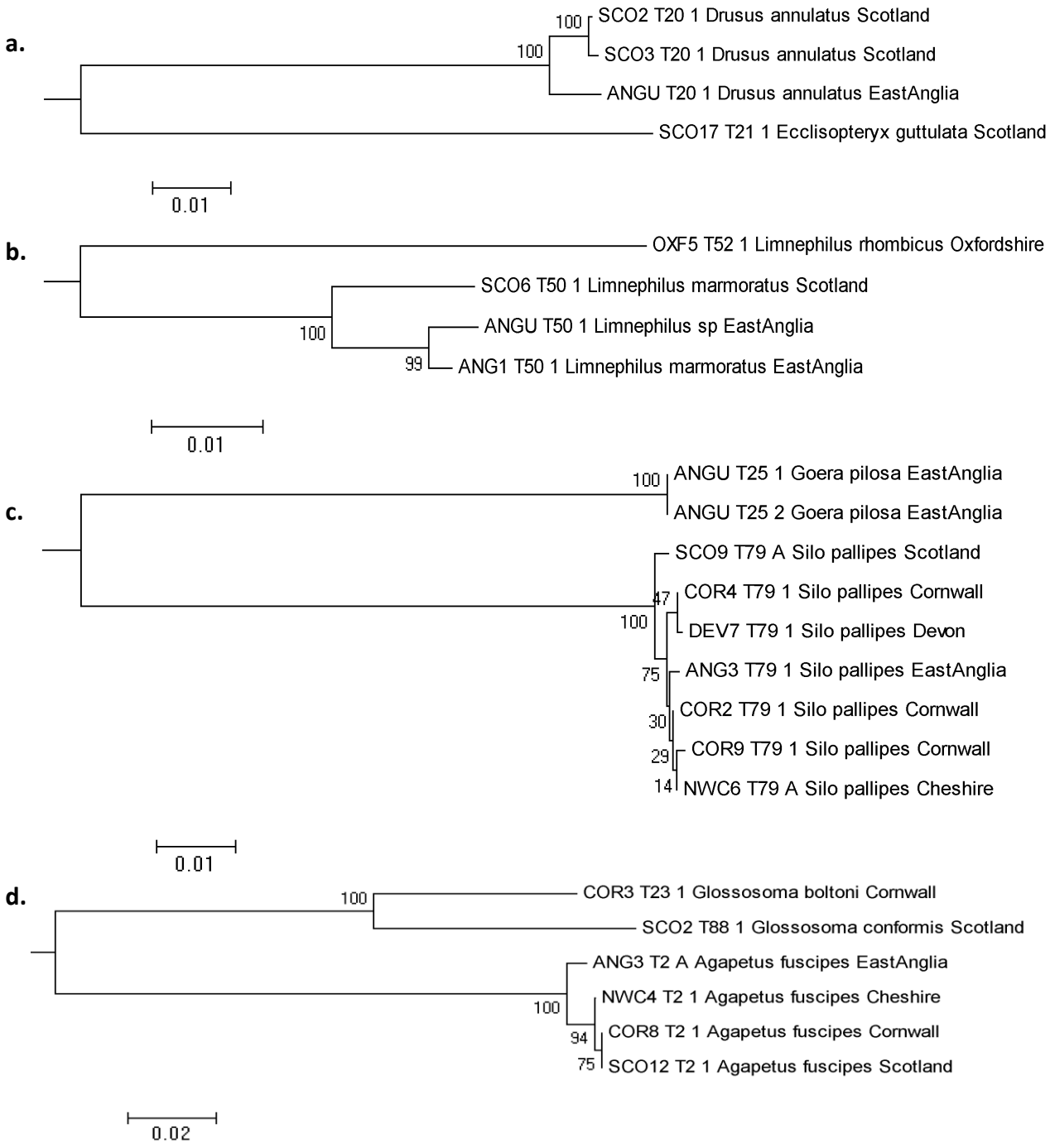
Supplementary Figure 2.1: Synoptic Neighbor-Joining phylogenetic tree for Trichoptera. All families are presented and congeneric species are collapsed in single groups. Values show bootstrap support. The arrow indicates the two main subgroupings presented in separate trees.



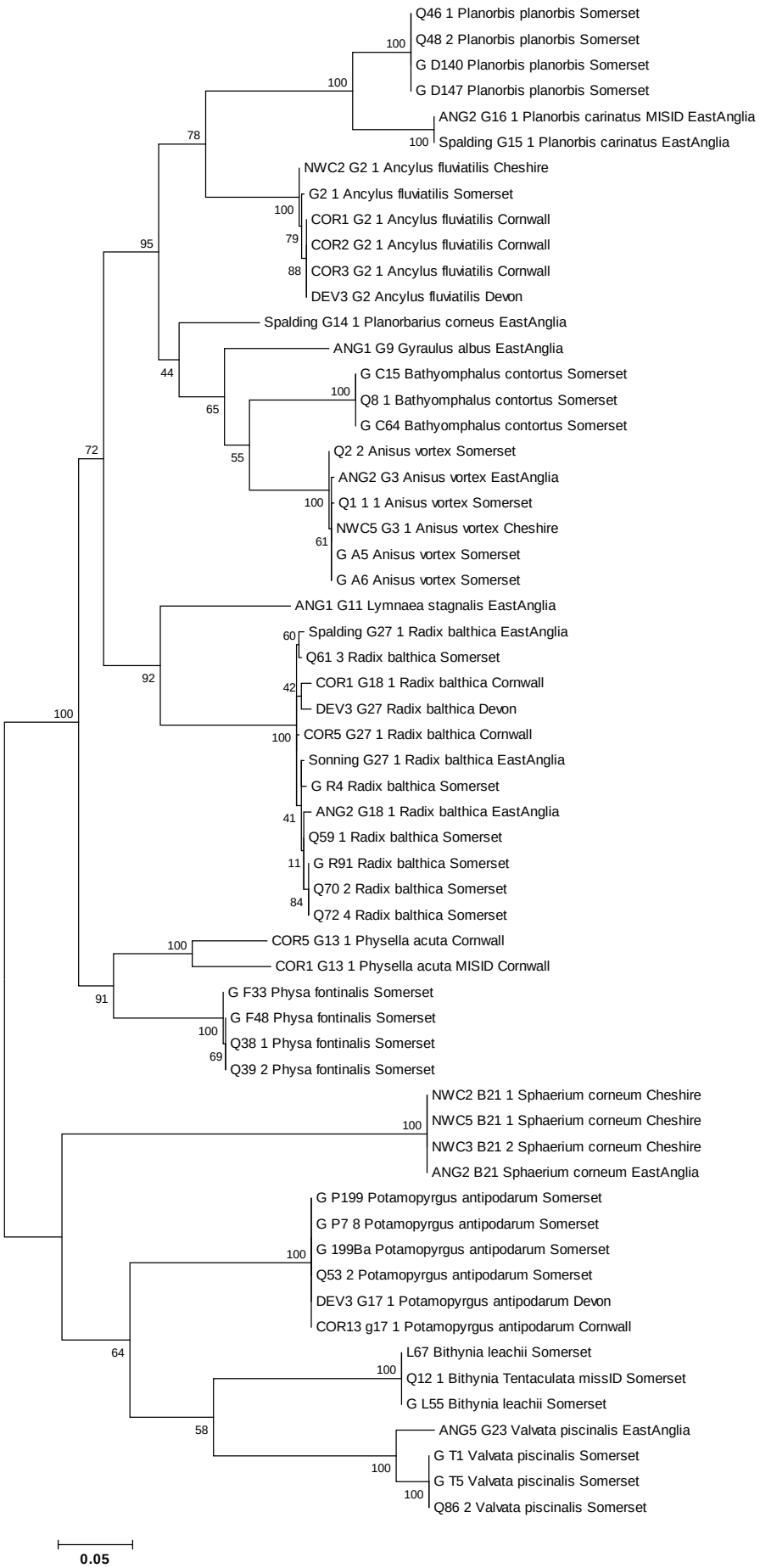
Supplementary Figure 2.2a: Maximum Likelihood phylogenetic tree for Trichoptera sequences constructed with 500 bootstrap replications (K2P distances) (part 1).



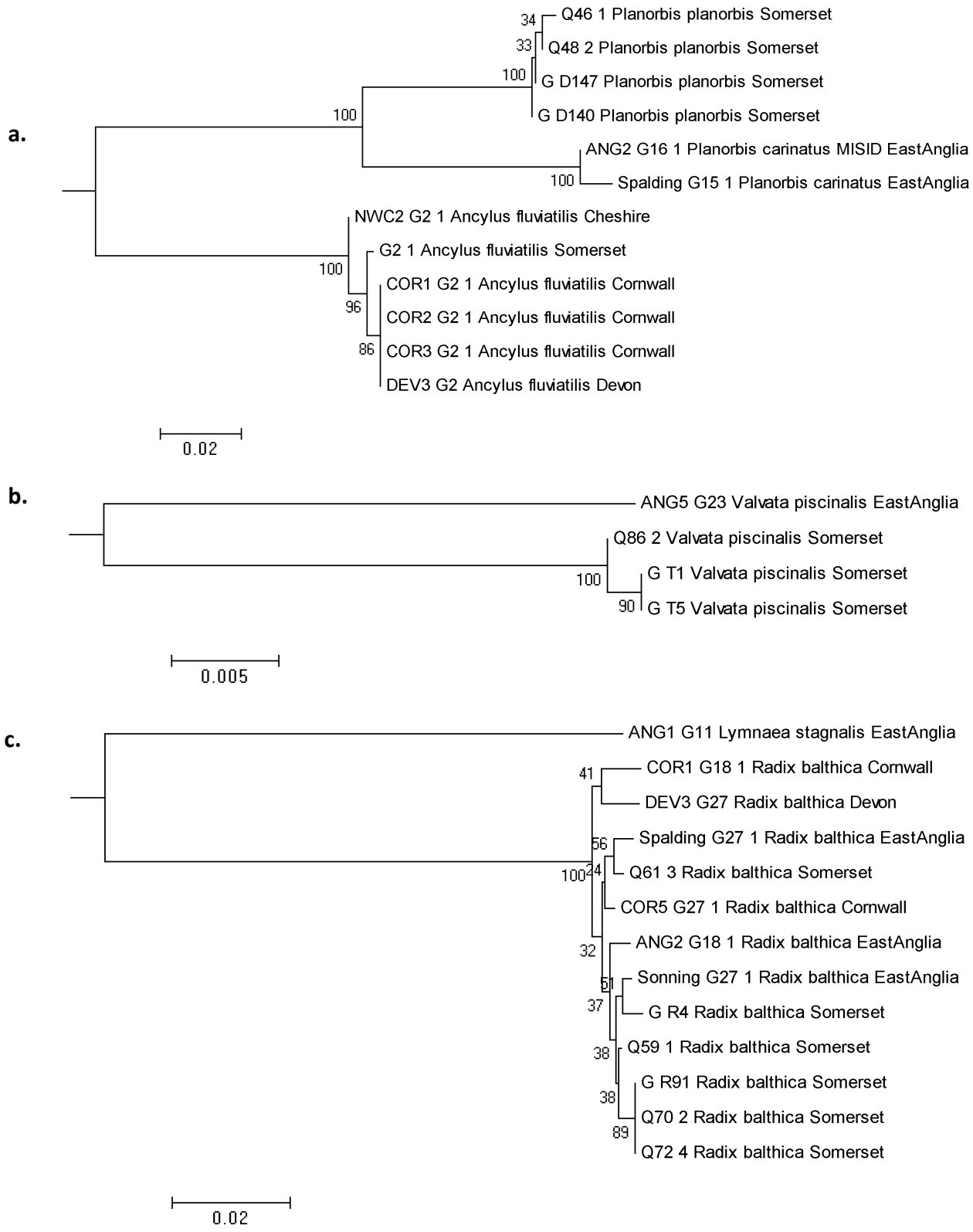
Supplementary Figure 2.2b: Maximum Likelihood phylogenetic tree for Trichoptera sequences, constructed with 500 bootstrap replications (K2P distances) (part 2).



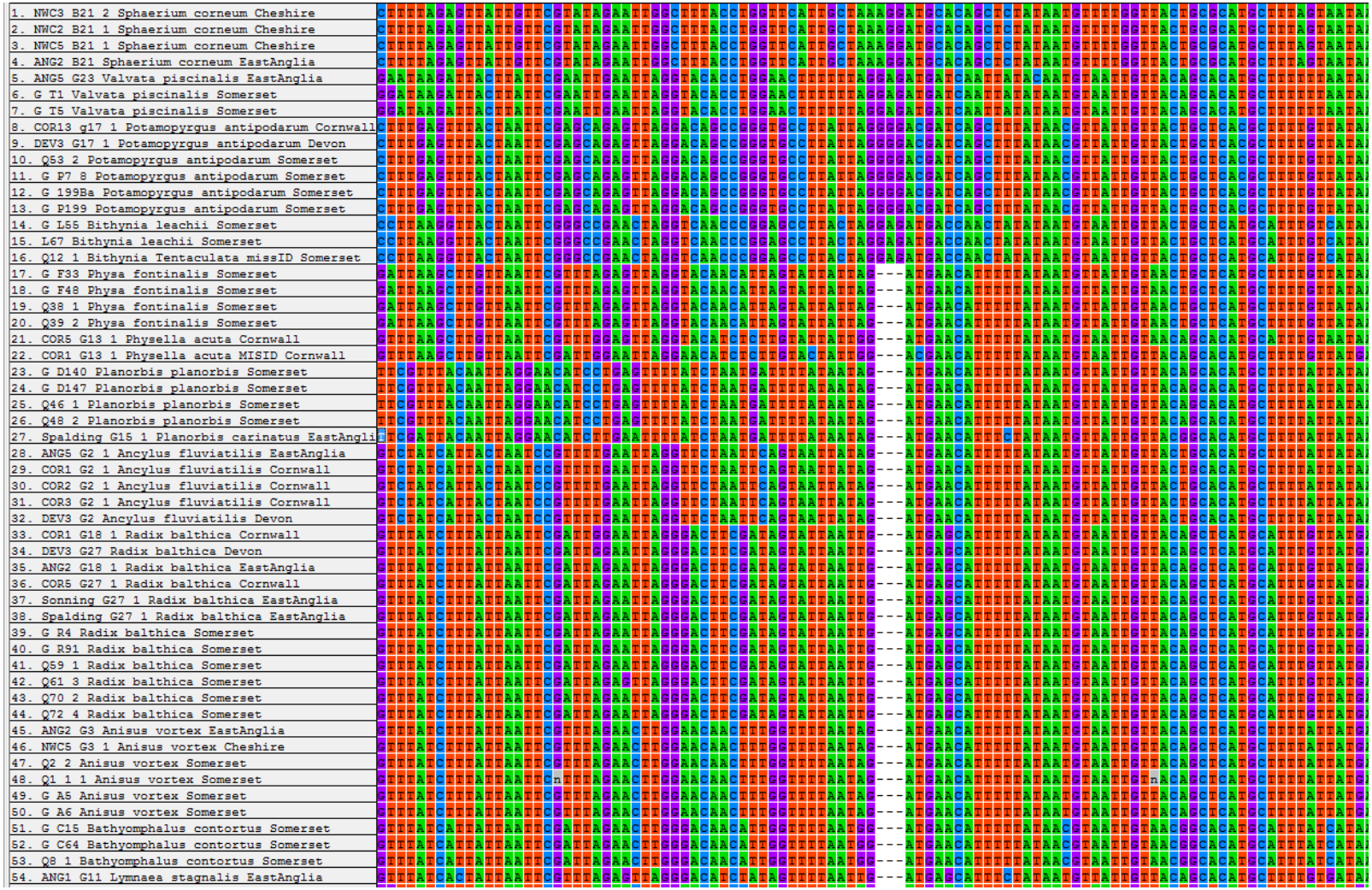
Supplementary Figure 2.3: Trichoptera species NJ subtrees.
For species a) *D. annulatus*, b.) *L. marmoratus*, c.) *S. palipes* and d.) *A. fuscipes*. Values on branches represent bootstrap support.



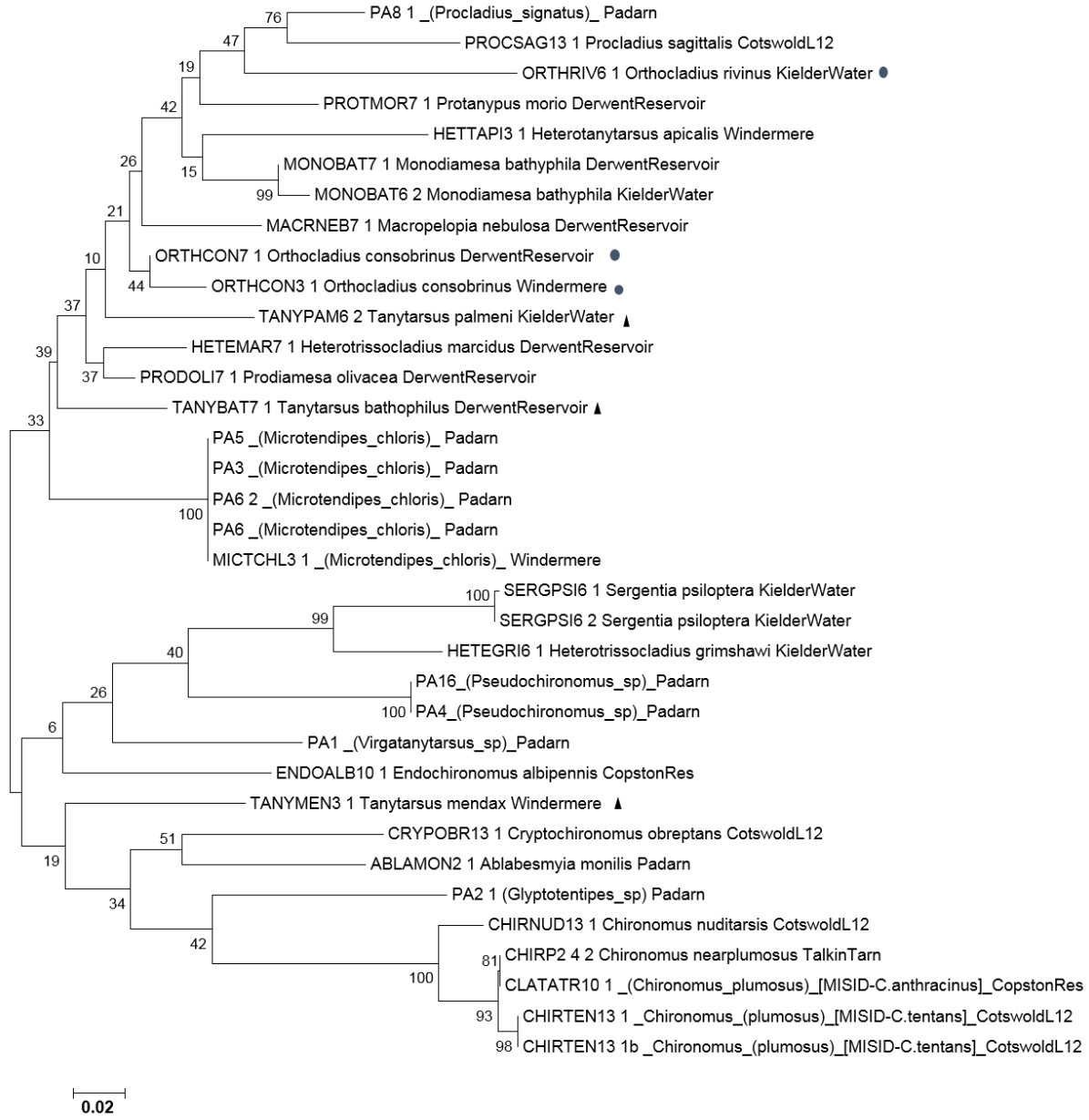
Supplementary Figure 2.4: Maximum Likelihood phylogenetic tree for Gastropoda sequences, constructed with 500 bootstrap replications (K2P distances).



Supplementary Figure 2.5: **Gastropoda species NJ subtrees.**
For species a.) *A. fluviatilis*, b.) *R. balthica* and c.) *V. piscinalis.* Values on branches represent bootstrap support.



Supplementary Figure 2.6: Alignment of Gastropoda sequences.
Sequences are showing a 3bp insertion occurring in all members of the Planorbidae, Physidae and Limneidae families. Lines 1-16 sequences from remaining families, which do not present the deletion (ClustalW, MEGA5).



Supplementary Figure 2.7: Maximum Likelihood phylogenetic tree for Chironomidae.
Tree constructed with 500 bootstrap replications (K2P distances).
Non-monophyletic genera are marked (triangles: Tanytarsus, circles: Orthocladius).

References

Anderson, R. (2005). An annotated list of the non-marine Mollusca of Britain and Ireland. *Journal of Conchology*, **38**, 607–637.

Armitage, P.D., Pinder, L.C. & Cranston, P. (2012). *The Chironomidae: biology and ecology of non-biting midges*. Chapman and Hall, London.

Baird, D.J. & Sweeney, B.W. (2011). Applying DNA barcoding in benthology: the state of the science. *Journal of the North American Benthological Society*, **30**, 122–124.

Bensasson, D., Zhang, D., Hartl, D.L. & Hewitt, G.M. (2001). Mitochondrial pseudogenes: evolution ' s misplaced witnesses. *TRENDS in Ecology & Evolution*, **16**, 314–321.

Bergsten, J., Bilton, D.T., Fujisawa, T., Elliott, M., Monaghan, M.T., Balke, M., Hendrich, L., Geijer, J., Herrmann, J., Foster, G.N., Ribera, I., Nilsson, A.N., Barraclough, T.G. & Vogler, A.P. (2012). The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology*, **61**, 851–869.

Borges, L.M.S., Hollatz, C., Lobo, J., Cunha, A.M., Vilela, A.P., Calado, G., Coelho, R., Costa, A.C., Ferreira, M.S.G., Costa, M.H. & Costa, F.O. (2016). With a little help from DNA barcoding: investigating the diversity of Gastropoda from the Portuguese coast. *Scientific reports*, **6**, 20226.

Brodin, Y., Ejdung, G., Strandberg, J. & Lyrholm, T. (2013). Improving environmental and biodiversity monitoring in the Baltic Sea using DNA barcoding of Chironomidae (Diptera). *Molecular Ecology Resources*, **13**, 996–1004.

Cameron, S.L., Rubinoff, D. & Will, K. (2006). Who will actually use DNA barcoding and what will it cost? *Systematic biology*, **55**, 844–847.

Carew, M., Pettigrove, V., Metzeling, L. & Hoffmann, A. (2013). Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology*, **10**, 45.

Collins, A., Ohandja, D.G., Hoare, D. & Voulvoulis, N. (2012). Implementing the Water Framework Directive: A transition from established monitoring networks in England and Wales. *Environmental Science and Policy*, **17**, 49–61.

Costa, F.O. & Carvalho, G.R. (2010). New insights into molecular evolution: Prospects from the barcode of life initiative (BOLI). *Theory in Biosciences*, **129**, 149–157.

Costa, F.O. & Carvalho, G.R. (2007). The Barcode of Life Initiative : synopsis and prospective societal impacts of DNA barcoding of Fish. *Genomics Society and Policy*, **3**, 29–40.

Costa, F.O., Landi, M., Martins, R., Costa, M.H., Costa, M.E., Carneiro, M., Alves, M.J., Steinke, D. & Carvalho, G.R. (2012). A Ranking System for Reference Libraries of DNA Barcodes: Application to Marine Fish Species from Portugal (R. DeSalle, Ed.). *PLoS ONE*, **7**, e35858.

Cranston, P.S. (1990). Biomonitoring and invertebrate taxonomy. *Environmental Monitoring and Assessment*, **14**, 265–273.

Creer, S., Fonseca, V.G., Porazinska, D.L., Giblin-Davis, R.M., Sung, W., Power, D.M., Packer, M., Carvalho, G.R., Blaxter, M.L., Lambshead, P.J.D. & Thomas, W.K. (2010). Ultrasequencing of the meiofaunal biosphere: Practice, pitfalls and promises. *Molecular Ecology*, **19**, 4–20.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., Taberlet, P., Taberlet, P., Coissac, E., Hajibabaei, M., Rieseberg, L., Yu, D., Ji, Y., Emerson, B., Wang, X., Ye, C., Yang, C., Ding, Z., Ficetola, G., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessiere, J., Taberlet, P., Pompanon, F., Geller, J., Meyer, C., Parker, M., Hawk, H., Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glockner, F., Bru, D., Martin-Laurent, F., Philippot, L., Schloss, P., Gevers, D., Westcott, S., Clarke, L., Soubrier, J., Weyrich, L., Cooper, A., Ji, Y., Barba, M. De, Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., Taberlet, P., Leray, M., Yang, J., Meyer, C., Mills, S., Agudelo, N., Ranwez, V., Boehm, J., Machida, R., Little, D., Deagle, B., Kirkwood, R., Jarman, S., Zhou, X., Shokralla, S., Gibson, J., Nikbakht, H., Janzen, D., Hallwachs, W. & Hajibabaei, M. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology letters*, **10**, 1789–1793.

Demin, A.G., Polukonova, N. V. & Mugue, N.S. (2011). Molecular phylogeny and the time of divergence of minges (Chironomidae, Nematocera, Diptera) inferred from a partial nucleotide sequence of the cytochrome oxidase I gene (COI). *Russian Journal of Genetics*, **47**, 1168–1180.

Dick, M.W. (1970). Saprolegniaceae on insect exuviae. *Transactions of the British Mycological Society*, **55**, 449–458.

Dobson, S.L., Bourtzis, K., Braig, H.R., Jones, B.F., Zhou, W., Rousset, F. & O'Neill, S.L. (1999). Wolbachia infections are distributed throughout insect somatic and germ line tissues. *Insect Biochemistry and Molecular Biology*, **29**, 153–160.

Dunigan, E.P. (1988). *Biological Indicators of Freshwater Pollution and Environmental Management*. Springer Netherlands, Dordrecht.

Ekrem, T., Stur, E. & Hebert, P.D.N. (2010). Females do count: Documenting chironomidae (Diptera) species diversity using DNA barcoding. *Organisms Diversity and Evolution*, **10**, 397–408.

Ekrem, T., Willassen, E. & Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution*, **43**, 530–542.

Elder, J.F. & Collins, J.J. (1991). Freshwater molluscs as indicators of bioavailability and toxicity of metals in surface-water systems. *Rev Environ Contam Toxicol*, **122**, 37–79.

Ferrington JLC, Blackwood MA, Wright CA, Crisp NH, Kavanaugh JL, Sc.F. (1991). A protocol for using surface-floating pupal exuviae of Chironomidae for rapid bio assessment of changing water quality. *Sediment and stream water quality in a changing environment: trends and explanation*, pp. 181–190. International Association of Hydrological Sciences Publication.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan

invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

Geraci, C.J., Al-Saffar, M. a. & Zhou, X. (2011). DNA barcoding facilitates description of unknown faunas: a case study on Trichoptera in the headwaters of the Tigris River, Iraq. *Journal of the North American Benthological Society*, **30**, 163–173.

Goldberg, C.S., Sepulveda, A., Ray, A., Baumgardt, J. & Waits, L.P. (2013). Environmental DNA as a new method for early detection of New Zealand mudsnails (*Potamopyrgus antipodarum*). *Freshwater Science*, **32**, 792–800.

Gray, C., Bista, I., Creer, S., Demars, B.O.L., Falciani, F., Don, T.M., Sun, X. & Woodward, G. (2015). Freshwater conservation and biomonitoring of structure and function: Genes to ecosystems. *Aquatic Functional Biodiversity: An Ecological and Evolutionary Perspective* (eds A. Belgrano, G. Woodward & U. Jacob), pp. 241–271. Elsevier.

Gunderina, L.I. (2010). Species-specific PCR primers for identification of the sibling species Chironomus plumosus (Linnaeus, 1758) and Chironomus balatonicus (Devai, Wuelker et Scholl, 1983) (Chironomidae, Diptera). *Fauna Norvegica*, **31**, 151–157.

Gunderina, L.I., Kiknadze I I, Istomina, A.G. & Butler, M. (2009). Geographic differentiation of genomic DNA of Chironomus plumosus (Diptera, Chironomidae) in natural holarctic populations. *Russian Journal of Genetics*, **45**, 54–62.

Hajibabaei, M., Singer, G.A.C., Hebert, P.D.N. & Hickey, D.A. (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, **23**, 167–172.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & deWaard, J.R. (2003). Biological identifications through DNA barcodes. *Proceedings. Biological sciences / The Royal Society*, **270**, 313–21.

Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A. & Werren, J.H. (2008). How many species are infected with Wolbachia? - A statistical analysis of current data. *FEMS Microbiology Letters*, **281**, 215–220.

Hopkins, S.R., Wyderko, J.A., Sheehy, R.R., Belden, L.K. & Wojdak, J.M. (2013). Parasite predators exhibit a rapid numerical response to increased parasite abundance and reduce transmission to hosts. *Ecology and Evolution*, **3**, 4427–4438.

Hubert, N. & Hanner, R. (2015). DNA Barcoding, species delineation and taxonomy: a historical perspective. *DNA Barcodes*, **3**, 44–58.

Hurst, G.D.D. & Jiggins, F.M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings. Biological sciences / The Royal Society*, **272**, 1525–1534.

Joly, S., Davies, T.J., Archambault, A., Bruneau, A., Derry, A., Kembel, S.W., Peres-Neto, P., Vamosi, J. & Wheeler, T.A. (2014). Ecology in the age of DNA barcoding: The resource, the promise and the challenges ahead. *Molecular Ecology Resources*, **14**, 221–232.

Kenney, M.A., Sutton-Grier, A.E., Smith, R.F. & Gresens, S.E. (2009). Benthic macroinvertebrates as indicators of water quality: The intersection of science and policy. *Terrestrial Arthropod Reviews*, **2**, 99–128.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Kjer, K.M., Blahnik, R.J. & Holzenthal, R.W. (2001). Phylogeny of Trichoptera (caddisflies): characterization of signal and noise within multiple datasets. *Systematic biology*, **50**, 781–816.

Kranzfelder, P., Ekrem, T. & Stur, E. (2016). Trace DNA from insect skins: A comparison of five extraction protocols and direct PCR on chironomid pupal exuviae. *Molecular Ecology Resources*, **16**, 353–363.

Kress, W.J., García-Robledo, C., Uriarte, M. & Erickson, D.L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology and Evolution*, **30**, 25–35.

Krosch, M.N., Baker, A.M., Mather, P.B. & Cranston, P.S. (2011). Systematics and biogeography of the Gondwanan Orthocladiinae (Diptera: Chironomidae). *Molecular Phylogenetics and Evolution*, **59**, 458–468.

Layton, K.K.S., Martel, A.L. & Hebert, P.D.N. (2014). Patterns of DNA barcode variation in canadian marine molluscs. *PLoS ONE*, **9**, 1–9.

Lenat, D.R. & Resh, V.H. (2001). Taxonomy and stream ecology - The benefits of genus- and species-level identifications. *J. N. Am. Benthol. Soc.*, **20**, 287–298.

Meyer, C.P. & Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology*, **3**, 1–10.

Morse, J.C. (1997). Phylogeny of Trichoptera. *Annual Review of Entomology*, **42**, 427–450.

Nei, M. & Kumar, S. (2000). *Molecular Ecology and Phylogenetics*. Oxford University Press, New York.

Oreska, M.P.J. & Aldridge, D.C. (2011). Estimating the financial costs of freshwater invasive species in Great Britain: A standardized approach to invasive species costing. *Biological Invasions*, **13**, 305–319.

Park, G.S. & Marshall, H.G. (2000). The Trophic Contributions of Rotifers in Tidal Freshwater and Estuarine Habitats. *Estuarine, Coastal and Shelf Science*, **51**, 729–742.

Pfenninger, M., Nowak, C., Kley, C., Steinke, D. & Streit, B. (2007). Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic Chironomus (Diptera) species. *Molecular Ecology*, **16**, 1957–1968.

Pfenninger, M., Salinger, M., Haun, T. & Feldmeyer, B. (2011). Factors and processes shaping the population structure and distribution of genetic variation across the species range of the freshwater snail Radix balthica (Pulmonata, Basommatophora). *BMC Evolutionary Biology*, **11**, 135.

Pilgrim, E.M., Jackson, S. a., Swenson, S., Turcsanyi, I., Friedman, E., Weigt, L. & Bagley, M.J. (2011). Incorporation of DNA barcoding into a large-scale biomonitoring program: opportunities and pitfalls. *Journal of the North American Benthological Society*, **30**, 217–231.

Polukonova, N. V. & Beljanina, S.I. (2002). On the Possibility of Hybridogenesis in the Origin of Midge Chironomus usenicus Loginova et Beljanina (Chironomidae, Diptera). *Russian Journal of Genetics*, **38**, 1385–1390.

Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, **21**, 1864–1877.

Ratnasingham, S. & Hebert, P.D.N. (2007). BOLD: The Barcode of Life Data System: Barcoding. *Molecular Ecology Notes*, **7**, 355–364.

Raunio, J., Heino, J. & Paasivirta, L. (2011). Non-biting midges in biodiversity conservation and environmental assessment: Findings from boreal freshwater ecosystems. *Ecological Indicators*, **11**, 1057–1064.

Remigio, E.A. & Hebert, P.D.N. (2003). Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. *Molecular Phylogenetics and Evolution*, **29**, 641–647.

Ruiter, D.E., Boyle, E.E. & Zhou, X. (2013). DNA barcoding facilitates associations and diagnoses for Trichoptera larvae of the Churchill (Manitoba, Canada) area. *BMC ecology*, **13**, 5.

Ruse, L.P. (2013). Chironomid (Diptera) species recorded from UK lakes as pupal exuviae. *Journal of Entomological and Acarological Research*, **45**, 13.

Ruse, L. (2010). Classification of nutrient impact on lakes using the chironomid pupal exuvial technique. *Ecological Indicators*, **10**, 594–601.

Ruse, L. (2011). Lake acidification assessed using chironomid pupal exuviae. *Fundamental and Applied Limnology / Archiv für Hydrobiologie*, **178**, 267–286.

S.A.Miller, D.D.D. and H.F.P. (1988). A simple salting out procedure for extractin DNA from humam nnucleated cells. *Nucleic acids research*, **15**, 1215.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, **4**, 406–25.

Schmidt-Kloiber, A. & Nijboer, R.C. (2004). The effect of taxonomic resolution on the assessment of ecological water quality classes. *Hydrobiologia*, **516**, 269–283.

Smith, M.A., Bertrand, C., Crosby, K., Eveleigh, E.S., Fernandez-Triana, J., Fisher, B.L., Gibbs, J., Hajibabaei, M., Hallwachs, W., Hind, K., Hrcek, J., Huang, D.W., Janda, M., Janzen, D.H., Li, Y., Miller, S.E., Packer, L., Quicke, D., Ratnasingham, S., Rodriguez, J., Rougerie, R., Shaw, M.R., Sheffield, C., Stahlhut, J.K., Steinke, D., Whitfield, J., Wood, M. & Zhou, X. (2012). Wolbachia and DNA barcoding insects: Patterns, potential, and problems. *PLoS ONE*, **7**, e36514.

Smith, S. A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C.S., Rouse, G.W., Giribet, G. & Dunn, C.W. (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, **480**, 364–367.

Song, H., Buhay, J.E., Whiting, M.F. & Crandall, K.A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial

pseudogenes are coamplified. *Proceedings of the National Academy of Sciences*, **105**, 13486–13491.

Stein, E.D., Martinez, M.C., Stiles, S., Miller, P.E. & Zakharov, E. V. (2014). Is DNA barcoding actually cheaper and faster than traditional morphological methods: Results from a survey of freshwater bioassessment efforts in the United States?. *PLoS ONE*, **9**, e95525.

Stein, E.D., White, B.P., Mazor, R.D., Miller, P.E. & Pilgrim, E.M. (2013). Evaluating Ethanol-based Sample Preservation to Facilitate Use of DNA Barcoding in Routine Freshwater Biomonitoring Programs Using Benthic Macroinvertebrates. *PLoS ONE*, **8**, e51273.

Strong, E.E., Gargominy, O., Ponder, W.F. & Bouchet, P. (2008). Global diversity of gastropods (Gastropoda; Mollusca) in freshwater. *Hydrobiologia*, **595**, 149–166.

Sunnucks, P. & Hales, D.F. (1996). Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus Sitobion (Hemiptera: Aphididae). *Molecular biology and evolution*, **13**, 510–524.

Sweeney, B.W., Battle, J.M., Jackson, J.K. & Dapkey, T. (2011). Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society*, **30**, 195–216.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Valentini, A., Pompanon, F. & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology and Evolution*, **24**, 110–117.

Wallace, I.. (1991). A review of the Trichoptera of Great Britain. 59.

Wiberg-Larsen, P. (2008). Overall distributional patterns of European Trichoptera. *Ferrantia*, **55**, 143–154.

Wiemers, M. & Fiedler, K. (2007). Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in zoology*, **4**, 8.

Wilson, R.S & McGill, J.D. (1979). The use od chironomid pupal exuviae for biological surveillance of water quality.

Wilson, R. & Ruse, L. (2005). *A guide to the identification of genera of chironomid pupal exuviae occurring in Britain and Ireland*. Freshwater Biological Association Special Publication no. 13.

Zhou, X., Adamowicz, S.J., Jacobus, L.M., Dewalt, R.E. & Hebert, P.D. (2009). Towards a comprehensive barcode library for arctic life - Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. *Frontiers in zoology*, **6**, 30.

Zhou, X., Robinson, J.L., Geraci, C.J., Parker, C.R., Flint, O.S., Etnier, D.A., Ruiter, D., DeWalt, R.E., Jacobus, L.M. & Hebert, P.D.N. (2011). Accelerated construction of a regional DNA-barcode reference library: caddisflies (Trichoptera) in the Great Smoky Mountains

National Park. *Journal of the North American Benthological Society*, **30**, 131–162.

Chapter 3

Annual time series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity

---

"If you don't hope, you won't find the impossible;
that which is hidden and unexplored"
Heraclitus

Chapter 3: Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity.

3.1 Abstract

The use of environmental DNA (eDNA) in biodiversity assessments offers a step-change in sensitivity, throughput and simultaneous measures of ecosystem diversity and function. There remains, however, a need to examine eDNA persistence in the wild through simultaneous temporal measures of eDNA and biota. We used metabarcoding of two markers of different lengths, derived from an annual time-series of aqueous lake eDNA to examine temporal shifts in ecosystem biodiversity and in an ecologically important group of macroinvertebrates (Diptera: Chironomidae). The analyses allow different levels of detection and validation of taxon richness and community composition (β-diversity) through time, with shorter eDNA fragments dominating the eDNA community. Comparisons between eDNA, community DNA, taxonomy and UK species abundance data further show significant relationships between diversity estimates derived across the disparate methodologies. Our results reveal the temporal dynamics of eDNA and validate the utility of eDNA metabarcoding for tracking seasonal diversity at the ecosystem scale.

Note:

This chapter has been submitted for publication to the journal Nature Communications, and is presented here with the formatting required for submission in this journal, which requires that the methods section is presented at the end of the document.

Co-authors:

Iliana Bista, Gary R. Carvalho, Kerry Walsh, Mathew Seymour, Mehrdad Hajibabaei, Delphine Lallias, Martin Christmas, Simon Creer

3.2 Introduction

The maintenance of biodiversity underpins the stability of ecosystem processes in constantly changing environments (Loreau & de Mazancourt 2013). Consequently, biodiversity loss not only affects ecosystem function and services, but also society as a whole (Cardinale *et al.* 2012). One major impediment for elucidating the relationship between biodiversity and ecosystem health is a need for robust and detailed understanding of biodiversity processes and dynamics in time and space (Thomsen & Willerslev 2015). To halt or reverse contemporary species loss and habitat degradation, there is a need for increasingly reliable and cost effective methods for biodiversity assessment, since widely employed traditional approaches fall short in many cases (Lawson Handley 2015). Currently, species identification of individuals at immature life stages and among closely related species is difficult and requires high-level, labour-intensive taxonomic expertise, thereby rendering large scale ecosystem-wide assessments expensive, time consuming and potentially unrepresentative of the ecosystem sampled (Yu *et al.* 2012). However, recent advancements in molecular detection techniques, most notably the application of environmental DNA (eDNA), offer exciting new opportunities to improve existing biodiversity assessment procedures.

Environmental DNA (eDNA) is DNA extracted directly from an environmental sample (e.g., water, soil or air), without prior isolation of the organisms themselves (Dejean *et al.* 2011). Sources of eDNA include sloughed skin cells, urine, faeces, saliva or other bodily secretions (Rees *et al.* 2014), and consist of both free molecules (extracellular DNA) and free cells (Barnes & Turner 2016). Furthermore, eDNA collected from water samples has highly sensitive detection capability and is non-invasive to the sampled biota (Bohmann *et al.* 2014), thereby potentially improving environmental management and assessment of freshwater ecosystems (Kelly *et al.* 2014b; Lawson Handley 2015).

Previous work with eDNA of aquatic invertebrates is dominated by targeted PCR-based approaches (e.g. qPCR), which are limited in assessing biodiversity (Goldberg *et al.* 2013; Mächler *et al.* 2014; Deiner & Altermatt 2014). However, high throughput sequencing (HTS) applications, such as metabarcoding, are already advancing prospects in ecology (Chave

2013), offering comprehensive and efficient tools for measuring and assessing total biodiversity (Ji *et al.* 2013). High throughput sequencing has successfully been used for sequencing whole communities of invertebrates (bulk samples) (Hajibabaei *et al.* 2011; Gibson *et al.* 2014, 2015), though only a few studies have employed metabarcoding of aqueous eDNA (Valentini *et al.* 2016; Hänfling *et al.* 2016). Additionally, most aqueous eDNA studies have focused on macroorganisms, including fish and amphibians (Evans *et al.* 2016; Valentini *et al.* 2016; Hänfling *et al.* 2016). The limited number of studies which have addressed invertebrate detection, only targeted specific species such as for example two arthropod species in Thomsen *et al.* (2012b), four invertebrate species in (Deiner *et al.* 2015) (but see recently published work on river macroinvertebrate diversity by Deiner *et al.* 2016). Nevertheless, the combination of HTS and eDNA is poised to become a prominent tool for ecosystem assessment (Thomsen *et al.* 2012b; Kelly *et al.* 2014b) by simultaneously assessing a plethora of organisms, including associated organism interactions, with a throughput sufficient for rapid whole community assessment.

Regardless of the increasing number of eDNA studies, several factors of eDNA research demand clarification, including persistence of eDNA (Lodge *et al.* 2012). Persistence of eDNA is the time that eDNA remains detectable (e.g., in the water) after removal or loss of the organism from the environment, which influences the timeframe for biodiversity assessment(Dejean *et al.* 2011). Investigating the temporal relationship between community DNA and eDNA is vital, since accurate (extant) biodiversity assessment requires detection of contemporary, and ecologically relevant, biodiversity. The persistence of eDNA for several different species has been studied mainly in artificial systems, including aquaria and mesocosms (Dejean *et al.* 2011; Thomsen *et al.* 2012b; Goldberg *et al.* 2013; Strickler *et al.* 2015). Notably, persistence of short eDNA fragments, in artificial environments, was found to vary between days to weeks after removal of the study organisms, depending upon biotic and abiotic factors (Barnes *et al.* 2014).

Species identity by eDNA is typically undertaken by detection of short DNA fragments (Rees *et al.* 2014), a practise possibly influenced by ancient DNA work, which utilises highly fragmented DNA (Taberlet *et al.* 2012b). For the detection of rare and evasive species, short DNA fragments might indeed increase detection, although with some risk of errors if not

properly analysed. Possible biases when using short fragments include inadvertently sampling old eDNA fragments which have demonstrated remarkable persistence (Barnes & Turner 2016), especially when bound to sediments where degradation rate is slower, due to protection of DNA molecules and inactivation of extracellular nucleases (Barnes *et al.* 2014). Conversely, DNA fragments of several hundred base pairs length are less likely to persist long after release into the environment due to rapid degradation (Lindahl 1993) and may represent a less abundant, but more contemporary, biodiversity signal (Deagle *et al.* 2006).

While the ecological value of collecting temporal data is established, most ecological studies focus on spatial data (Magurran *et al.* 2010). Similarly, many existing eDNA studies have focused on spatial detection, such as early detection of invasive species (Dejean *et al.* 2012; Goldberg *et al.* 2013) and rare, or endangered species (Biggs *et al.* 2015). Temporal estimates have been relatively neglected by eDNA studies (but see Biggs *et al.,* 2015 for repeated seasonal sampling), and an understanding of temporal relationships between eDNA and community biodiversity remains a knowledge gap (Thomsen & Willerslev 2015). Additionally, there are no published studies, to our knowledge, employing temporally collected data that incorporate seasonal variation across an annual cycle from aqueous eDNA for ecosystem-wide biodiversity level analysis.

Furthermore, overall ecosystem biodiversity characterisation, using indicator taxonomic groups, can facilitate comparisons between taxonomically identified biodiversity over time (e.g. collection of invertebrate samples) and eDNA detection. One such indicator group is the Chironomidae or non-biting midges (Diptera: Chironomidae), which exhibit specialised responses to ecological stressors and are acknowledged as one of the most important macroinvertebrate groups for monitoring lake ecosystem health (Wilson & Ruse 2005; Ruse 2011). Importantly, samples can be collected after adult emergence in the form of shed skins of the pupae (pupal exuviae) that float on the water surface. The exuviae technique allows for integrated sampling of lake ecosystems from all aquatic microhabitats of the lake, and sample identification can yield insights on ecosystem-wide biodiversity (Wilson & Ruse 2005).

Accordingly, here we a.) Investigate whether metabarcoding of lake eDNA is effective for the detection of community diversity and temporal shifts in an ecologically important sentinel group of macroinvertebrates, via comparison to the molecular and morphological analysis of chironomid exuvial bulk samples. b.) Investigate the use of eDNA analyses for characterising whole-ecosystem biodiversity patterns and c.) Explore the effects of amplicon length on detection of contemporary diversity. Collectively, we examine the ecological relevance of eDNA by exploring mechanisms underpinning the temporal dynamics of eDNA and the biological community at the ecosystem scale in nature.

3.3 Results

3.3.1 Sequencing results

After stringent filtering and quality control, 13,100,236 reads were obtained for: 1.) the full-length COI barcoding region (658bp) (amplicon COIF 6,659,598 reads) and 2.) a 235bp fragment on the 5'region of the COI barcoding region (amplicon COIS 6,440,638 reads), from 32 samples comprising 16 eDNA and 16 invertebrate community DNA samples. Data for these two amplicons were extracted from a larger dataset including additional amplicon libraries, sequenced on two lanes of MiSeq. Overall, the eDNA samples (extracted from filtered water samples) achieved good sequence coverage (mean number of reads per sample (±SD): COIF: 269,769 ± 57,427; COIS: 259,723 ± 85,437) (for exact number of reads per sample, see Supplementary Table ST1). Some of the community DNA samples that contained only small amounts of pupal exuviae resulted in a lower number of reads for both amplicons.

3.3.2 Control samples

During PCR screening of negative controls, no band (no amplification) was observed on agarose gels. Regardless of no visual proof of amplification, each sample was sequenced and a very low number of reads was returned (Supplementary Results SR1). The positive controls yielded good results for both amplicons, with 547,730 (COIS) and 393,341 (COIF) reads after quality control. Detection success was 100% for COIS (all 30 species detected)

and 87% for COIF (26 species detected) (Supplementary Results SR1, Supplementary Table ST2). BLAST identification and screening of positive control reads resulted in >99.9% of the reads being assigned to the target species known to be present in the positive control. The relative abundance of OTUs found in the positive control which were attributed to non-target taxa was 0.026% for the COIS and 0.007% for the COIF (Supplementary Table ST3).

3.3.3 Abundance filtering and rarefaction analysis

Following investigations of how screening different levels of abundance of rare OTUs affected overall OTU richness (including no filtering, and removal of OTUs that were present at less than 0.01% and 0.02%), a filtering level of 0.01% was set for all ecological analyses. Removal of OTUs present at less than 0.01% yielded equitable levels of OTU genus richness for the community DNA (37 genera) and eDNA (43 genera) according to 2014 Chironomidae records of Llyn Padarn (31 genera) (Fig. 1), and was within the limits of a small number of non-target reads detected in the positive control samples. The genus richness comparisons employed COIS data to ensure comparability between eDNA and community DNA for the Chironomidae below. According to the analysis of OTU accumulation curves versus sequence coverage, a rarefaction depth of 57,869 reads was applied across all water samples (Supplementary Fig. SF1a). To subsample Animalia OTUs in our samples a rarefaction depth of 24,914 reads per sample was used (Supplementary Fig. SF1b). These levels of rarefaction depth were selected based on a combination of accumulation curve results (Supplementary Fig. SF1, SF2) and the lowest number of reads achieved for a single sample.

3.3.4 Total taxonomic diversity

OTU clustering of the combined eDNA and community DNA datasets at 97% similarity cut-off (after removal of low abundance OTUs) yielded: 442 (eDNA) and 309 (community DNA) OTUs for COIF, and 482 (eDNA) and 394 (community DNA) OTUs for COIS. Taxonomic assignment through BLAST identified the majority of OTUs from Animalia and Protista (Supplementary Fig. SF3). From the eDNA samples, COIF identified 170 (35.3%) Animalia OTUs, of which 91 comprised Arthropoda (including 42 Insecta), whilst COIS identified 251

Animalia OTUs (56.8%), of which 212 were Arthropoda (including 167 Insecta) (Supplementary Fig. SF4). For the community DNA samples, COIF detected 219 (43.6%) Animalia OTUs, of which 171 were Arthropoda (including 132 Insecta), whilst COIS recovered 227 (73.5%) Animalia OTUs, of which 212 consisted of Arthropoda (including 184 Insecta).

Although not the focus of the study, metabarcoding of the eDNA samples (COIS used here as an example) also yielded matches to fish (*Phoxinus phoxinus*), amphibian and terrestrial OTUs represented at high read frequencies or distributed across numerous independent samples. Of the terrestrial taxa, spider OTUs from the Segestriidae (3,753 reads) and Thomisidae (1,858 reads) families, a millipede OTU (7,312 reads), orthopteran OTU (14,237 reads) and 2,114 reads from *Bos taurus* were recovered from multiple samples throughout the year, in addition to a broader diversity of terrestrial groups represented at lower frequencies in the dataset.

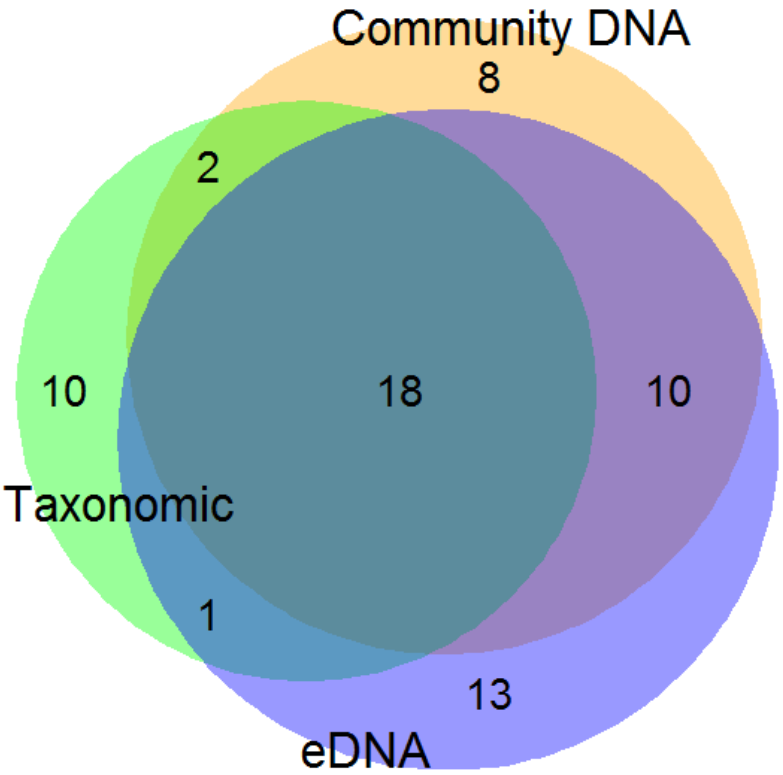


Figure 1: Number of Chironomidae genera per sample type.
The overlap area shows the number of genera common between sample types (purple: eDNA, orange: community DNA, green: taxonomic identification).

3.3.5 Temporal trends of OTU richness from eDNA samples (Total diversity)

Measures of OTU richness were calculated exclusively for eDNA samples and plotted against time to detect possible seasonal variations (Supplementary Fig. SF5). All samples were rarefied at an equal depth appropriate for each amplicon (total diversity dataset: 57,869 reads per sample, animal diversity dataset: 24,914 reads per sample, for all water samples).

Mean Animalia richness for COIS (±SD) was 37.8 (±10.4), and for COIF, 31.4 (±11.4) (Supplementary Fig. SF5a). A significant correlation was detected (Spearman's correlation, $p<0.05$) between the OTU Animalia richness estimates derived from COIF with time and temperature, but not with pH or dissolved oxygen (D.O.). Additionally, mean total richness for COIS (±SD) was 73.1 (±21.2), and for COIF, 88.1 (±26.9) (Supplementary Fig. SF5b). A significant correlation was detected (Spearman's correlation, $p<0.05$) between the COIF (total richness), time, temperature and D.O., but not pH. No significant correlation was found for COIS for the Animalia and total richness and any of the above parameters.

3.3.6 Community structure (β-diversity) from eDNA samples

We used eDNA samples to look into possible changes in community structure over time, for the Animalia identified diversity as well as the total diversity in the dataset. For the eDNA samples, nMDS analysis (Sørensen index) of total diversity for both amplicons (Fig. 2), delimited patterns of seasonal variations driving community composition. More biologically coherent patterns were presented by the COIF amplicon (Fig. 2a), while for COIS smaller subgroupings were also detected, including two outlier samples (Nov 25 & Dec 17). ANOSIM analyses also supported two main groupings, "winter" (Nov-April) and "summer" samples (April–Oct) (COIF: ANOSIM sig. level=0.1%, Global R = 0.717, COIS: ANOSIM sig. level = 0.2%, Global R = 0.475, with outlying samples from winter sampling). Additional analysis of the total diversity supports similar findings [two main groupings: "winter" (Nov-April) and "summer" samples (April–Oct) (COIF: ANOSIM sig. level=0.1%, Global R = 0.777, COIS: ANOSIM sig. level = 0.1%, Global R = 0.703)] (Supplementary Fig. SF6).

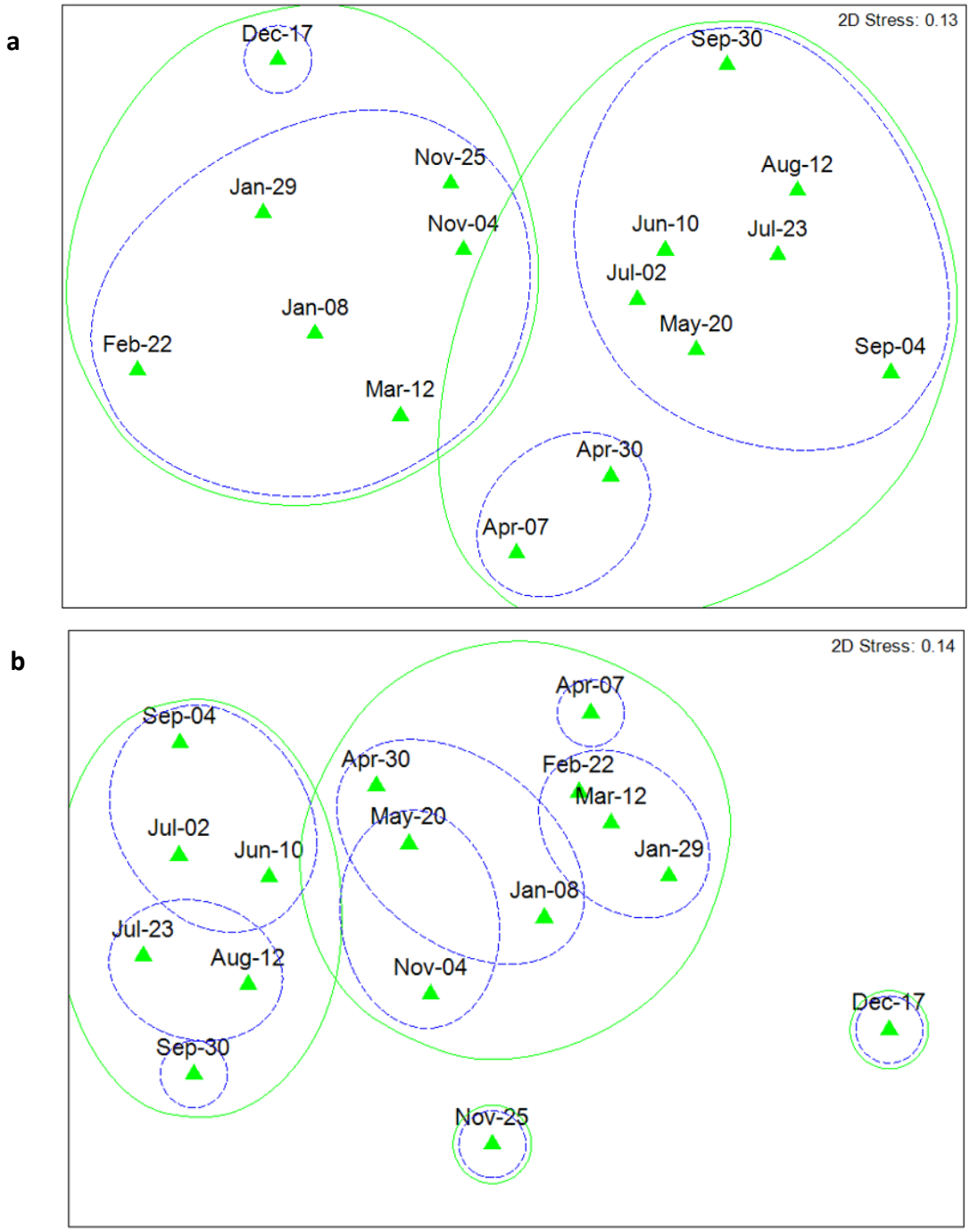


Figure 2: Animal eDNA β-diversity – nMDS (Sørensen index).
a. COIF, **b**. COIS amplicon (eDNA samples only) (N = 32). Solid green circles: 30% similarity cut-off (corresponding to "winter" –"summer" groups), dashed blue circles: 40% similarity cut-off.

3.3.7 Temporal trends in Chironomidae richness (community DNA and eDNA)

Analyses of un-trimmed COIF Chironomidae data suggested that temporal richness patterns between eDNA and community DNA samples were comparable to those of COIS (Spearman's $p<0.01$ correlation between eDNA and community DNA for COIF un-trimmed data) (Supplementary Figs. SF7). Nevertheless, the sequencing coverage of Chironomidae from the eDNA samples were approximately an order of magnitude lower than for COIS (Supplementary Fig. SF2). Subsequently, in order to maintain a sufficient sequencing depth across samples, COIF was not retained for further Chironomidae related analyses and rarefied incidence based data were used with 4,000 sequencing reads per sample for COIS only (Supplementary Fig. SF2).

For the Chironomidae assigned OTUs, COIS identified 103 OTUs from eDNA and 94 OTUs from community DNA samples (138 unique OTUs in total). Using a combination of BLAST ID ≥99% and the online Barcode of Life Database (BOLD) species assignment tool(Ratnasingham & Hebert 2007), 73 OTUs (53% out of 138 unique) were assigned species level taxonomic information. Analysis of historical species occurrence data collected by the Environment Agency (EA) (summer surveys 2003 – 2013) in Llyn Padarn (N. Wales, UK) indicated the presence of ≥99 Chironomidae species from 57 genera. Moreover, Fig.1 illustrates the qualitative overlap between the number of chironomid genera delimited by the current community DNA (65%), eDNA (61%) and taxonomy approaches.

To visualise the empirically derived annual diversity patterns, OTU and genus richness was assessed against time (Fig. 3) using a polynomial model. Observed OTU richness ranged from 5-27 OTUs for eDNA and 1-27 OTUs for community DNA over time (Fig. 3a). Conversely, genus level richness ranged from 5-19 for eDNA and 1-16 for community DNA. For the data derived from taxonomic identification of invertebrate (exuviae) community samples, genus level richness ranged from 10-18 (green points, restricted to 4 summer sampling times) (Fig 3b). Please also note that sampling points spanning the winter months (days 36 -190), which did not yield data, represented samples which contained very low physical numbers of exuviae. Consequently, they were not sequenced to an adequate depth in a mixed Illumina sequencing library, and could not be retained for analysis.

Significant associations were detected between time and Chironomidae OTU and genera richness derived from community DNA (OTU richness: R^2 =0.890, p-value = <0.01; Genera richness: R^2 = 0.849, p-value = 0.017). However Chironomidae OTU and genera richness derived from eDNA samples did not differ significantly over time (OTU: R^2 = 0.187, p-value = 0.460; Genera: R^2 = 0.128, p-value = 0.635) (Fig. 3). Taxonomic richness (genus level) also did not differ significantly over the limited time points available from seasonal sampling.

3.3.8 Temporal variation of OTU Abundance

We assessed the annual variation in OTU abundance from metabarcoding sequencing reads between eDNA and community DNA sampling methods using a generalized additive model (GAM). To allow across method comparisons we compared OTU abundances for Chironomidae OTUs occurring in both eDNA and community DNA datasets (45 OTUs). Abundances differed significantly among different OTUs (p-value <0.01) with a significant effect of the temporal smoothing term (p = 0.047) (Table 1). Additionally, abundances did not differ significantly between methods (p-value = 0.908), but a significant OTU identity x method interaction (p-value = 0.003) was found. The abundance of OTU reads was also found to be significantly positively correlated with expected species frequency (ranging from 0.01 to 0.79) across 97 sites in the United Kingdom (UK) (p-value = 0.003) (Table 1), using previously catalogued Chironomidae species frequency data (Ruse 2013) (Fig. 4).

Table 1: Generalized additive model (GAM).
The model explains OTU sequence abundance relative to OTU taxonomic ID (OTU) and sampling method (eDNA or community DNA - Method) over time. Model estimates and significances of the smoothing terms are given for the most parsimonious models. (R^2 = 0.18, df: degrees of freedom, edf: estimated degrees of freedom).

	df	F	p-value
OTU	44	4.688	<0.01
Method	1	0.013	0.908
OTU x Method	44	1.733	0.003
Approximate significance of smooth terms:	**edf**	**F**	**p-value**
s(Time)	2.899	2.561	0.047

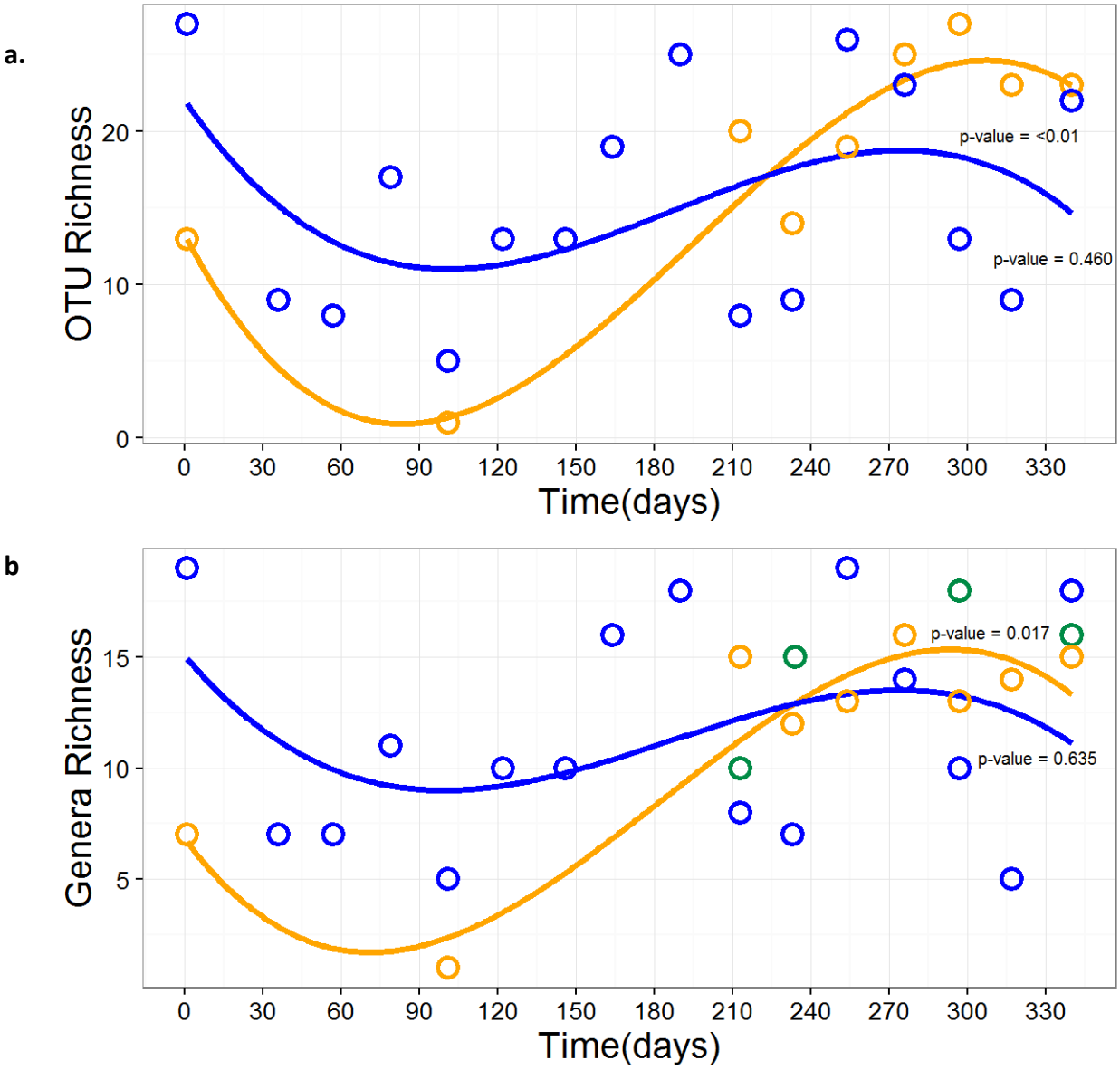


Figure 3: Richness patterns for Chironomidae OTUs and genera.
a.) OTU richness. **b**.) Genera richness. Points represent richness values to individual sampling points for eDNA (blue), community DNA (orange) and taxonomic identification of chironomid exuviae (green). Sampling points spanning the winter months (days 36 -190) did not yield data due to very low physical numbers of exuviae. Best fitted, significant lines from polynomial regressions for eDNA samples (blue) and community DNA (orange), plotted against time (x –axis: Sep. 2013 – Sep 2014).

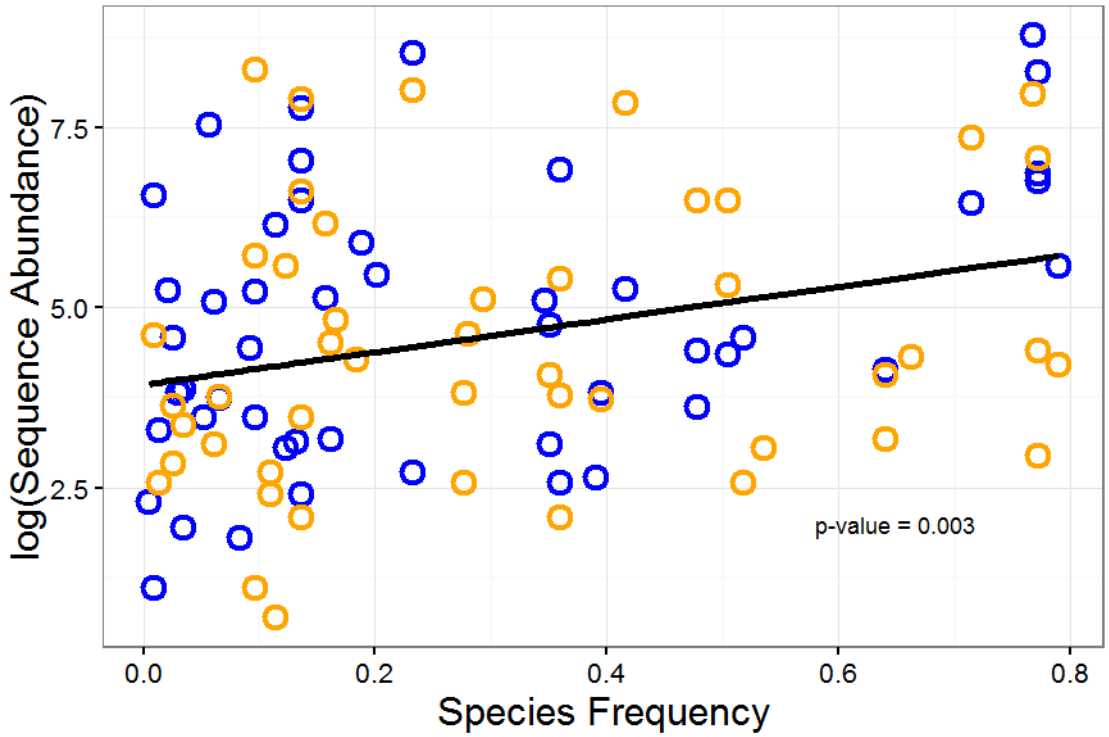


Figure 4: Sequence abundance patterns for Chironomidae OTUs against species frequency across the UK according to historical data, showing eDNA samples (blue) and community DNA (orange) along with the best fitted, significant, linear regression model (black line).

3.4 Discussion

We present here one of the first temporal studies of aqueous eDNA and community DNA biodiversity from a lake ecosystem, in addition to targeting a specific group of ecological sentinel macroinvertebrates. In contrast to previous analyses that have used PCR (qPCR) to infer presence/absence of a small number of target species (e.g., macroinvertebrates) from eDNA samples (Mächler *et al.* 2014; Deiner & Altermatt 2014), we employed HTS of amplicon libraries (metabarcoding) to assess temporal total biodiversity. Such methodology allows for the characterisation of the entire community, which is not possible through targeted individual-species sequencing that employs taxon specific primers. Simultaneously, we provide among the first accounts of temporally collected biodiversity data from an annual series of eDNA samples, compared with a series of invertebrate community DNA samples. Our findings yield an informative characterisation of temperate lake ecosystem-wide biodiversity, through detection of multiple groups of organisms from invertebrates to macro-organisms, of primarily freshwater, but also terrestrial origins. Furthermore, the biodiversity of the indicator taxon group used (Chironomidae) was successfully detected throughout the year, from both eDNA and community DNA samples, exhibiting substantial overlap with traditional taxonomy data. In addition, OTU sequence abundances were significantly positively associated with expected chironomid species abundance based on UK taxa occurrence data (Table 1, Fig. 4). Such direct coincidence, despite potential biotic and abiotic variability in the release, transport and persistence of eDNA (Barnes & Turner 2016), demonstrates the value of eDNA metabarcoding for biodiversity characterisation and ecosystem monitoring (Baird & Hajibabaei 2012).

Both metabarcoding amplicons detected large amounts of Animal phylum level diversity from eDNA samples, showing broad representation across the freshwater taxonomic biosphere, including the broadly studied Arthropoda (Supplementary Fig. SF4). Within the Arthropoda, the dominance of Insecta, Maxillopoda and Malacostraca (Crustacea) also demonstrates the utility of eDNA metabarcoding for characterisation of freshwater ecosystem-wide biodiversity. There is increasing exposure of the use of eDNA metabarcoding for the detection of fish and amphibians (Valentini *et al.* 2016; Hänfling *et al.* 2016), as also recorded here. However, a more novel concept is the ability of freshwater

systems to integrate biodiversity information from terrestrial sources. Terrestrial species found in our dataset, such as spider, millipede and orthopteran species, or the ubiquitous *Bos taurus* (domesticated cow), are all commonplace in the surrounding area of the study site and were detected by the analysis of eDNA residing in the lake water samples. The ability of freshwater catchments to contain eDNA from broader habitat biodiversity therefore presents an opportunity for further research regarding the relationship between aqueous eDNA and biodiversity at the landscape scale.

Focusing on the Chironomidae richness estimates derived from the analysis of the short COI fragment (Fig. 3), we can see that the COIS amplicon yielded 138 unique OTUs from both sample types throughout the year. The analysis of the COIS amplicon therefore provided valuable comparative qualitative and quantitative data both within the metabarcoding datasets and between the historically collected data for Llyn Padarn and the rest of the UK (Ruse 2013). Other eDNA studies have focused mainly on macro-organisms such as fish or amphibians whereby skin cells and mucus are a likely primary source of eDNA (Barnes & Turner 2016). While aquatic invertebrates such as chironomids are individually typically much smaller, the accumulated biomass of the community clearly produces sufficiently detectable and persistent amounts of eDNA (from natural shedding, moulting and death) for meaningful biodiversity assessment. Additional quantitative studies are required to determine the effects of invertebrate community biomass on levels of eDNA in environmental samples (Evans *et al.* 2016).

Sequencing of the complete COI region (COIF ~658bp) from eDNA samples was successful in detecting several genera of chironomids and provided biodiversity estimates comparable with community DNA biodiversity patterns (Supplementary Fig. SF7). However, it was not possible to retain the COIF locus throughout all analyses after applying strict abundance filtering of OTUs. Low sequence coverage of the COIF for the Chironomidae (primarily in the water eDNA and not the community DNA samples (Supplementary Fig. SF2) meant that more robust, ecological comparisons were more effectively achieved using the short eDNA fragment (COIS). Possible reasons for the discrepancies in coverage of the two amplicons could be related to variations in primer specificity, with the COIS primers being more successful than COIF primers in amplifying Chironomidae (Carew *et al.* 2013) (please also

see the limitations of the Folmer COI barcoding primers for metabarcoding analyses in Deagle *et al.* (2014)). Nevertheless, we did not detect substantial phylogenetic biases in OTUs recovered from the two primer pairs (Supplementary Fig. SF8) and coverage of the Chironomidae was only depleted in the water eDNA samples for the COIF. Alternatively, the discrepancy in different amplicon success may be due to the reduced availability of longer sized eDNA fragments in a natural ecosystem (Deagle *et al.* 2006).

After DNA is released into the environment, the degradation process likely begins, breaking down DNA and yielding shorter fragments. It has been shown that ~400bp length fragments remain detectable in water for days to weeks (Dejean *et al.* 2011; Goldberg *et al.* 2013), with the rate of degradation depending upon various biotic and abiotic factors (Barnes *et al.* 2014). Overall, smaller fragments degrade slower compared to longer fragments, suggesting an enhanced probability of detection by studies targeting shorter DNA fragments (Taberlet *et al.* 2012a). The present data support the enhanced detection of shorter eDNA fragments, as evidenced by higher sequence coverage of the Chironomidae by the shorter COIS amplicon in the water eDNA samples. Nevertheless, the data additionally show that longer fragments are available at likely lower concentrations in the wild (Deagle *et al.* 2006) (represented by the COIF amplicon) (Supplementary Fig. SF2). Using time vs. DNA fragmentation as a working hypothesis for eDNA degradation, longer fragments are predicted to represent more recently living cellular material. It is also therefore noteworthy that among the water eDNA analyses, only the biodiversity delimited by the COIF amplicon yielded significant associations with time/temperature (Spearman's correlation, $p<0.05$) (Supplementary Fig. SF5), most likely representing more rapid breakdown of longer eDNA fragments in the lake environment. Nevertheless, higher sequence coverage, or methods that preferentially amplify longer amplicons, are needed to enhance amplification probability for potentially smaller concentrations of longer eDNA fragments in natural systems. Such solutions include the combination of multiple primer pairs (Gibson *et al.* 2014), or use of taxon specific/blocking primers. Other suggested strategies for enhancing HTS of eDNA (where concentrations are sufficiently high) involve direct shotgun sequencing or use of capture probes (Taberlet *et al.* 2012b; Liu *et al.* 2016).

Amongst the concerns regarding the utility of eDNA to assess biodiversity, is whether or not species detection represents living or recently living organisms, or communities of "zombie" DNA (i.e. historically distant DNA from organisms that previously lived in the ecosystem a substantial time ago) (Baird & Hajibabaei 2012). If eDNA did have long persistence times in the wild, temporal patterns of β-diversity would be predicted to be extremely low (i.e., non-existent), especially when derived from smaller fragments. However, here we have clearly shown that temporal turnover (β-diversity) was observed for both the animal level (Fig. 2), and total diversity derived eDNA biodiversity analysis (Supplementary Fig. SF6), including temporal patterns of seasonal biodiversity groupings over the year. Similar temporal results were observed for both amplicons, with the short eDNA amplicon providing higher temporal resolution. Some winter samples (Nov 25th and Dec 17th) in the COIS nMDS analysis displayed high levels of β-diversity, since they either contained higher richness (Supplementary Fig. SF5, days 57 and 79) or additional cohorts of taxa not present in the remaining samples (Supplementary Fig. SF4). In the absence of technical artefacts, the additional turnover in β-diversity observed could be the consequence of extreme storm events that coincided with the winter 2013-2014 sampling (Met Office 2016), inputting additional allochthonous eDNA from outside the study area. The time points defining the separation of the two main seasonal biodiversity groups were identified over November and late April, times which also correspond to water temperature below 8 $^{\circ}$C (winter samples) and above 10 $^{\circ}$C (summer samples). Changes in observed community composition (β-diversity) over April and November (Fig. 2, Supplementary Fig. SF6) most likely reflect seasonal turnover, possibly attributed to lake inversion effects (Moss 2010). It is known that changes in water temperature around these times of the year (Spring and Autumn), can trigger the loss of water column stratification by mixing due to changes in surface water temperature(Moss 2010). Collectively, the demonstration of seasonal turnover of lake eDNA β-diversity supports empirical studies using model ecosystems (Moss 2010). Previous laboratory and mesocosm studies have demonstrated the short-term temporal decay of eDNA in artificial environments (e.g. 2-6 weeks) (Thomsen *et al.* 2012b; Strickler *et al.* 2015; Barnes & Turner 2016) and the present data show that the eDNA signal in the wild is of a contemporary nature.

Metabarcoding sequencing of invertebrate communities directly reveals the presence/absence of living, or recently living communities (Taberlet *et al.* 2012b). Hence, the insights provided by community DNA samples here offered an essential benchmark to serve as a proxy for the contemporary invertebrate community. The biodiversity estimates derived from metabarcoding of the community DNA (Fig. 3, Supplementary Fig. SF7, orange lines) matched literature-based estimations of seasonal variation of Chironomidae for Northern Hemisphere temperate latitudes (Armitage *et al.* 2012) (Supplementary Fig. SF9), with a decrease in species richness over winter (often represented by "null" samples due to low numbers of collected exuviae) and a summer increase related to rising water temperature (Fig. 3). Since the emergence patterns of Chironomidae through the year are strongly related to changes in temperature and photoperiod (Armitage *et al.* 2012) (Supplementary Methods SI.1), rapid turnover in emerging communities are apparent and can yield biased estimates of ecological status due to short-term shifts of species emergence (Raunio *et al.* 2010). One of the advantages of metabarcoding over traditional analysis is the ability to analyse many samples simultaneously, and so using molecular approaches for biodiversity assessment presents the opportunity to intensify ecological assessment and derive greater precision in ecosystem health assessment (Thomsen & Willerslev 2015).

The companion analysis of the chironomid eDNA did not follow the expected emergence pattern, despite detecting Chironomidae turnover throughout the year from community DNA samples (Fig. 3). The combination of the β-diversity turnover in *eDNA composition* (Fig. 2), seasonally fluctuating *community DNA richness* (Fig. 3, orange lines) and a lack of coherent seasonal shifts in *eDNA richness* (Fig. 3, blue line) thereby provides an annual model of "community DNA – eDNA" dynamics. The data thereby suggest that there will likely be standing persistent sources of eDNA for biodiversity detection in lake ecosystems that experience annual species turnover (Moss 2010) (Fig. 2). Compositional turnover is thereby expected to result from seasonal variation in species abundances, increasing sources of contemporary eDNA, and environmental degradation decreasing levels of past eDNA accumulation.

Using GAM modelling facilitated comparison between read abundances of individual OTUs derived from eDNA and community DNA analyses. Numbers of read abundances differed

between OTUs over time and between eDNA and community DNA abundances at the individual OTU level (Table 1). There was also a significant positive association between the abundance of sequencing reads derived from the present study and species frequency at the national scale (Fig. 4). Therefore, lower frequency OTUs from the present study occur at lower abundances and higher frequency OTUs are more common, according to an extensive database of Chironomidae occurrence across the UK (Ruse 2013) (Fig. 4).

In combination, the analyses provide an overview of chironomid lake eDNA dynamics. Some species will inevitably yield higher levels of eDNA than others, in relation to life history stage, moulting rates/frequency, abundance, biomass, or cellular content/mitochondrial densities(Rees *et al.* 2014; Thomsen & Willerslev 2015; Barnes & Turner 2016). In addition, the relationship between eDNA and community DNA is affected by biophysical characteristics and interactions between biotic and abiotic factors (e.g. microbial activity, UV radiation and temperature) that affect persistence and degradation rates throughout the year(Barnes *et al.* 2014; Barnes & Turner 2016). Despite such dynamic interactions, numerous broad quantitative associations have been reported for a range of taxa and their eDNA profiles, including data from artificial, semi-natural and natural aquatic ecosystems (Thomsen *et al.* 2012a; Minamoto *et al.* 2012; Pilliod *et al.* 2013; Kelly *et al.* 2014a; Klymus *et al.* 2015; Lacoursière-Roussel *et al.* 2016). Here also, regardless of which methodology was employed, metabarcoding of both eDNA and community DNA reflected general Chironomidae species frequencies across the UK (Ruse 2013) (Fig. 4) and overlapped with biodiversity estimates derived from taxonomy analyses (Fig. 1).

In summary, we have shown that eDNA from water samples collected consecutively over an annual cycle in a lake ecosystem reveals ecologically representative species and community-level shifts in diversity. Importantly, such patterns were validated both by independent assessments of changes in physical presence in a key indicator group of macroinvertebrates, as well as coinciding with established seasonal trends in indicator species emergence and traditional taxonomy. Collectively, the findings address key outstanding questions related to the ecological relevance and temporal persistence of freshwater eDNA in a natural ecosystem, with significant implications for biomonitoring and the future investigation of biodiversity ecosystem functioning relationships.

3.5 Methods

3.5.1 Field sampling

Samples (chironomid pupal exuviae and water samples) were collected during Sept 2013 – Sept 2014 from Llyn Padarn, UK (Supplementary Methods SI.2), an oligotrophic lake ecosystem located in Snowdonia National Park, N. Wales, UK (Supplementary Fig. SF10). The site has been monitored regularly by the UK Environment Agency (EA), and more recently by Natural Resources Wales (NRW) for indicator species of Chironomidae and other invertebrate communities, providing important historical data. Two sites at opposite sides of the lake were selected for sampling: Site 1 (S1) and Site 2 (S2) (Supplementary Fig. SF10). Using two locations increases potential for species detection based on both eDNA and invertebrate sampling. Sampling was conducted at approximately three-week intervals for 1 year (16 time points), using standardised sampling methodology, and collecting simultaneously water and Chironomidae samples. The two sites were sampled always in the same sequence (S1, then S2) between 8:30am–11:30am, including consecutive collection of water samples, invertebrate samples, followed by water metadata (pH, Dissolved Oxygen (D.O.), conductivity and water temperature), using a calibrated YSI Pro Plus multi-meter. As only water and exuviae (shed skins) were collected and the work was performed in collaboration with the EA and NRW, a permit was not required.

3.5.2 Chironomid Exuviae Collection and eDNA filtration

Invertebrate samples in the form of chironomid exuviae (shed pupal skins) were collected using the field collection protocols for the Chironomid Pupal Exuviae Technique (CPET) (Ruse 2010), using a 250µm mesh collection net (Supplementary Methods SI. 1). The floating insect skins were collected on the leeward side (accumulation area) of each sampling site following described methods (Wilson & Ruse 2005) and placed in a sterile container. Upon returning to the lab, the sample was coarsely sorted to remove excessive plant debris, fixed in 100% ethanol and stored at 4°C on the same day of collection, until further processing.

For eDNA samples, one litre of surface water was collected using sterile glass Nalgene bottles from each site, which was transferred on ice and placed at 4°C immediately after return to the laboratory. Filtration was completed within 6 hours in a PCR-free separate room. Sterilised, reusable funnel filtration units (Nalgene filter holders with funnel) were used with 0.45µm cellulose nitrate filter membranes and a high-pressure vacuum pump. The filter membranes were stored in sterile 15ml falcon tubes at - 80°C until DNA extraction. All equipment used was thoroughly sterilised (including Trigene soaks, UV cross-linking and autoclaving) before each sampling event (Supplementary Methods SI. 3).

3.5.3 DNA extractions for eDNA filter membranes and invertebrate samples

Environmental DNA (eDNA) was extracted from the filter membranes, using a modified Phenol Chloroform protocol (PCI), adopted from Renshaw et al (Renshaw *et al.* 2015), with an added digestion step with the addition of 20µl Proteinase K (20mg/µl) (Sigma – Aldrich) and incubation at 60°C for 1 hour. This protocol was selected after rigorous in-house testing of available eDNA capture and extraction protocols (Supplementary Methods SI. 4). In Renshaw et al. (Renshaw *et al.* 2015) it was demonstrated that the latter protocol yielded the highest number of DNA copies of targeted eDNA fragments. Furthermore, the combination of filtration and PCI has been shown to optimise DNA yields, performing equally well in eukaryotes and prokaryotes, with enhanced detection of diversity than other methods (Deiner *et al.* 2015). Two individual extractions were performed for each sample, which were subsequently pooled. Extractions were performed in a different building to PCR library construction where no invertebrate DNA had been handled previously. Extracts were stored in a clean room with no post PCR processing.

DNA extraction from the bulk pupal exuviae samples (community DNA) was performed using a modified QIAmp Blood Maxi Kit protocol, with an added Proteinase K overnight incubation step. Due to seasonal variation of chironomid emergence (Armitage *et al.* 2012), the mass of the collected invertebrate skin material varied, with some of the winter samples containing smaller amounts of tissue. In order to optimise extraction efficiency, 1g of dry invertebrate material was subsampled from large samples. Conversely, for some low-density winter samples, 1g of exuviae was not available and so in these instances, the whole

sample was used for analysis. DNA extraction was performed in standard Qiagen Blood and Tissue kit columns for small winter samples and QIAmp Blood Maxi Kit columns for all other samples with an added 20µl Proteinase K (20mg/µl) overnight incubation step. Both kits are verified by Qiagen to use the same chemistry and differ with respect to the use of columns of different volume capacity to prevent clogging of the membrane. Following separation from the ethanol preservative, the community samples were allowed to air-dry for approximately 1 hour and then were homogenised using a sterile mechanical drill and pestle. For detailed information on each extracted sample, see Supplementary Tables ST4 & ST5.

3.5.4 Primer selection and MiSeq Library preparation

To fulfil the overarching aims of the study, we required (a.) metabarcoding primers that would amplify across a broad range of taxa (in particular, lake occurring taxa), (b.) a marker enabling the best annotation power for macroinvertebrates and in particular, the Chironomidae, (c.) a combination of two primer pairs providing different length amplicons.

Accordingly, two amplicons of different sizes of the mitochondrial Cytochrome Oxidase I gene (COI) were selected for sequencing. The full-length COI barcoding region (658bp), using the universal Folmer primers LCO1490 - HCO2198 (Folmer *et al.* 1994) (amplicon COIF) and a 235bp fragment (amplicon COIS) using the forward primer LCO1490 and the reverse COIA-R primer (reversed forward COI-A primer by Carew *et al.* 2013). The forward COI-A primer was designed by (Carew *et al.* 2013) specifically for amplification of Chironomidae from environmental samples. Two Illumina MiSeq dual indexed amplicon libraries were prepared using a two-step PCR protocol (Miya *et al.* 2015). The first round amplification was performed using template-specific primers with 5' Illumina tails (TruGrade, by IDT, Integrated DNA Technologies (Coralville, USA)), followed by Agencourt AMPure magnetic bead purification. A second round amplification was performed using Illumina adapters with 8-nucleotide Nextera indexes (see Supplementary Table ST6). A 5N sequence was implemented between the forward universal tail and the template specific primer, which is known to improve clustering and cluster detection on MiSeq sequencing platforms (Miya *et al.* 2015). Using primers with identical tails in the first step and indexed primers in the

second, is a protocol specifically developed by Illumina to reduce bias caused by variable index sequences in mixed environmental samples (Berry *et al.* 2011; O'Donnell *et al.* 2016).

Each sample was amplified in triplicate, the final products were pooled and purified with AMPure beads and quantified using a dsQubit assay. Final library pooling was performed in equimolar quantities for all samples. Sequencing was performed at the Liverpool Centre for Genome Research, distributed across two independent lanes (for the COIS and COIF amplicons) of Paired-end Illumina MiSeq (2x300) sequencing (detailed PCR amplification protocols are provided in Supplementary Methods SI.5).

3.5.5 Sequencing quality control

To control for quality of eDNA capture methods, negative controls (blanks) were collected during water filtration, which were sequenced on the MiSeq along with reagent and filter blank extractions (for details on collection of blank samples see Supplementary Methods SI.3). To account for efficiency of amplification protocols and sequencing, a composite positive control sample comprising 30 invertebrate DNA extracts was also amplified in triplicate with both primer pairs, and sequenced alongside eDNA and community samples on MiSeq (for details on preparation of positive control samples see Supplementary Methods SI. 3).

3.5.6 Bioinformatics and statistical analysis

Sequences, including positive and negative controls, were de-multiplexed and Illumina adapters trimmed using Cutadapt (Martin 2011) and Sickle (Joshi & Fass 2011). A 10% level of mismatch (2 bases) was allowed for primer removal. Filtering and quality control were then performed using USEARCH v7 (Edgar 2010). Sequence quality was visualised using FastQC (www.bioinformatics.babraham.ac.uk) and only sequences with a Phred quality score >25 were retained for analysis. Using USEARCH (fastq_maxee = 1) sequences with a maximum expected error (maxee) > 1 were discarded. Maxee is the expected number of errors as sum of the error probabilities (provided by Phred scores). Filtering was performed after merging of R1 and R2 reads (minimum overlap 25bp), which allows recalculation of the error probabilities for the combined sequences and increased accuracy. Sequences

shorter than 100bp were discarded. The remaining sequences were de-replicated and sorted by cluster size (cluster abundance) and sequences with <2 clusters (singletons) were removed. For the COIF amplicon, the whole barcoding region was amplified and sequenced, but because of the current limitations of MiSeq sequencing read lengths, only the forward reads (R1) were used for analysis. Consequently, the per base quality drop expected in Illumina MiSeq data at the tail of the forward reads was inspected in FastQC and all reads were truncated at 250bp and then quality filtered as above. Next, chimeras were removed (uchime_denovo) using a *de novo* delimitation approach. An operational taxonomic unit (OTU) table was created using OTU clustering at 97% similarity (USEARCH). Clustering at 97% similarity level was chosen based on existing knowledge of intraspecific diversity for Chironomidae (Carew *et al.* 2013), since previous studies suggest that chironomid intraspecific diversity ranges between 0-4.2% (Carew *et al.* 2013) or 0-4.9% (Ekrem *et al.* 2007).

Taxonomy was then assigned to the OTU table using BLAST+ (megablast) (Camacho *et al.* 2009) against a reference COI database. The reference library was compiled from NCBI GenBank, by downloading all COI sequences, >100bp, excluding environmental sequences (20th June 2015, N = 807,388 sequences) and higher taxonomic level information was edited using the GALAXY online software platform (Goecks *et al.* 2010). Taxonomic assignment of the OTU tables and subsequent analysis was performed in QIIME (Caporaso *et al.* 2010). All analyses involving USEARCH, QIIME and BLAST+ were performed using the High Performance Computing (HPC) Wales systems.

Given the potentially sensitive nature of eDNA metabarcoding, low frequency sequences can either represent less abundant taxa, or possible false positives and low level contaminant OTUs (Murray *et al.* 2015). In order to reduce the error associated with low frequency sequences, and also focus analyses on predicted levels of richness (Fonseca *et al.* 2010), we used two types of analysis. First, we identified the frequency of potential contaminant reads in the positive control. Second, we compared chironomid eDNA richness with variable levels of relative abundance filtering (no filtering, 0.01% and 0.02%), against historical records of richness (genus level only available) for Llyn Padarn (based on summer surveys for Llyn Padarn, 2003 – 2013). Consequently, abundance filtering was performed

on the OTU tables at the level that most closely emulated expected chironomid richness and within the limits associated with empirically observed low-level contamination in the sequencing dataset.

The validity of the Chironomidae OTUs identified by BLAST and retained after abundance filtering was checked using a phylogenetic approach. The BLAST identified Chironomidae OTUs were aligned with barcodes from 24 Chironomidae and 40 Trichoptera species obtained herein, sequenced from UK samples using universal primers (Folmer *et al.* 1994). Alignment, testing for the presence of stop codon and insertions and bootstrapped phylogenetic tree construction were performed in MEGA (Tamura *et al.* 2007). Ultimately, only the OTUs that grouped closely with known chironomid sequences on the phylogenetic tree were included in further analysis.

For downstream analyses, the appropriate depth of coverage per sample was determined according to OTU accumulation vs. sequence coverage curves generated in QIIME. Samples were subsequently normalised using rarefaction in QIIME at appropriate depth for each amplicon (Magurran & McGill 2011).

3.5.7 Taxonomic identification of invertebrate community samples

To provide a comparison with community DNA and eDNA sequenced samples, chironomid exuviae community samples from 4 time points (T10: April 30, T11: May 20, T14: July 23, T16: September 04) were taxonomically identified according to standard CPET methodology used by the EA. More specifically, 200 chironomid exuviae were subsampled from the total community sample and identified to the highest possible level (genus or species) by specialised EA staff. The results of the taxonomic identification were used to compare chironomid richness at the genus level with metabarcoding-generated richness (see below).

3.5.8 Calculation of diversity measures

OTU richness (total diversity and Chironomidae diversity) was calculated in QIIME. Furthermore, for Chironomidae with good taxonomic identification, richness was also calculated at the genus level. To assess variation of richness over time polynomial regression was performed using R version 3.2.4 (2016).

The PRIMER-E software (Clarke & Gorley 2006) was used to calculate β-diversity based on the Sørensen index for total diversity and Animalia only diversity detected from aqueous eDNA samples and for Chironomidae OTUs for both sample types. Non-metric multi-dimensional scaling (nMDS) and Hierarchical Clustering (HC) analysis were used to represent community similarity between samples. Analysis of similarity (ANOSIM) was used to test for significant effects of time in relation to community composition.

3.5.9 Chironomidae OTU read abundance (eDNA vs community DNA)

In order to explore relationships between the numbers of metabarcoding sequence reads, individual OTUs and methodology (eDNA vs. community DNA), we used a generalized additive model (GAM), with time as a smoothing term, using the R-package mgcv (Wood 2011). In the GAM model, abundance, calculated as total normalised reads per OTU and standardized per method (to allow for across method comparison), was assessed in relation to OTU identity and method (eDNA vs community DNA). Additionally, we assessed the ecological relationship between OTU abundance (log transformed) in Llyn Padarn and species frequency (i.e. abundances derived from ecological assessment) across the UK, by performing a two-way ANOVA, using the lm function in R. UK species frequencies were derived from a Chironomidae inventory of 435 species across 220 UK lakes (Ruse 2013). We restricted the species frequency data to 97 sites where species frequency was inventoried at the national level and observed in this study.

3.6 Acknowledgements

This work was funded by the Environment Agency (EA) UK, a Knowledge Economy Skills Scholarship (KESS), a Natural Environment Research Council (NERC) NBAF pilot project grant (NBAF824 2013-14) and the Freshwater Biological Association (FBA) (Gilson Le Cren Memorial Award 2014). We thank the EA and Bangor University for support and in particular, Wendy Grail, John Evans, Emlyn Roberts and EA staff for facilitating provision of eDNA grade laboratory working spaces, equipment, and taxonomic identification of chironomid specimens; HPC Wales for allowing use of their systems; Les Ruse and APEM for identification of Chironomidae specimens for Barcoding; Natural Resources Wales for

providing historical data. We also acknowledge the support of NERC Highlight Topic grant NE/N006216/1. Knowledge Economy Skills Scholarships (KESS) is a pan-Wales higher-level skills initiative led by Bangor University on behalf of the HE sector in Wales. It is part funded by the Welsh Government's European Social Fund (ESF) convergence programme for West Wales and the Valleys.

3.7 Author contributions

IB, SC, GRC: Designed experiment. IB: Performed lab work, fieldwork, bioinformatics and parts of statistical analysis. MS: Performed statistical analysis and data modelling, DL: Contributed in optimisation of analytical pipelines. MH, MC, KW: Participated in experimental design. IB, SC, GRC: Wrote manuscript. IB, SC, GRC, MS, KW, DL, and MH: Edited manuscript.

3.8 Additional information

Competing financial interests. The authors declare no competing financial interests.

Data deposition. Sequencing data reported here have been deposited in GenBank (Submission IDs: 1966226, 1966195) and the European Nucleotide Archive (ENA) (Accession number: PRJEB13009).

3.9 Supplementary Information

SI.1 Emergence patterns of Chironomidae and the Chironomid Pupal Exuviae Technique (CPET).

Chironomids exhibit specialised responses to ecological stressors and are acknowledged as one of the most important macroinvertebrate groups for monitoring lake ecosystem health (Wilson & Ruse 2005). However, benthic larvae collected with traditional kick-net sampling are notoriously difficult to identify, even by specialists. To overcome these problems lentic Chironomidae biodiversity is assessed via the identification of shed exuviae (skins) of emerging adults that float and accumulate on the leeward edge of lentic ecosystems (Wilson & Ruse 2005; Ruse 2011). Exuvial samples therefore offer a unique advantage to simultaneously compare the diversity of recent lentic invertebrate communities and eDNA and to explore how eDNA is related to ecosystem wide biodiversity. Additionally, using the CPET technique compared to traditional kick-net sampling, allows for integrated collection of specimens from a wide range of habitats rather than only the profundal zone. The collection and sorting process is fast and the identification of the exuviae is easier than identification of larvae, while the sample collected is also fresh, as the exuviae remain floating for only about 48h (Wilson & Ruse 2005).

The emergence patterns of Chironomidae are known to differ in different latitudinal zones, due to variations in temperature and photoperiod (Armitage *et al.* 2012). In the tropics, the emergence cycles are accelerated, following the lunar cycles, with species emerging all year round. On the contrary, closer to the Arctic, emergence of adults occurs over a limited window over the summer period. Emergence is limited also by surface freezing of the water bodies. For the temperate zones, emergence is higher over the summer but not limited to that time. Species are known to emerge across all seasons, but with less intensity in winter months. Hence an episodic pattern occurs, with lower emergence over winter, which increases gradually over time.

SI.2 Sampling sites on Llyn Padarn, N. Wales (UK)

Llyn Padarn is located in Snowdonia Nature Reserve (53.130051, -4.135567), adjacent to Llanberis. Approximate surface area is 97.6 ha with maximum depth 27m. The two sites used for sample collection are shown on the map: Site 1 (S1) (NW: 53.139106, -4.153975) and Site 2 (S2) (SW: 53.122414, -4.126761) (Supplementary Fig. SF9). In the past, the lake has been monitored by the Environment Agency (EA) and more recently by its successor Natural Resources Wales (NRW).

SI.3 Equipment Sterilization and control samples

All equipment was thoroughly sterilized between sampling visits. The glass Nalgene bottles used for water collection, filtration units and forceps would undergo consecutive cleaning rounds including wash and overnight soak with 10% Trigene (Ammonium chloride & hydrochloride, Medichem Int.), thorough rinse, UV treatment for 5 min and autoclaving. All additional equipment used for invertebrate collection (net, meters, boots) was also thoroughly washed with 10% Trigene. For eDNA extractions, single-use pre-sterilised scissors and forceps were used to handle the filter membranes, and the exterior of storage tubes was wiped with 10% Trigene before handling. During field surveys, to minimise cross contamination from consecutive sampling points, the water samples were collected first, before any other samples or measurements were taken and prior to invertebrate collection.

Negative controls were collected by filtration of distilled water. The negative control equipment would undergo the same cleaning steps (Nalgene bottles filled with distilled water) along with all other equipment. A litre of distilled water was filtered through the filtration funnels (prior to sample filtration), and the filter membranes were collected and stored same as the rest of the samples. Further to distilled water negative controls, blank extractions of reagents (reagent controls) and filters (filter controls) were extracted with the same Phenol Chloroform extraction protocol (PCI) (Renshaw *et al.* 2015). All negative controls were amplified with both primer pairs and MiSeq library preparation steps (see Methods), and sequenced on Illumina MiSeq.

Positive controls were used to account for efficiency of amplification protocols and sequencing. A composite sample was prepared using DNA extracts from 30 invertebrate samples including Amphipoda, Coleoptera, Diptera, Ephemeroptera, Gastropoda, Hemiptera, Isopoda and Trichoptera (**Table S1**). The sample contained 11 Chironomidae extracts (Diptera). This sample was amplified with the matching protocols for COIS and COIF accordingly and sequenced on Illumina MiSeq.

SI.4 Testing of capture and extraction protocols for eDNA

Rigorous testing of eDNA capture and extraction protocols was performed prior to commencing the experiment. For testing of filtration methods, two types of filtration membranes at different pore sizes were used: glass fibre at 0.7µm and cellulose nitrate at 0.45µm and 0.2µm. Two volumes of water samples were used at 1L and 2L. Ethanol precipitation and centrifugation, using 15ml water samples was also tested, as well as direct centrifugation of 50ml water samples (no precipitation or filtration). For the latter two, varying centrifugation speeds and centrifugation times were also tested. The extraction protocols included the DNeasy Blood & Tissue kit (QIAGEN), Power Water DNA Isolation kit (MoBio) and Phenol Chloroform extraction protocol (PCI) as per (Renshaw *et al.* 2015) with an added Proteinase K step.

From all the above, the collection of eDNA using 0.45µm cellulose filter membranes (2lt water) coupled with a PCI extraction protocol was considered optimal, due to the following: 1) Higher concentrations of collected DNA as per spectrophotometric quantification (NanoDrop) and quality of DNA from agarose gel visualization. 2) Possibility for collection of larger water sample (2L). 3) Ease of storage of collected samples (filter membrane) until DNA extraction (storage at -80^{o}C). 4) Optimal pore size for collection of smaller DNA molecules (compared to glass fibre 0.7µm) and filtration time efficiency (compared to cellulose 0.2µm). 5) Good performance in PCR amplification of long COI amplicons.

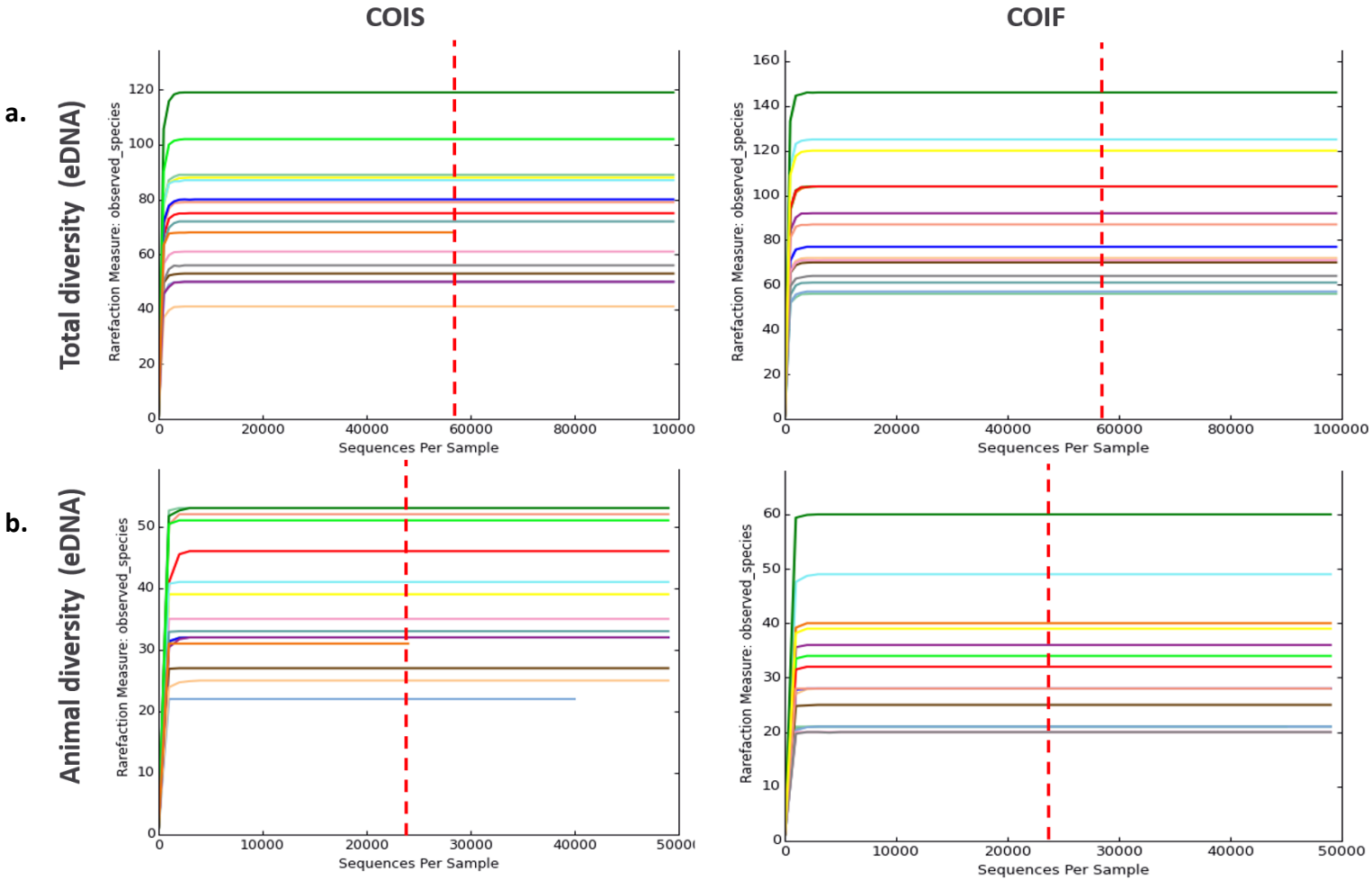
SI.5 PCR protocols for MiSeq Library Preparation

PCRs were performed in 25µl reaction volumes containing, for **Round 1**: 12.5µl Q5® Hot Start High-Fidelity 2X Master Mix, 10.5µl PCR water, 0.5µl (10nmole/µl) of each forward and reverse primer and 1µl DNA (10ng/µl). For **Round 2**: 12.5µl Q5® Hot Start High-Fidelity 2X Master Mix, 6.5µl PCR water, 0.5µl of each forward and reverse primer and 5µl Purified PCR product from Round 1. The following thermo-cycling parameters were used: **Round 1**: **COIF**: Denaturation at 98°C for 30 sec, 20 cycles of: 98°C for 10 sec, 46°C for 30 sec, 72°C for 40 sec, followed by a 10min extension at 72°C, hold at 4°C. **COIS:** Denaturation at 98°C for 30 sec, 20 cycles of: 98°C for 10 sec, 45°C for 30 sec, 72°C 30 sec, followed by a 10min extension at 72°C, hold at 4°C. **Round 2: both amplicons**: Denaturation at 98°C for 30 sec, 15 cycles of: 98°C for 10 sec, 55°C for 30 sec, 72°C for 30 sec, followed by a 10min extension at 72°C, cool at 4°C for 10min. Round 1 PCRs were performed using Illumina-tailed primers and Round 2 using Illumina indexes.

SI.6 Positive and negative control results

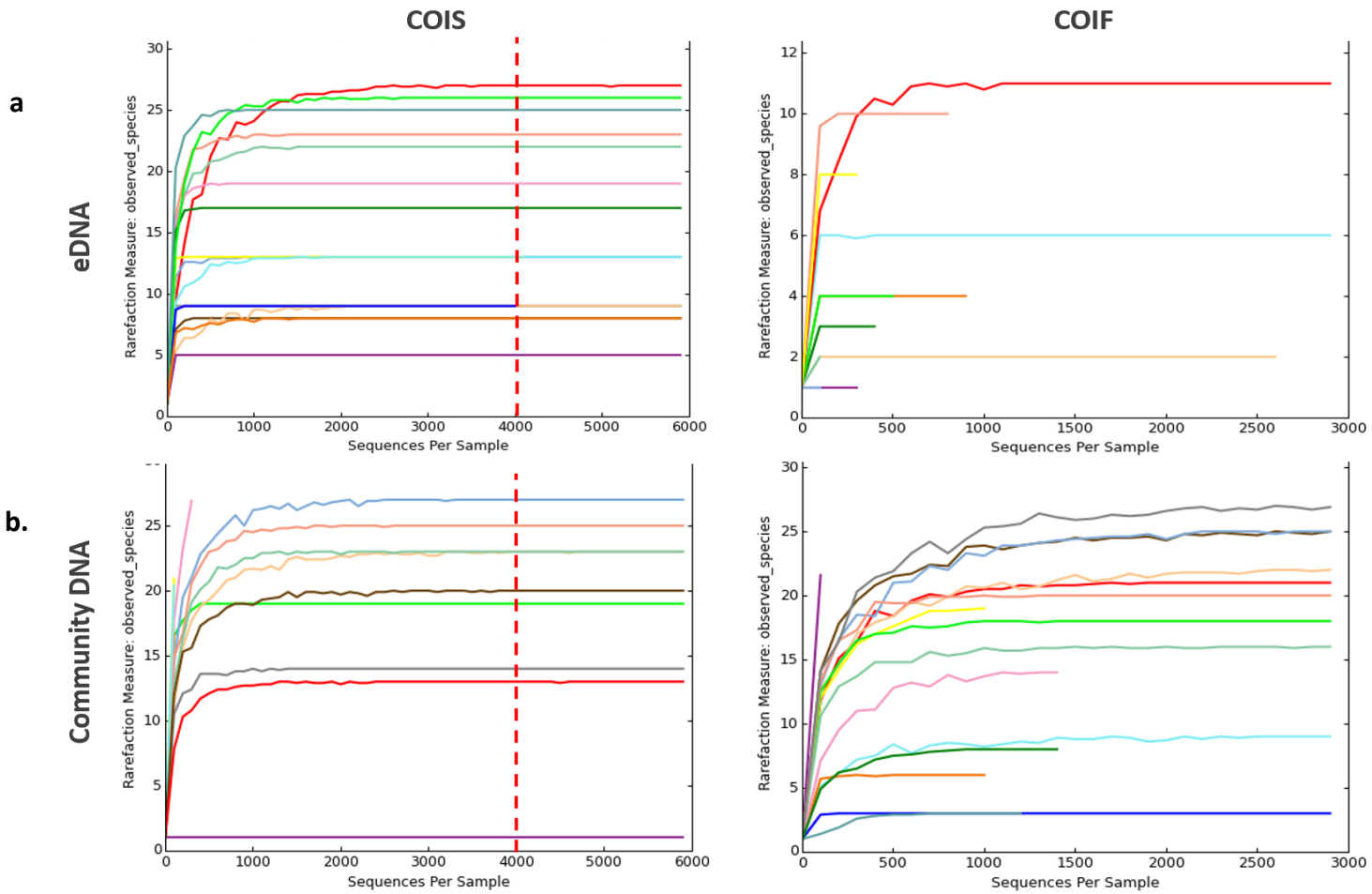
Negative Controls. After PCR and sequencing of the negative control samples, COIS detected only two OTUs, which were BLAST-identified as bacteria. For COIF, again only two OTUs were detected, identified as Gastropoda and Diptera. The Gastropoda OTU presented up to 240 reads in one of the controls while the Dipteran OTU only presented 10 reads in total across all types of negative controls.

Positive controls. Sequencing of the positive control samples resulted in 100% detection success for COIS, which detected all 30 taxa present. The COIF amplicon failed to detect four taxa (87% success rate). Amongst the species that were not detected was a mayfly species (*E. danica*) which also failed to amplify and sequence during individual barcoding of specimens, using the same primer pair.



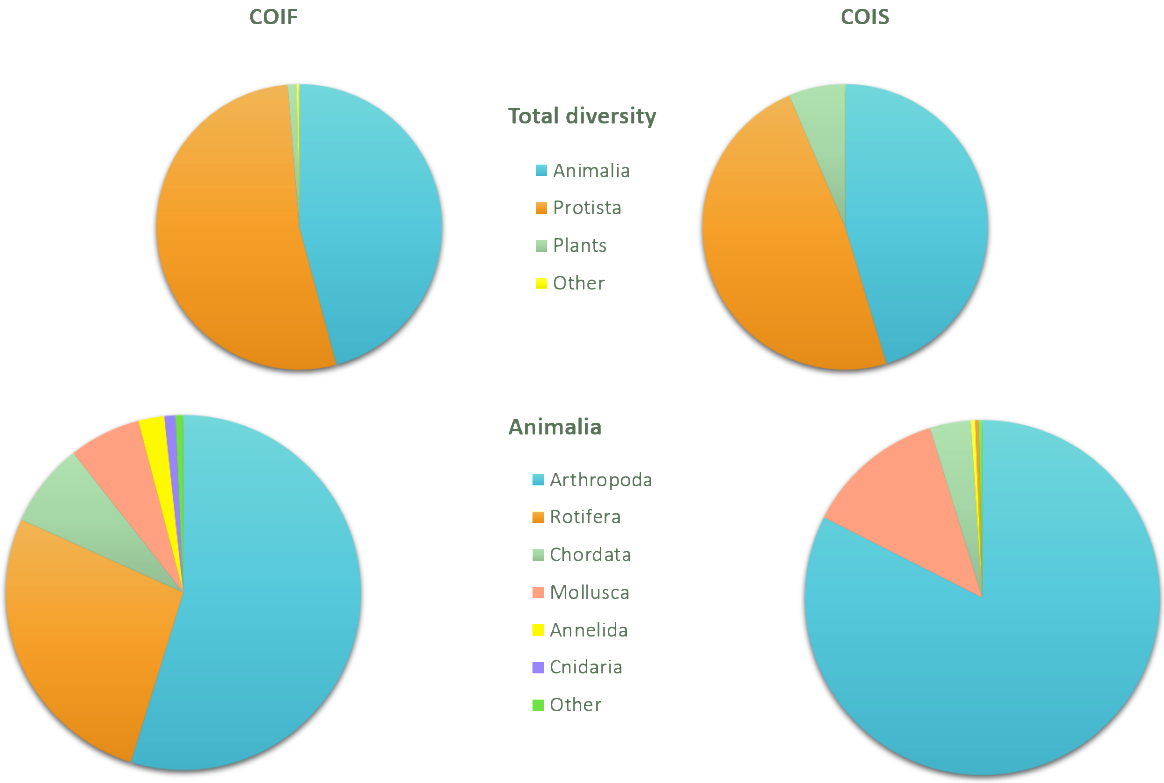
Supplementary Figure 1: Rarefaction plots (Total diversity).

The figure shows (a) total taxa and (b) animal taxa only, based on water extracted eDNA samples only for both amplicons (COIS and COIF). Dashed red lines indicate the rarefaction depth used for analysis (a. total taxa 57,869 reads, b. animal taxa 24,914 reads), x-axis: reads per sample, y-axis: OTU richness (N=64).

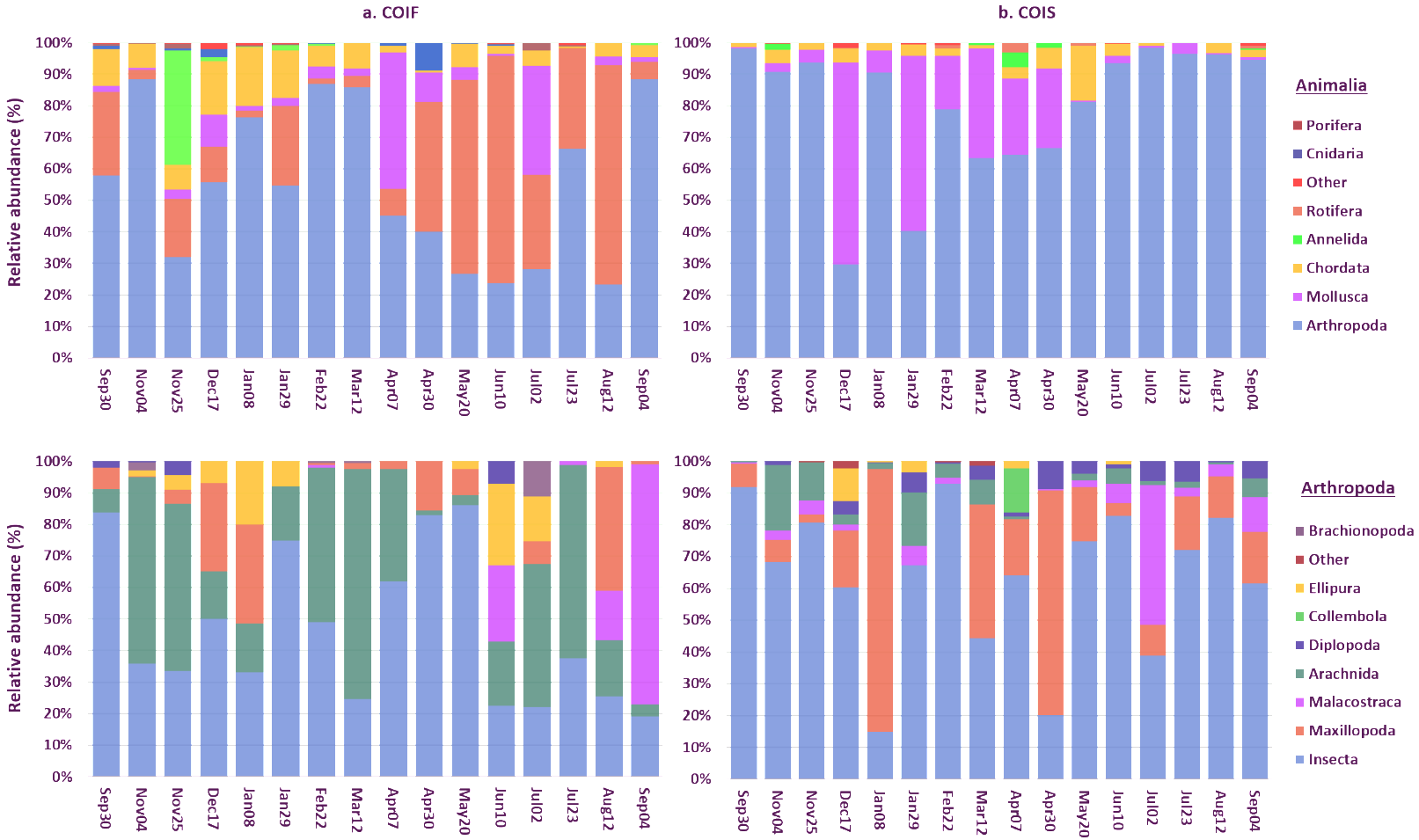


Supplementary Figure SF 2: Rarefaction plots (Chironomidae).

The figure shows Chironomidae identified OTUs, (a) eDNA samples and (b) community DNA samples, for both amplicons (COIS and COIF). Dashed red lines indicate the rarefaction depth used for analysis (COIS: 4,000 reads). Due to low coverage of COIF eDNA samples (a-top), this amplicon was excluded from further analysis. x-axis: reads per sample, y-axis: OTU richness (N=64).

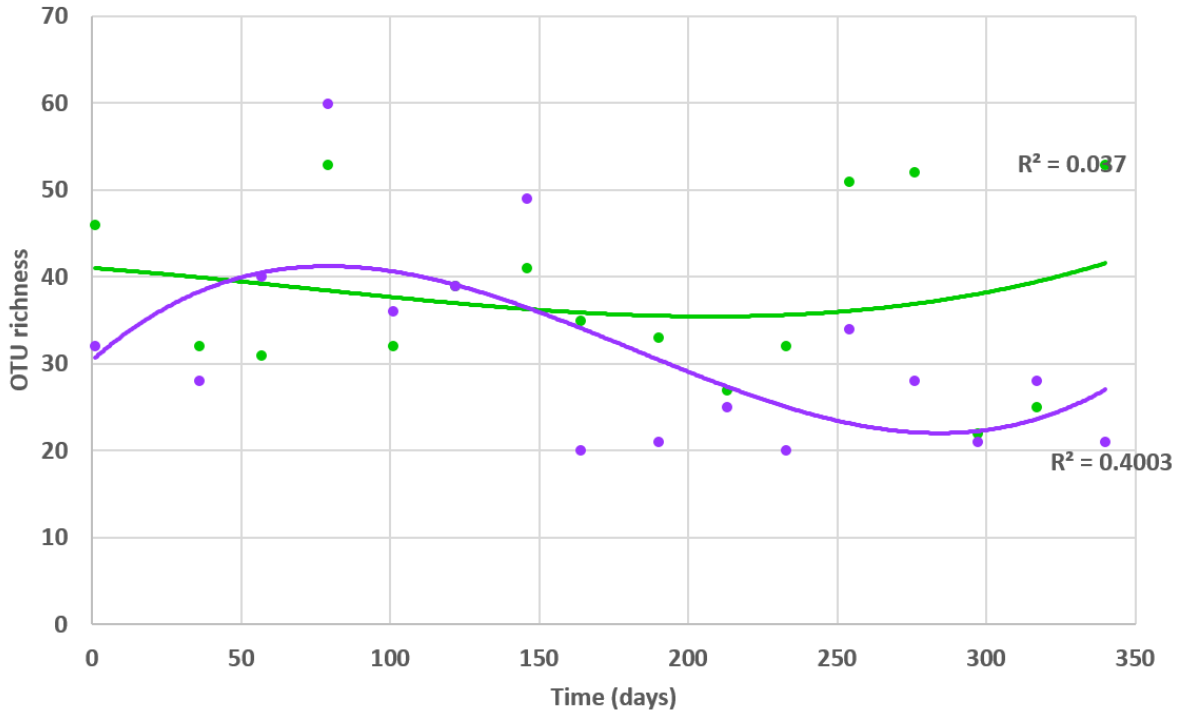


Supplementary Figure SF 3: Summary representation of taxa detected.
Results shown for eDNA samples for both amplicons (COIF, COIS). Top: Kingdoms, bottom: phylum Animalia.

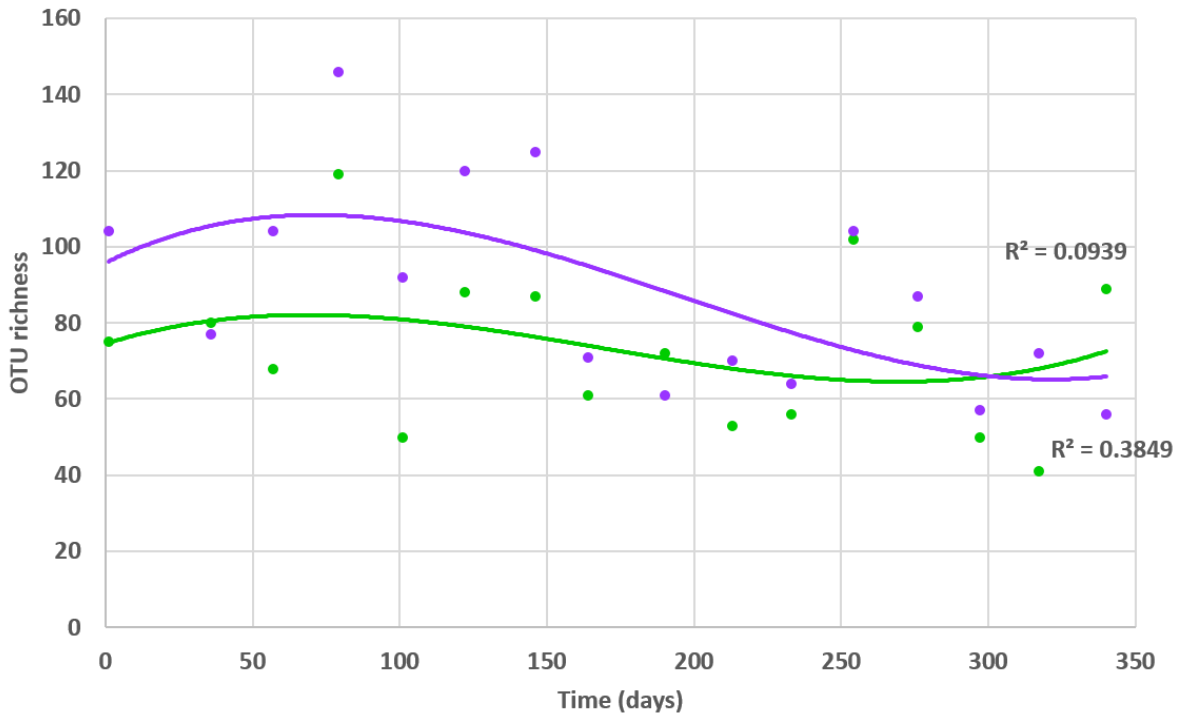


Supplementary Figure SF 4: Histogram presenting taxonomic relative abundance for both amplicons.
a.) COIF, b.) COIS, for all animal (top) and all arthropod (bottom) taxa in eDNA samples through the year (x-axis: sampling dates). All samples were rarefied at 24,914 read depth.

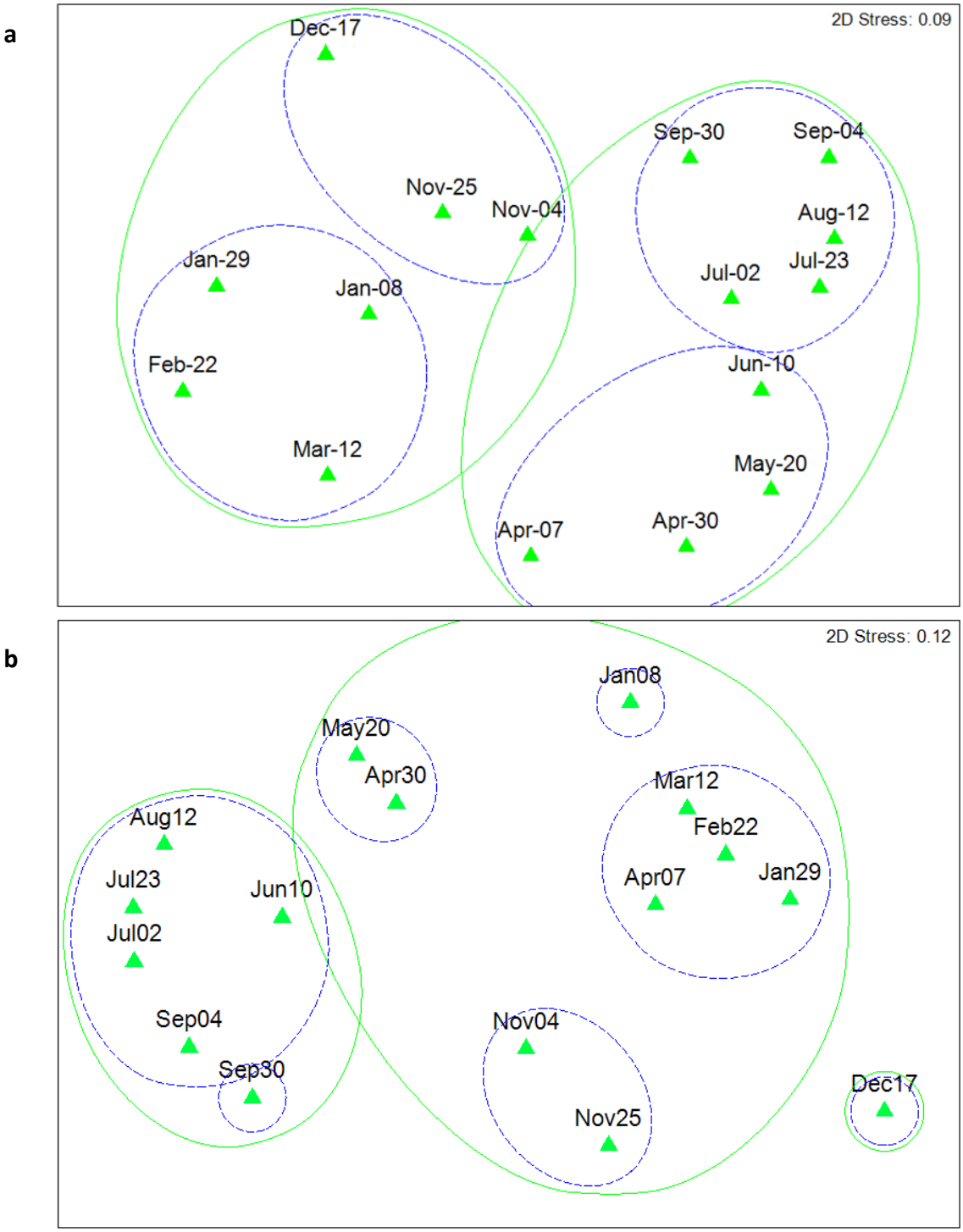
a.



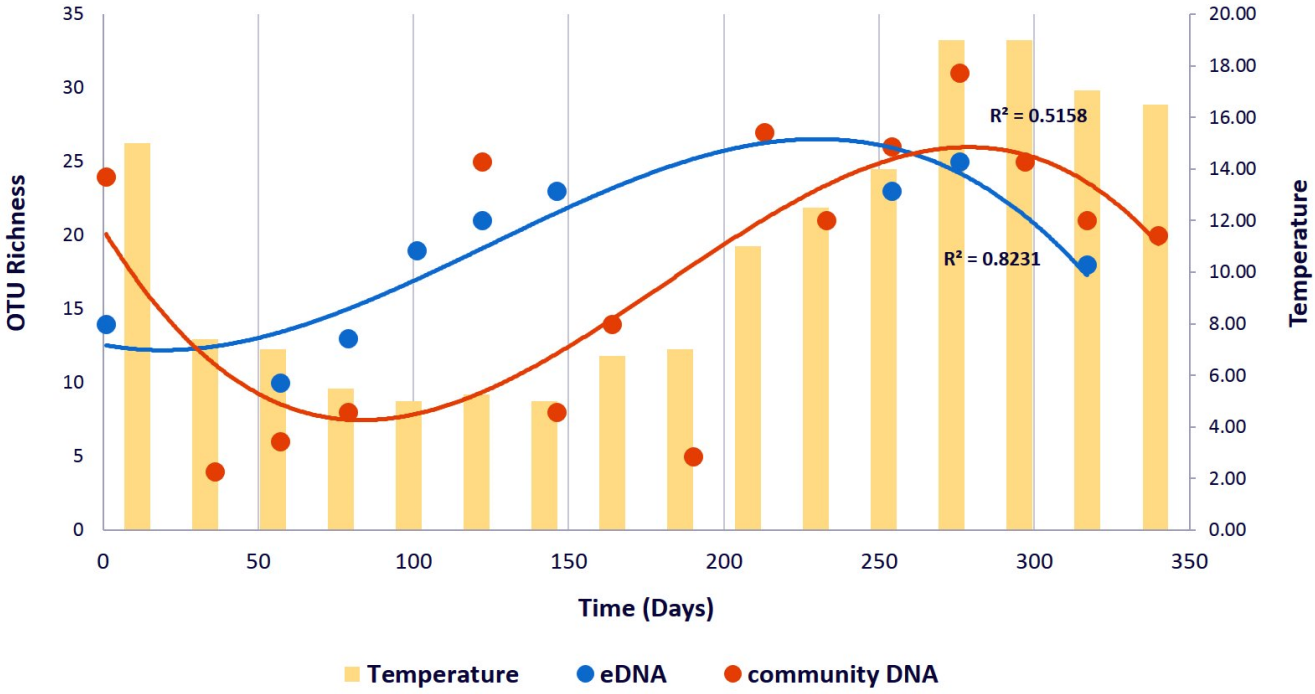
b.



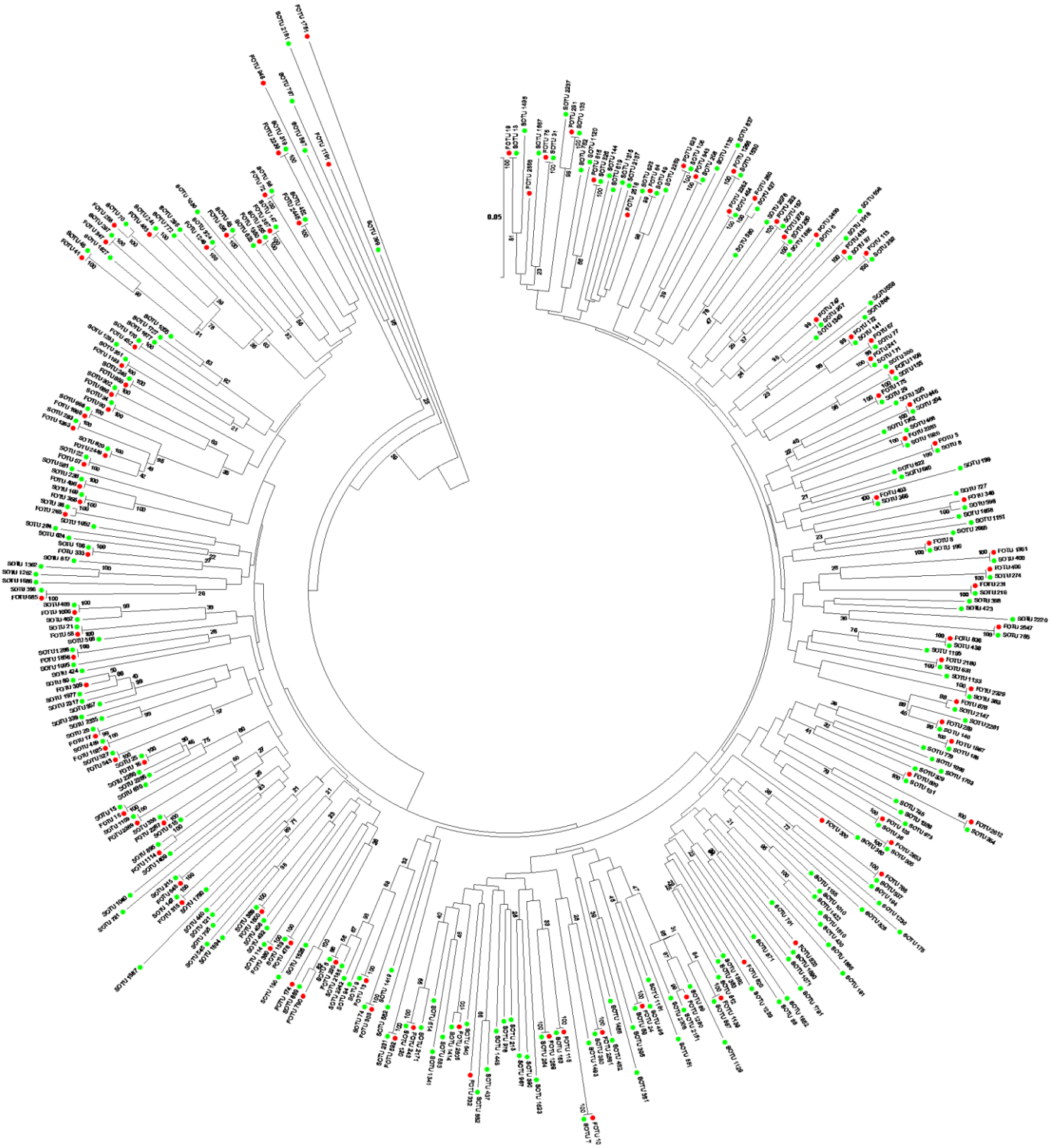
Supplementary Figure SF 5: Yearly trends of OTU richness.
a.) animal diversity b.) total diversity, detected by eDNA samples for both COIS (green) and COIF (purple). X-axis: time in days (Sep 30th 2014- Sep 4 2015), y-axis: OTU richness.



Supplementary Figure SF 6: nMDS plots of β-diversity (Sørensen index).
For eDNA samples only. a.) COIF, b.) COIS (N = 32). Solid green circles: 30% similarity cut-off (corresponding to "winter" –"summer" groups), dashed blue circles: 40% similarity cut-off (N=32).



Supplementary Figure SF 7: OTU richness patterns for Chironomidae OTUs for the COIF amplicon (raw data un-trimmed). Points represent richness values to individual sampling points for eDNA (blue) and community DNA (orange). Best fitted lines from polynomial regressions for eDNA samples (blue) and community DNA (orange), plotted against time (x – axis: Sep. 2013 – Sep 2014).



Supplementary Figure SF 8: Neighbour-Joining phylogenetic tree.
The tree comprises all OTUs identified as Chironomidae prior to abundance filtering, for both amplicons (COIF: red markers (FOTU), COIS: green markers (SOTU)). Distances calculated using the p-distance method 1000 bootstrap replications (N = 351).

Supplementary Figure SF 9: Map of Llyn Padarn, N. Wales (UK).
Marked with red the two sites used for sample collection (S1: Site 1, NW: 53.139106, -4.153975, S2: Site 2, SW: 53.122414, -4.126761). Google Earth, August 2016.

Supplementary Table ST 1: Summary table of number of reads obtained per sample.
Triplicate PCRs from each time point were pooled and sequenced as one. (EXCOI: exuviae community DNA samples, WCOI: water eDNA samples, COIS; 235bp amplicon, COIF: 658bp amplicon).

#Sample	Sample name	Number of reads		Sample Type	Collection date	Time point
		COIS	COIF			
1	1_EXCOI	159874	383161	Pupal exuviae	30/09/2013	T1
2	2_EXCOI	464	3603	Pupal exuviae	04/11/2013	T2
3	3_EXCOI	442	3808	Pupal exuviae	25/11/2013	T3
4	4_EXCOI	349	2844	Pupal exuviae	17/12/2013	T4
5	5_EXCOI	203602	507	Pupal exuviae	08/01/2014	T5
6	6_EXCOI	387	2406	Pupal exuviae	29/01/2014	T6
7	7_EXCOI	262	365700	Pupal exuviae	22/02/2014	T7
8	8_EXCOI	411	1825	Pupal exuviae	12/03/2014	T8
9	9_EXCOI	475	2755	Pupal exuviae	07/04/2014	T9
10	10_EXCOI	165644	468915	Pupal exuviae	30/04/2014	T10
11	11_EXCOI	139771	363563	Pupal exuviae	20/05/2014	T11
12	12_EXCOI	289842	336948	Pupal exuviae	10/06/2014	T12
13	13_EXCOI	168006	347443	Pupal exuviae	02/07/2014	T13
14	14_EXCOI	343465	15231	Pupal exuviae	23/07/2014	T14
15	15_EXCOI	489950	25608	Pupal exuviae	12/08/2014	T15
16	16_EXCOI	500181	18963	Pupal exuviae	04/09/2014	T16
17	1_WCOI	240086	273799	Water	30/09/2013	T1
18	2_WCOI	189255	260032	Water	04/11/2013	T2
19	3_WCOI	62109	253590	Water	25/11/2013	T3
20	4_WCOI	288282	302474	Water	17/12/2013	T4
21	5_WCOI	261100	346620	Water	08/01/2014	T5
22	6_WCOI	272002	253954	Water	29/01/2014	T6
23	7_WCOI	157903	280711	Water	22/02/2014	T7
24	8_WCOI	253438	263482	Water	12/03/2014	T8
25	9_WCOI	314163	245330	Water	07/04/2014	T9
26	10_WCOI	282801	253024	Water	30/04/2014	T10
27	11_WCOI	224307	154471	Water	20/05/2014	T11
28	12_WCOI	281971	430025	Water	10/06/2014	T12
29	13_WCOI	252773	249347	Water	02/07/2014	T13
30	14_WCOI	276285	285992	Water	23/07/2014	T14
31	15_WCOI	311891	203605	Water	12/08/2014	T15
32	16_WCOI	309147	259862	Water	04/09/2014	T16

Supplementary Table ST 2: Positive control contents.
Extracts used for preparation of a positive control sample and taxonomic information of the specimens used for extraction (species level information was not available for some of the specimens). The last two columns show the success of the amplicons in detecting each extract (v: detected, x: not detected).

Positive Control Contents					**Amplicon**	
Number	**Extract Code**	**Order**	**Family**	**Species**	**COIF**	**COIS**
1	C_pseudQ24_1	Amphipoda	Crangonyctidae	*Crangonyx pseudogracilis*	v	v
2	G_pulexQ29_2	Amphipoda	Gammaridae	*Gammarus pulex*	v	v
3	G_marinusQ33_2	Coleoptera	Gyrinidae	*Gyrinus marinus*	v	v
4	PA6	Diptera	Chironomidae	*Chironomidae sp.*	v	v
5	PA8	Diptera	Chironomidae	*Chironomidae sp.*	**x**	v
6	PA16	Diptera	Chironomidae	*Chironomidae sp.*	v	v
7	PA17	Diptera	Chironomidae	*Chironomidae sp.*	v	v
8	SERGPSI6_2	Diptera	Chironomidae	*Sergentia psiloptera*	v	v
9	ABLAMON2	Diptera	Chironomidae	*Ablabesmyia monilis*	v	v
10	CHIRTEN13_1	Diptera	Chironomidae	*Chironomus tentans*	v	v
11	CRYPPSI13_1	Diptera	Chironomidae	*Cryptochironomus psittacinus*	v	v
12	MONOBAT6A	Diptera	Chironomidae	*Monodiamesa bathyphila*	v	v
13	CLATATR10A	Diptera	Chironomidae	*Cladotanytarsus atridorsum*	v	v
14	POLYNUC7B	Diptera	Chironomidae	*Polypedilum nubeculosum*	v	v
15	E_danicaE130	Ephemeroptera	Ephemeridae	*Ephemera danica*	**x**	v
16	COR1_G18_1	Gastropoda	Lymnaeidae	*Radix sp.*	v	v
17	DEV3_G27_1	Gastropoda	Lymnaeidae	*Radix balthica*	v	v
18	ANG5_G2_1	Gastropoda	Planorbidae	*Ancylus fluviatilis*	v	v
19	A_vortexQ2_2	Gastropoda	Planorbidae	*Anisus vortex*	v	v
20	N_glaucaN10	Hemiptera	Notonectidae	*Notonecta glauca*	v	v
21	A_aquaticus	Isopoda	Asellidae	*Asellus aquaticus*	**x**	v
22	SCO12_T2_1	Trichoptera	Glossosomatidae	*Agapetus fuscipes*	**x**	v
23	COR2_T79_1	Trichoptera	Goeridae	*Silo pallipes*	v	v
24	WALE13_T33_1	Trichoptera	Hydropsychidae	*Hydropsyche instabilis*	v	v
25	HE1_T4_1	Trichoptera	Hydroptilidae	*Agraylea sexmaculata*	v	v
26	YO2_T3_1	Trichoptera	Hydroptilidae	*Agraylea multipunctata*	v	v
27	HEA_T37_1	Trichoptera	Hydroptilidae	*Hydroptila vectis*	v	v
28	ANG5_T43_1	Trichoptera	Lepidostomatidae	*Lepidostoma hirtum*	v	v
29	SCOT2_T27_1	Trichoptera	Limnephilidae	*Halesus radiatus*	v	v
30	ANG5_T77_1	Trichoptera	Leptoceridae	*Athripsodes albifrons*	v	v

Supplementary Table ST 3: Positive control sequencing results.
Summary table of sequencing results obtained from positive control samples for 235bp COIS and 658bp COIF amplicon. Shown the number of reads, number of OTUs and relative abundance assigned to our target species (target), unidentified OTUs (unknown) and identified OTUs not present in our target species (Non – target).

Positive controls	COIS			COIF		
	reads	%	OTUs	reads	%	OTUs
Target	547569	99.971	33	393068	99.931	29
Unknown	18	0.003	3	246	0.063	16
Non - Target	143	0.026	14	27	0.007	6
Total	547730	100.000	50	393341	100	51

Supplementary Table ST 4: Summary of eDNA extracts from filter membranes.
Two extractions were performed for each time point which were combined for PCR and sequencing.

Extract Number	Collection date	Extraction date	Site	DNA concentration (ng/µl)	Time point	DNA concentration - Combined (ng/µl)
1	30/09/2013	18/10/2014	1	72.08	**T1**	53
2	30/09/2013	18/10/2014	2	25.34		
3	04/11/2013	18/10/2014	1	34.92	**T2**	37
4	04/11/2013	18/10/2014	2	34.09		
5	25/11/2013	18/10/2014	1	39.03	**T3**	25
6	25/11/2013	18/10/2014	2	11.81		
7	17/12/2013	06/10/2014	1	24.68	**T4**	56
8	17/12/2013	06/10/2014	2	90.58		
9	08/01/2014	06/10/2014	1	45.5	**T5**	46
10	08/01/2014	06/10/2014	2	46.06		
11	29/01/2014	06/10/2014	1	25.73	**T6**	24
12	29/01/2014	06/10/2014	2	21.24		
13	22/02/2014	07/10/2014	1	58.81	**T7**	52
14	22/02/2014	07/10/2014	2	46.87		
15	12/03/2014	06/10/2014	1	36.62	**T8**	36
16	12/03/2014	06/10/2014	2	37.69		
17	07/04/2014	07/10/2014	1	77.77	**T9**	76
18	07/04/2014	07/10/2014	2	75.19		
19	30/04/2014	06/10/2014	1	47.33	**T10**	49
20	30/04/2014	06/10/2014	2	49.72		
21	20/05/2014	18/10/2014	1	80.05	**T11**	68
22	20/05/2014	18/10/2014	2	52.55		
23	10/06/2014	07/10/2014	1	44.49	**T12**	47
24	10/06/2014	07/10/2014	2	50.33		
25	02/07/2014	18/10/2014	1	37.74	**T13**	48
26	02/07/2014	18/10/2014	2	44.94		
27	23/07/2014	18/10/2014	1	66.18	**T14**	62
28	23/07/2014	18/10/2014	2	45.93		
29	12/08/2014	18/10/2014	1	90.28	**T15**	68
30	12/08/2014	18/10/2014	2	35.4		
31	04/09/2014	18/10/2014	1	80.02	**T16**	65
32	04/09/2014	18/10/2014	2	41.33		

Supplementary Table ST 5: Summary of DNA extracts from exuviae community samples.

Extract Number	Collection date	Extraction date	DNA concentration (ng/µl)	Time point	Method
1	30/09/2013	23/11/2014	36.98	E1	QIAmp Blood Maxi
2	04/11/2013	19/11/2014	9.81	E2	Qiagen B & T Kit
3	25/11/2013	19/11/2014	6.59	E3	Qiagen B & T Kit
4	17/12/2013	19/11/2014	12.26	E4	Qiagen B & T Kit
5	08/01/2014	19/11/2014	9.57	E5	Qiagen B & T Kit
6	29/01/2014	19/11/2014	9.01	E6	Qiagen B & T Kit
7	22/02/2014	19/11/2014	6.13	E7	Qiagen B & T Kit
8	12/03/2014	19/11/2014	12.15	E8	Qiagen B & T Kit
9	07/04/2014	19/11/2014	16.5	E9	Qiagen B & T Kit
10	30/04/2014	23/11/2014	35.7	E10	QIAmp Blood Maxi
11	20/05/2014	23/11/2014	31.31	E11	QIAmp Blood Maxi
12	10/06/2014	23/11/2014	30.15	E12	QIAmp Blood Maxi
13	02/07/2014	23/11/2014	18.15	E13	QIAmp Blood Maxi
14	23/07/2014	23/11/2014	19.9	E14	QIAmp Blood Maxi
15	12/08/2014	23/11/2014	19.89	E15	QIAmp Blood Maxi
16	04/09/2014	23/11/2014	25.42	E16	QIAmp Blood Maxi

Supplementary Table ST 6: Primers used for library preparation.
Round 1: forward / reverse universal tail and template specific primer. A multi N region inserted in forward primer to assist cluster formation. Round 2: a forward or reverse Illumina adapter and an i5 or i7 Nextera index with the appropriate universal tail.

Primer pair	Round 1	Direction
LCO1490	**Forward Universal tail** **Template specific primer** ACACTCTTTCCCTACACGACGCTCTTCCGATCT **NNNNN** GGTCAACAAATCATAAAGATATTGG	**Forward**
HC02198	**Reverse Universal tail** **Template specific primer** GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAAACTTCAGGGTGACCAAAAAATCA	**Reverse**
COI_A_rev	**Reverse Universal tail** **Template specific primer** GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCARAAWCTTATATTATTTATTCGDGG	**Reverse**
	Round 2	
All Forward	**P5 Illumina adapter** **Index 2 (i5)** **Forward Universal tail** 5' AATGATACGGCGACCACCGAGATCTACAC - i5 Index - ACACTCTTTCCCTACACGACGCTC 3'	**Forward**
All Reverse	**P7 Illumina adapter** **Index 1 (i7)** **Reverse Universal tail** 5' CAAGCAGAAGACGGCATACGAGAT - i7 Index - GTGACTGGAGTTCAGACGTGTGCTC 3'	**Reverse**

References

Armitage, P.D., Pinder, L.C. & Cranston, P. (2012). *The Chironomidae: biology and ecology of non-biting midges*. Chapman and Hall, London.

Baird, D.J. & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.

Barnes, M.A. & Turner, C.R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, **17**, 1–17.

Barnes, M.A., Turner, C.R., Jerde, C.L., Renshaw, M.A., Chadderton, W.L. & Lodge, D.M. (2014). Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science and Technology*, **48**, 1819–1827.

Berry, D., Mahfoudh, K. Ben, Wagner, M. & Loy, A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology*, **77**, 7846–7849.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R.A., Foster, J., Wilkinson, J.W., Arnell, A., Brotherton, P., Williams, P. & Dunn, F. (2015). Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). *Biological Conservation*, **183**, 19–28.

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., Yu, D.W. & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*, **29**, 358–367.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009). BLAST plus: architecture and applications. *BMC Bioinformatics*, **10**, 1.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., Mcdonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. & Knight, R. (2010). QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nature methods*, **7**, 335–336.

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., A.Wardle, D., Kinzig, A.P., Daily, G.C., Loreau, M., Grace, J.B., Larigauderie, A., Srivastava, D.S. & Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, **489**, 326–326.

Carew, M., Pettigrove, V., Metzeling, L. & Hoffmann, A. (2013). Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology*, **10**, 45.

Chave, J. (2013). The problem of pattern and scale in ecology: What have we learned in 20 years? *Ecology Letters*, **16**, 4–16.

Clarke, K.R. & Gorley, R.N. (2006). Primer v6: User Manual/Tutorial. 192.

Deagle, B.E., Eveson, J.P. & Jarman, S.N. (2006). Quantification of damage in DNA recovered from highly degraded samples--a case study on DNA in faeces. *Frontiers in zoology*, **3**, 11.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., Taberlet, P., Taberlet, P., Coissac, E., Hajibabaei, M., Rieseberg, L., Yu, D., Ji, Y., Emerson, B., Wang, X., Ye, C., Yang, C., Ding, Z., Ficetola, G., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessiere, J., Taberlet, P., Pompanon, F., Geller, J., Meyer, C., Parker, M., Hawk, H., Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glockner, F., Bru, D., Martin-Laurent, F., Philippot, L., Schloss, P., Gevers, D., Westcott, S., Clarke, L., Soubrier, J., Weyrich, L., Cooper, A., Ji, Y., Barba, M. De, Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., Taberlet, P., Leray, M., Yang, J., Meyer, C., Mills, S., Agudelo, N., Ranwez, V., Boehm, J., Machida, R., Little, D., Deagle, B., Kirkwood, R., Jarman, S., Zhou, X., Shokralla, S., Gibson, J., Nikbakht, H., Janzen, D., Hallwachs, W. & Hajibabaei, M. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology letters*, **10**, 1789–1793.

Deiner, K. & Altermatt, F. (2014). Transport distance of invertebrate environmental DNA in a natural river. *PLoS ONE*, **9**, e88786.

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

Deiner, K., Walser, J.C., Mächler, E. & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, **183**, 53–63.

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P. & Miaud, C. (2011). Persistence of environmental DNA in freshwater ecosystems. *PLoS ONE*, **6**, e23398.

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E. & Miaud, C. (2012). Improved detection of an alien invasive species through environmental DNA barcoding: The example of the American bullfrog Lithobates catesbeianus. *Journal of Applied Ecology*, **49**, 953–959.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Ekrem, T., Willassen, E. & Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution*, **43**, 530–542.

Evans, N.T., Olds, B.P., Renshaw, M.A., Turner, C.R., Li, Y., Jerde, C.L., Mahon, A.R., Pfrender, M.E., Lamberti, G.A. & Lodge, D.M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, **16**, 29–41.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for

amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

Fonseca, V.G., Carvalho, G.R., Sung, W., Johnson, H.F., Power, D.M., Neill, S.P., Packer, M., Blaxter, M.L., Lambshead, P.J.D., Thomas, W.K. & Creer, S. (2010). Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature communications*, **1**, 98.

Gibson, J.F., Shokralla, S., Curry, C., Baird, D.J., Monk, W.A., King, I. & Hajibabaei, M. (2015). Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS ONE*, **10**, 1–15.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 8007–12.

Goecks, J., Nekrutenko, A. & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, **11**, R86.

Goldberg, C.S., Sepulveda, A., Ray, A., Baumgardt, J. & Waits, L.P. (2013). Environmental DNA as a new method for early detection of New Zealand mudsnails (*Potamopyrgus antipodarum*). *Freshwater Science*, **32**, 792–800.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C. & Baird, D.J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, **6**, e17497.

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

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P., Edwards, F.A., Larsen, T.H., Hsu, W.W., Benedick, S., Hamer, K.C., Wilcove, D.S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B.C. & Yu, D.W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Joshi, N. & Fass, J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. *Available at https://github.com/najoshi/sickle.*, 2011.

Kelly, R.P., Port, J.A., Yamahara, K.M. & Crowder, L.B. (2014a). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS ONE*, **9**, e86175.

Kelly, R.P., Port, J. a., Yamahara, K.M., Martone, R.G., Lowell, N., Thomsen, P.F., Mach, M.E., Bennett, M., Prahler, E., Caldwell, M.R. & Crowder, L.B. (2014b). Harnessing DNA to improve environmental management. *Science*, **344**, 1455–1456.

Klymus, K.E., Richter, C.A., Chapman, D.C. & Paukert, C. (2015). Quantification of eDNA

shedding rates from invasive bighead carp Hypophthalmichthys nobilis and silver carp Hypophthalmichthys molitrix. *Biological Conservation*, **183**, 77–84.

Lacoursière-Roussel, A., Rosabal, M. & Bernatchez, L. (2016). Estimating fish abundance and biomass from eDNA concentrations: variability among capture methods and environmental conditions. *Molecular Ecology Resources*, **16**, 1401–1414.

Lawson Handley, L. (2015). How will the 'molecular revolution' contribute to biological recording? *Biological Journal of the Linnean Society*, **115**, 750–766.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, **362**, 709–715.

Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., Zhang, H., Misof, B., Kjer, K.M., Tang, M., Niehuis, O., Jiang, H. & Zhou, X. (2016). Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, **16**, 470–479.

Lodge, D.M., Turner, C.R., Jerde, C.L., Barnes, M.A., Chadderton, L., Egan, S.P., Feder, J.L., Mahon, A.R. & Pfrender, M.E. (2012). Conservation in a cup of water: Estimating biodiversity and population abundance from environmental DNA. *Molecular Ecology*, **21**, 2555–2558.

Loreau, M. & de Mazancourt, C. (2013). Biodiversity and ecosystem stability: A synthesis of underlying mechanisms. *Ecology Letters*, **16**, 106–115.

Mächler, E., Deiner, K., Steinmann, P. & Altermatt, F. (2014). Utility of environmental DNA for monitoring rare and indicator macroinvertebrate species. *Freshwater Science*, **33**, 1174–1183.

Magurran, A.E., Baillie, S.R., Buckland, S.T., Dick, J.M., Elston, D.A., Scott, E.M., Smith, R.I., Somerfield, P.J. & Watt, A.D. (2010). Long-term datasets in biodiversity research and monitoring: Assessing change in ecological communities through time. *Trends in Ecology and Evolution*, **25**, 574–582.

Magurran, A.E. & McGill, B.J. (2011). *Biological diversity: frontiers in measurement and assessment* (A.E. Magurran & B.J. McGill, Eds.). Oxford University Press.

Martin, M. (2011). Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011. Date of access 05/08/2015. *EMBnet*, **17**, 10–12.

Met Office. (2016). Hadley Centre for Climate Prediction and Research (JCHMR), Maclean Building, Wallingford OX10 8BB, UK.

Minamoto, T., Yamanaka, H., Takahara, T., Honjo, M.N. & Kawabata, Z. (2012). Surveillance of fish species composition using environmental DNA. *Limnology*, **13**, 193–197.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J.Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M. & Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science*, **2**, 150088.

Moss, B.R. (2010). *Ecology of Fresh Waters: A View for the Twenty-First Century*. Wiley-

Blackwell.

Murray, D.C., Coghlan, M.L. & Bunce, M. (2015). From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. *PLoS ONE*, **10**, e0124671.

O'Donnell, J.L., Kelly, R.P., Lowell, N.C. & Port, J.A. (2016). Indexed PCR primers induce template- Specific bias in Large-Scale DNA sequencing studies (A.R. Mahon, Ed.). *PLoS ONE*, **11**, e0148698.

Pilliod, D.S., Goldberg, C.S., Arkle, R.S., Waits, L.P. & Richardson, J. (2013). Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences*, **70**, 1123–1130.

Ratnasingham, S. & Hebert, P.D.N. (2007). BOLD: The Barcode of Life Data System: Barcoding. *Molecular Ecology Notes*, **7**, 355–364.

Raunio, J., Paasivirta, L. & Hämäläinen, H. (2010). Assessing lake trophic status using spring-emerging chironomid pupal exuviae. *Fundamental and Applied Limnology / Archiv für Hydrobiologie*, **176**, 61–73.

Rees, H.C., Maddison, B.C., Middleditch, D.J., Patmore, J.R.M. & Gough, K.C. (2014). The detection of aquatic animal species using environmental DNA - a review of eDNA as a survey tool in ecology (E. Crispo, Ed.). *Journal of Applied Ecology*, **51**, 1450–1459.

Renshaw, M.A., Olds, B.P., Jerde, C.L., Mcveigh, M.M. & Lodge, D.M. (2015). The room temperature preservation of filtered environmental DNA samples and assimilation into a phenol-chloroform-isoamyl alcohol DNA extraction. *Molecular Ecology Resources*, **15**, 168–176.

Ruse, L.P. (2013). Chironomid (Diptera) species recorded from UK lakes as pupal exuviae. *Journal of Entomological and Acarological Research*, **45**, 13.

Ruse, L. (2010). Classification of nutrient impact on lakes using the chironomid pupal exuvial technique. *Ecological Indicators*, **10**, 594–601.

Ruse, L. (2011). Lake acidification assessed using chironomid pupal exuviae. *Fundamental and Applied Limnology / Archiv für Hydrobiologie*, **178**, 267–286.

Strickler, K.M., Fremier, A.K. & Goldberg, C.S. (2015). Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biological Conservation*, **183**, 85–92.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012a). Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012b). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Møller, P.R., Rasmussen, M. & Willerslev, E. (2012a). Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples. *PLoS ONE*, **7**, e41732.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Wiuf, C., Rasmunsen, M., Gilbert, M.T.P., Orlando, L. & Willerslev, E. (2012b). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, **21**, 2565–2573.

Thomsen, P.F. & Willerslev, E. (2015). Environmental DNA - An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, **183**, 4–18.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P.F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G.H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.M., Peroux, T., Crivelli, A.J., Olivier, A., Acqueberge, M., Le Brun, M., M??ller, P.R., Willerslev, E. & Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, **25**, 929–942.

Wilson, R. & Ruse, L. (2005). *A guide to the identification of genera of chironomid pupal exuviae occurring in Britain and Ireland*. Freshwater Biological Association Publication 13, Ambleside, UK., Ambleside, Cumbria.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **73**, 3–36.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Chapter 4

Investigating the performance of amplicon vs. shotgun sequencing for biomass estimation in macroinvertebrate community samples

Chapter 4: Investigating the performance of amplicon vs. shotgun sequencing for biomass estimation in macroinvertebrate community samples

4.1 Abstract

New applications of DNA and RNA sequencing are emerging and expanding in the field of ecological monitoring, yet questions remain regarding their precision and efficiency. Due to primer bias issues, the ability of metabarcoding to depict with accuracy the relative abundances of taxa from mixed communities has been questioned, while PCR-free whole mito-genome sequencing has been suggested as a possibly more reliable alternative. Here we used a set of carefully designed mock communities comprising 13 species of freshwater macroinvertebrates (precisely measured for biomass content), to compare the accuracy of COI metabarcoding (3 amplicons) vs. shotgun mito-metagenome sequencing. Additionally, COI barcoding and shotgun mito-genome sequencing for individual specimens was performed, to provide reference sequences for OTU assignment and mito-metagenome assembly respectively.

We found that even though both methods occasionally failed to recover very low abundance species, metabarcoding was more inconsistent by failing to recover some species with higher abundance as well, probably due to primer bias. Shotgun sequencing results provided highly significant correlations between read number and biomass in all but one species. Conversely, the read-biomass relationships obtained from amplicon sequencing were not significant for 5 out of 13 (amplicons B1FR-450bp, FF130R-130bp) or 8 out of 13 (amplicon FFFR, 658bp) species. Combining the results of all three amplicons (multi-amplicon approach), improved the read-biomass correlations for some of the species. Overall, we propose that shotgun mito-metagenomic sequencing outperforms metabarcoding in the accuracy of species biomass predictions for bulk communities of macroinvertebrates.

4.2 Introduction

4.2.1 Importance of accurate biodiversity assessment and the sequencing revolution

The accurate qualitative and quantitative assessment of biodiversity is essential in order to understand biodiversity and ecosystem function relationships, especially in the face of rapid biodiversity loss (Loreau & de Mazancourt 2013). However, the scale and intensity of contemporary biodiversity identification challenges are limited by the use of traditional taxonomic approaches (Jackson *et al.* 2014). Meanwhile, international directives require the application of sufficient monitoring of water bodies such as the Water Framework Directive (WFD), which is a legislation for management and protection of European aquatic ecosystems (Collins *et al.* 2012). In biomonitoring, the accurate quantification of community composition enables detection of both spatial and temporal variations in the biological community and by extension, the wider ecosystem (Cranston 1990). Traditional ecological assessment methods used for biomonitoring largely rely upon taxonomic identification of species, a practise that is labour intensive and inherently time consuming, requiring high-level taxonomic expertise for species-level identification and can be insufficient in case of damaged or immature specimens and certain life stages (Sweeney *et al.* 2011; Jackson *et al.* 2014).

The DNA sequencing revolution implemented by the advent of high throughput sequencing technologies (HTS) is revolutionising biomonitoring by increasing the throughput and taxonomic information that can be recovered (Baird & Hajibabaei 2012). The most commonly used taxonomic groups used for testing this work include various invertebrate taxa, such as benthic macroinvertebrates for freshwater ecosystem studies (e.g. Pfrender *et al.* 2010; Hajibabaei *et al.* 2011; Gibson *et al.* 2014, 2015; Shokralla *et al.* 2015). Similarly, terrestrial invertebrate taxa have been used, from soil or leaf litter (Yang *et al.* 2014), or from above ground invertebrate sampling (Malaise traps) (Ji *et al.* 2013). More recent work is also advancing into the detection of biodiversity from aqueous environmental DNA (eDNA), mainly through PCR-based detection (Mächler *et al.* 2014), and eDNA metabarcoding (fish and amphibian detection) (Valentini *et al.* 2016; Shaw *et al.* 2016; Hänfling *et al.* 2016).

A large majority of studies using HTS for diversity assessment of mixed samples to date utilise metabarcoding methodologies. Metabarcoding is a PCR based approach, where a selected

marker is amplified and sequenced with HTS, from bulk/environmental community samples, extracted from mixed tissue samples (Yu *et al.* 2012). Most commonly used markers for metabarcoding include the Cytochrome Oxidase Subunit I (COI) barcoding region, but also ribosomal RNA regions 16S (Epp *et al.* 2012), or RbcL and matK for plants (Hollingsworth *et al.* 2009).

4.2.2 Possible biases related to metabarcoding work

PCR based metabarcoding work has been the workhorse of contemporary biodiversity analysis. Due to intermediate PCR steps, it has been argued that the approach produces biases when it comes to accurately representing the diversity in bulk samples (Hajibabaei *et al.* 2012; Yu *et al.* 2012). In fact, it has been suggested that PCR biases might alter the biomass ratio of species, skew the relative abundance of species, or produce inaccurate representation of abundance of species in a given sample (Piñol *et al.* 2015). Furthermore, primer-template mismatches will also introduce biases through mis-representation of particular groups, as has been observed through metabarcoding of model invertebrate communities (Clarke *et al.* 2014; Elbrecht & Leese 2015). Other studies however, have reported significant relationships between biomass and number of reads for a selected number of species (Kelly *et al.* 2014; Hiiesalu *et al.* 2014). Moreover, the investigation of highly diverse samples from oyster reef communities (Leray & Knowlton 2015) found that metabarcoding OTU counts were strongly correlated with the amount of extracted DNA.

To deal with the uncertainties above, optimisation of metabarcoding work and use of multiple primer pairs has been suggested (Hajibabaei *et al.* 2012). The combination of multiple amplicons from the same region (Hajibabaei *et al.* 2012) or from different genes (Zhan *et al.* 2014; Gibson *et al.* 2014) has shown significant increases in the recovery of species richness compared to using individual primer pairs or single loci. While the use of multiple primer pairs in this work was mainly intended to investigate increase in richness detection, the same strategy should be investigated as a means of determining relative abundance of species as well.

4.2.3 Introducing Mito-metagenomics

Mitochondrial metagenomics (or mito-metagenomics) is a recently characterised research area involving the use of whole mitochondrial genome sequencing from bulk specimen samples (Crampton-Platt *et al.* 2016). Mito-metagenomics takes advantage of the high throughput of Illumina sequencing, producing millions of short reads, which are then assembled to provide shorter contigs, up to near complete mitochondrial genomes of the organisms in the mix. Current applications involve characterisation of bulk samples for ecological assessment (Tang *et al.* 2015) and phylogenetic reconstruction of multiple species simultaneously (Gillett *et al.* 2014). This approach, utilising shotgun sequencing of bulk invertebrate samples, has also been suggested by Zhou et al. (2013) as an alternative to PCR-based metabarcoding work. It is advocated that the absence of a PCR step will result in more accurate biomass to reads relationship; hence, this method could be more reliable for accurate representation of relative abundance of species in bulk samples.

For mito-metagenomic work, two distinct paths could be used for data analysis, described as "read-based" or "contig based" depending on whether reference sequences or a *de novo* approach are used (Gómez-Rodríguez *et al.* 2015). Though *de novo* assembly of mito-metagenomes is achievable, the use of reference mito-genomes or COI barcode reference databases have been found to increase the accuracy of the method (Gómez-Rodríguez *et al.* 2015). Nevertheless, the latter approach would of course increase the cost due to necessary steps for library construction at least at the initial stage, not to mention the time investment in the generation of DNA references.

4.2.4 Aims and hypothesis

The main aim of this chapter is to compare the efficiency and applicability of the two currently most prominent approaches for HTS of bulk invertebrate samples in relation to quantification of taxon biomass. To achieve a quantitative comparison between the two methods, a structured design of mock macroinvertebrate communities with known biomass content was

used, which were amplicon sequenced for three COI amplicons (MiSeq), as well as shotgun sequenced on HiSeq (for an overview of the experimental workflow see Figure 4.1).

We hypothesize that applying PCR-free mito-metagenomics sequencing would provide more accuracy in community composition quantification, as metabarcoding can be variably influenced by PCR-related issues, associated with primer bias and variable presence of DNA copies in mixed samples. Furthermore, the overall applicability of each method will be assessed while providing suggestions for future improvements. Ultimately, we aim to provide a comprehensive evaluation of the two methods and troubleshoot their future usage for ecological applications in biodiversity and freshwater ecosystem monitoring.

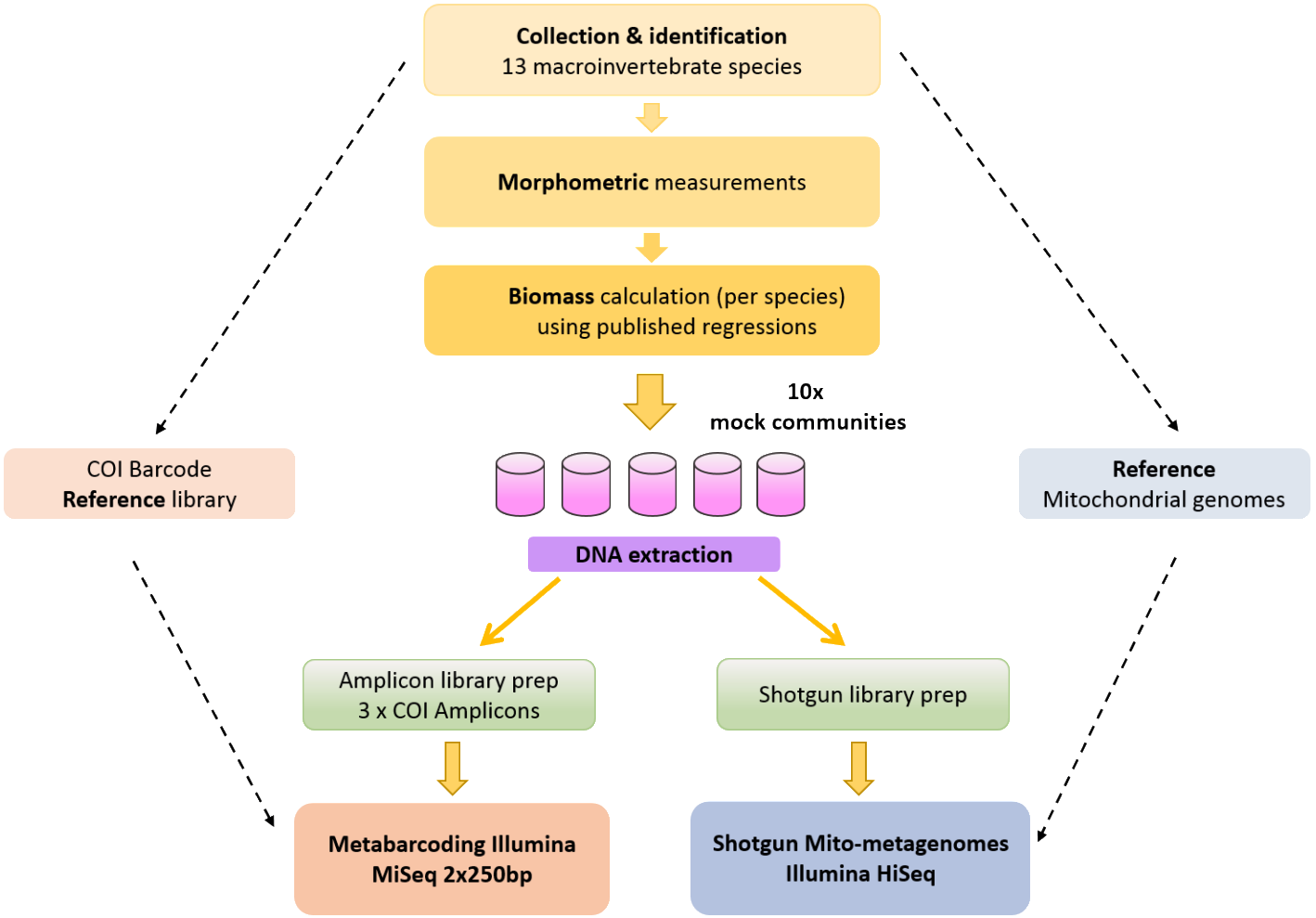


Figure 4.1: Brief overview of experimental workflow.
Yellow arrows indicate steps of laboratory work, including specimen processing, molecular work and sequencing for mock communities. Dashed lines indicate parallel laboratory steps for production of reference barcodes and mito-genomes from individual specimens. Acquisition of sequencing results was followed by bioinformatics and statistical analysis steps.

4.3 Methods

4.3.1 Sample collection

Specimens for this work were collected from the areas of Somerset and Suffolk by volunteer county surveyors over the period September – October 2014, and were identified to species level by the county surveyors in the first instance and preserved in absolute ethanol. The specimens were stored in replenished 100% ethanol and stored in a dark, dry and cool environment until morphological measurements and DNA extraction (smaller species were kept at 4°C). Subsequently, all specimens were sent to APEM Ltd., which is an ISO certified lab, for quality control (QC) of taxonomic identification. Misidentified specimens were removed from further work.

In total, 13 species were used for analysis, including eight species of Gastropoda and one of each from: Hemiptera, Isopoda, Amphipoda, Ephemeroptera and Coleoptera (Table 4.1, Figure 4.2). These species were selected to include a wide variety of taxonomic orders, resembling a natural community. Since a large number of specimens per species were required to allow sufficient differences in biomass among replicate communities, we also aimed for commonly occurring species. This also limited the number of different sites required for sample collection, aiming to limit intraspecific diversity in the emergent data.

4.3.2 Morphological measurements

Each species was measured morphologically according to published work, using appropriate body measurements that would produce an accurate representation of biomass. Different methodologies were used for measurement (See Table 4.1 for measurement taken for each species). Callipers were used for larger animals (*N. glauca, A. aquaticus, G. marinus, E. danica*), while smaller species were measured using a microscope fitted with an ocular micrometre (*P. antipodarum*). For the amphipod species (*G. pulex*), the software Image Pro paired with a stereoscopic microscope was used, to facilitate accurate measurements, by accounting for the curvature of specimens.

For estimation of biomass for each species, published regressions were used (Supplementary Table S 4.1). Conversion of length to mass is considered superior to other methodologies such

as determination of biovolume or weighing of specimens, due to increased precision and speed (Benke *et al.* 1999). Because for some species there was no equation available at the species level, the closest taxonomic group equation available was used. For *B. tentaculata* a species level regression was used (Baumgärtner & Rothhaupt 2003), which was also applied for *B. leachii* as a congeneric species (Bithynia). Similarly, a species level regression was used for *P. fontinalis*, adopted from Caquet (1993). Species specific regressions were used for *P. antipodarum* (Mährlein *et al.* 2016), *R. balthica* and *A. aquaticus* (Baumgärtner & Rothhaupt 2003). For the species *A. vortex*, *B. contortus* and *P. planorbis* a family level regression was used (Planorbidae), which was originally developed for species *Anisus rotundatus* (Family: Planorbidae) by Caquet (1993). For the remaining species, higher taxonomic level equations were used: family Gyrinidae for *G. marinus*, genus Gammarus (*G. minus*) for *G. pulex*, genus Ephemera for *E. danica*, and order level, Hemiptera for *N. glauca*, all adopted from Benke et al (1999).

Regression equations were selected for each species from studies that were as close as possible to the geographic region and ecosystem type in this study, as it has been suggested that these parameters could produce variation in within species development rates (Mährlein *et al.* 2016). Most specimens in this study were collected from shallow ponds hence using data from lake environments was preferred. Regarding the geographic region, in some cases we had to use equations developed from distant geographical areas for some species, as the number of studies available for European specimens is currently limited (Mährlein *et al.* 2016).

4.3.3 DNA barcode Reference Library

In addition to the bulk biomass community constructions, individual specimens were extracted and sequenced for the COI barcoding region using universal metazoan primers (Folmer *et al.* 1994) (see also Chapter 1). Different extraction protocols were employed according to tissue type: gastropod species were extracted with a CTAB chloroform protocol, and arthropods with a DNEasy Blood & Tissue (QIAGEN) extraction kit according to manufacturer's instructions. Good quality barcodes were obtained from all species (Table 4.1), except *E. danica* (Ephemeroptera), for which barcode sequencing was not successful.

The specimens selected for barcoding were representative of the different sampling locations to account for possible intraspecific variation. Sanger generated sequences were edited using CodonCode Aligner v.3.7.1 (CodonCode Corporation, Massachusetts). Alignment was performed using ClustalW in MEGA (Tamura *et al.* 2007), which was also used for detection of possible stop codons and insertion and deletion events.

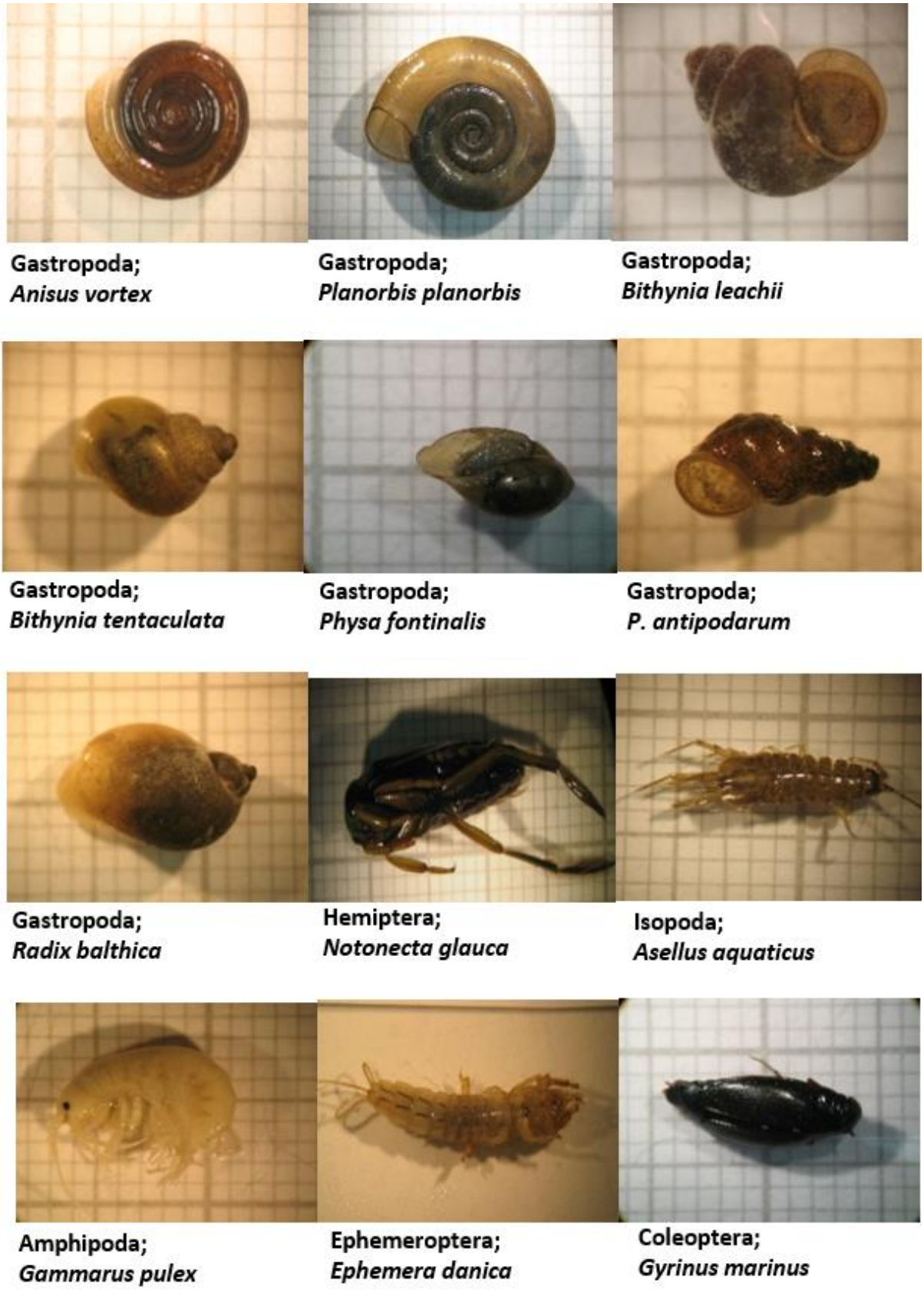


Figure 4.2: Species used for the construction of the mock communities.

Table 4.1: Species collected for construction of mock communities.
Taxonomic classification (Class/Order/Family/Species) and measurements taken (SW: Shell Width, AW: Aperture Width, BL: Body Length, HW: Head Width). The extraction method and number of individual COI barcodes sequenced are shown in the last two columns.

Number	Class/Order	Family	Species	Measureme	Barcode	Extraction
1	Mollusca/Gastropoda	Planorbidae	*Anisus vortex*	SW	6	CTAB
2	Mollusca/Gastropoda	Planorbidae	*Bathyomphalus contortus*	SW	3	CTAB
3	Mollusca/Gastropoda	Planorbidae	*Planorbis planorbis*	SW	4	CTAB
4	Mollusca/Gastropoda	Bithyniidae	*Bithynia leachi*	SH and AW	2	CTAB
5	Mollusca/Gastropoda	Bithyniidae	*Bithynia tentaculata*	SH and AW	2	CTAB
6	Mollusca/Gastropoda	Physidae	*Physa fontinalis*	SH and AW	4	CTAB
7	Mollusca/Gastropoda	Hydrobiidae	*Potamopyrgus antipodarum*	SH and AW	6	CTAB
8	Mollusca/Gastropoda	Lymnaeidae	*Radix balthica*	SH and AW	12	CTAB
9	Insecta/Hemiptera	Notonectidae	*Notonecta glauca*	BL	2	DNeasy
10	Crustacea/Isopoda	Asellidae	*Asellus aquaticus*	BL and HW	3	DNeasy
11	Crustacea/Amphipoda	Gammaridae	*Gammarus pulex*	BL	4	DNeasy
12	Insecta/Ephemeropter	Ephemeridae	*Ephemera danica*	BL	--	DNeasy
13	Insecta/Coleoptera	Gyrinidae	*Gyrinus marinus*	BL	3	DNeasy

4.3.4 Design of mock communities

The mock communities were designed to represent different sums of biomass per species and allow sufficient replication simultaneously. To increase statistical power, 10 communities were created containing either 13 or 14 species, with 136 to 156 specimens each (Table 4.2). Due to insufficient number of specimens, two of the species were present only in some of the communities (six for *N. glauca* and nine *P. fontinalis*). Every species was represented by a single specimen in only one occasion (a mean sized individual was used as a single representative). Depending on the number of available specimens, larger steps in number were implemented, while an effort was made to include specimens from a variety of body sizes in each community (including natural variability of body size and aiming for a similar mean body size across communities). An overview of the contents of each community in terms of numbers of species and corresponding specimens are presented in Table 4.2. For

detailed contents of communities based on morphological measurements and mass conversion, see Supplementary Table S4.2, and percentage contents of species Supplementary Table S4. 3. Supplementary Figure S4.1 provides a graphical representation of community composition as relative abundance of species contained in each community.

Positive controls. To assess the quality of sequencing performance across communities, three whole bodies of *D. melanogaster* were included in each community (prior to DNA extraction), to act as a positive control of extraction efficiency across all communities. Additionally, for the shotgun method, a second positive control was included. Here DNA extract of the Lepidopteran species *Mycalesis mineus* was added to the extracted community DNA at Shenzen, China by collaborators. This species had been previously sequenced for its mitochondrial genome by co-authors of this work, providing a reference mito-genome sequence. The species *D. melanogaster* was selected due to its model status and wide availability of mito-genome sequence information in public databases. We used D. melanogaster as a positive control of the efficiency of the DNA extraction method, while *M. mineus* (inserted at equal concentrations) was used to account for variability in shotgun sequencing efficiency.

Table 4.2: **Design of mock macroinvertebrate communities.**
See columns for the detailed contents of each community (1-10). The numbers refer to specimens from each species included in each community. Total number of specimens/ species (last column), and total number of specimens/community and number of species / community (bottom) are shown. Highlighted the species with lowest abundance (yellow) and highest abundance (grey) in each community, and five cases when the particular species was missing from that community (green).

		Community										
Number	**Species**	**1 Alpha**	**2 Bravo**	**3 Charlie**	**4 Delta**	**5 Echo**	**6 Fox**	**7 George**	**8 Henry**	**9 India**	**10 Julia**	**Specimens per species**
1	*Anisus vortex*	35	40	45	5	25	20	15	10	30	1	226
2	*Asellus aquaticus*	1	4	8	10	14	17	19	21	24	24	142
3	*Bathyomphalus contortus*	14	13	12	11	10	8	6	1	2	4	81
4	*Bithynia tentaculata*	24	10	6	25	26	1	27	15	20	26	180
5	*Ephemera danica*	16	3	1	6	8	12	10	18	14	20	108
6	*Gyrinus marinus*	2	1	3	10	4	8	5	9	6	7	55
7	*Planorbis planorbis*	24	25	19	22	1	4	7	10	13	16	141
8	*Potamopyrgus antipodarum*	10	32	28	25	21	33	14	17	1	5	186
9	*Radix balthica*	3	15	5	17	16	10	12	1	9	6	94
10	*Physa fontinalis*	1	3	4	6	8	10	12	13	13	0	70
11	*Notonecta glauca*	10	0	0	4	2	1	0	6	0	8	31
12	*Bithynia leachi*	12	3	5	1	9	11	8	7	13	14	83
13	*Gammarus pulex*	2	5	6	4	8	8	1	8	3	7	52
14	*Drosophila melanogaster*	3	3	3	3	3	3	3	3	3	3	30
	Total specimens	**157**	**157**	**145**	**149**	**155**	**146**	**139**	**139**	**151**	**141**	**1479**
	Total Nº of species	**14**	**13**	**13**	**14**	**14**	**14**	**13**	**14**	**13**	**13**	

4.3.5 DNA extraction for reference mito-genomes and bulk communities

For the construction of individual shotgun reference genomes for each species, high quality genomic DNA was extracted from a single specimen (where possible) using the Qiagen Blood and Tissue extraction kit. Final elution was performed using 50µl PCR Grade water (Roche). To minimise contamination of the target genomic sequences, we used either leg or muscle tissue, avoiding the specimens' guts. DNA quality and concentration was assessed with dsQubit assays and agarose gel electrophoresis. A minimum amount of 25µg total DNA was used for shotgun sequencing. For the species *A. vortex*, DNA extraction did not yield sufficient quality of genomic DNA and consequently this species was not sequenced for a reference mitochondrial genome.

For the mock communities, DNA was extracted from whole bodies of invertebrates (specimens previously used for barcoding were excluded so as not to alter biomass measurements). The re-combined communities were stored in 50ml falcon tubes in absolute ethanol. First, ethanol was carefully poured out and specimens were patted with blue roll, before they were allowed to dry at 37^{o}C for 2 hours in a clean plate. Sterile mortar and pestle sets were used to grind the dried specimens to as fine matter as possible, which was then transferred into 50ml Power Bead tubes from the Power Max Soil DNA Isolation Kit (MO-BIO) and vortexed at high speed for 5min. Subsequently, 450µl of Proteinase K, 20mg/ml (Sigma-Aldrich) was added and the bead tubes were placed at 65˚C in a shaker at medium speed to incubate for 3h. The manufacturer's protocol was followed for the next steps. For the final elution, the columns were allowed to incubate for 30min and were then centrifuged at 2500g for 5min. This step was repeated a second time to allow maximum recovery of DNA. All communities yielded between 48-98ng/µl DNA in 4ml final eluate (Supplementary Table S4.4) (Supplementary Figure S4.2).

4.3.6 Metabarcoding - Primer selection

For metabarcoding, a multi-amplicon approach was used, as it has been suggested that use of multiple primer pairs can increase diversity detection in bulk samples (Gibson *et al.* 2014). This approach was also used here to account for possible effects of multiple amplicons in accuracy of biomass estimations in mixed community samples. Overall, three primer pairs were selected covering different parts of the COI barcode region. (1) Whole barcoding region (amplicon FFFR) (658bp): universal Folmer primers (Folmer *et al.* 1994), (2) Folmer forward primer - 130R primer (amplicon FF130R) (130bp), (3) B1 forward primer - Folmer reverse (amplicon B1FR) (450bp) covering the length of the COI Barcoding region [B1 modified from (Hajibabaei *et al.* 2012), and 130R unpublished] (Figure 4.3) (see Table 4.3 for primer sequences). The three amplicons featuring in the final work were selected as the most successful in amplifying our target taxa, out of five possible amplicons of the COI (visual primer match was checked against aligned barcodes from our database plus NCBI downloaded sequences). Primers B1 and 130R are degenerate, specifically modified for use with macroinvertebrate communities.

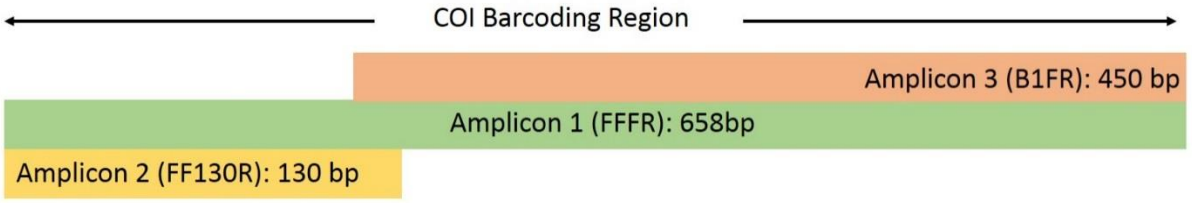


Figure 4.3: Positions of the sequenced amplicons on the COI Barcoding region.
Amplicons: 1. FFFR, 658bp (green), 2. FF130R, 130bp (yellow), 3. B1FR, 450bp (orange), according to the primer pair used.

4.3.7 Metabarcoding - Amplicon library preparation

Libraries were prepared using a three-step PCR protocol. For the first round, amplification was performed using only the target specific primer, then (purified) amplified product used as template for a second round of PCR using the template specific primers with added Illumina tails, and finally, a third round of PCR took place to add index sequences on the amplified product. The samples were sequenced on an Illumina MiSeq using 2x250bp chemistry.

PCRs were performed in 25µl reaction volumes containing, for **Round 1**: 5µl Buffer, 0.25µl Taq polymerase (Promega), 0.5µl BSA, 0.6µl (10nmole/µl) of each forward and reverse primer, 0.6µl dNTPs, 16.45µl PCR water and 1µl DNA (10ng/µl). For **Round 2**: 5µl Buffer, 0.25µl Taq polymerase (Promega), 0.5µl BSA, 0.6µl of each forward and reverse Illumina tailed primer, 0.6µl dNTPs, 12.45µl PCR water and 5µl purified PCR product from Round 1.

The following thermo-cycling conditions were used: **Round 1**: **FFFR**: Denaturation at 94°C for 2 min, 20 cycles of: 94°C for 30 sec, 45°C for 40 sec, 72°C for 1 min, followed by a 10min extension at 72°C, hold at 4°C. **B1FR, FF130R:** Denaturation at 94°C for 2 min, 23 cycles of: 94°C for 30 sec, 45°C for 40 sec, 72°C 1 min, followed by a 10min extension at 72°C, hold at 4°C. **Round 2: all amplicons**: Denaturation at 94°C for 2 min, 10 cycles of: 94°C for 30 sec, 45°C for 40 sec, 72°C for 1 min, followed by a 10min extension at 72°C, cool at 4°C for 10min. A third round of PCR was performed with product from Round 2, to attach Illumina indexes. Purification of PCR products between Round 1 and 2 was performed using an Exo-TSAP (Exonuclease – Thermosensitive Alkaline Phosphatase) protocol. A 3 step PCR protocol was selected to minimise the effects of variant index sequences on the amplification efficiency of each community (O'Donnell *et al.* 2016).

Table 4.3: COI primers used for metabarcoding.
Three amplicons were generated for the whole barcoding region as well as using combinations of new unpublished primers with the universal forward (F) and reverse (R) Folmer primers.

Primer Name	Primer Sequence	Direction	Citation
LCO1490	GGTCAACAAATCATAAAGATATTGG	F	Folmer *et al.* 1994
HC02198	TAAACTTCAGGGTGACCAAAAAATCA	R	Folmer *et al.* 1994
I-B1	CCHGATATAACITTYCCICG	F	Hajibabaei *et al*. 2012 (modified)
I-130R	GAAAATYATAAIGAAIGCRTGAGC	R	Not published

4.3.8 Amplicon data analysis

Sequences from the three COI amplicons were de-multiplexed and Illumina adaptors were trimmed using Cutadapt (Martin 2011) and Sickle (Joshi & Fass). Filtering and quality control

was performed in USEARCH v7 (Edgar 2010), and low quality sequences with Phred score <25, maximum expected error >1, and shorter than 100bp were discarded. High quality sequences were de-replicated, sorted by size and singletons were removed. For amplicons 2 (FF130R, 130bp) and 3 (B1FR, 450bp) the forward and reverse reads were merged with a 25bp minimum overlap. For amplicon 1 (FFFR, whole barcoding region 658bp) only the forward reads were used (R1). After visualization of read quality using FastQC (www.bioinformatics.babraham.ac.uk), the reads were truncated at 230bp length (>25 Phred score). This strategy was selected because the length of the original amplicon (658bp) did not allow sufficient overlap between the forward and reverse reads due to the current limitations of Illumina 2x250 MiSeq chemistry. Chimeras were removed with a *de novo* method, and a 97% similarity level was used for OTU clustering and generation of an OTU table in USEARCH. This level of similarity was used as a mean value for characterisation of the diverse taxa present in the bulk samples.

Taxonomy was assigned to the OTU table using Quantitative Insights In Microbial Ecology (QIIME) (Caporaso *et al.* 2010). Taxonomic identification of OTUs was performed in BLAST+ (megablast) (Camacho *et al.* 2009), against a reference COI database at a first instance. The database was compiled from NCBI GenBank, by downloading all COI sequences, longer than 100bp, with environmental sequences excluded (20th June 2015, N = 807,388 sequences), combined with our locally acquired barcode sequences (Table 4.1). Higher taxonomic level information was added using the GALAXY online software platform (Goecks *et al.* 2010). All analysis involving USEARCH, QIIME and BLAST was performed using High Performance Computing (HPC) Wales systems. The BLAST identified OTUs were aligned against our local barcode database and tested for the presence of stop codons and insertions in MEGA6 (Tamura *et al.* 2007). Alignment and phylogenetic analysis using a Neighbor-Joining (NJ) method (Saitou & Nei 1987) were also performed in MEGA6. Only the OTUs that BLASTed at >98% similarity with our reference barcodes and clustered closely with the known COI barcode sequences on the NJ tree were included in further analysis. When multiple OTUs were assigned to a single species, the total number of reads were collapsed into a sum per species.

4.3.9 Construction of reference mitogenomes

Genomic DNA extracted from individual species (all studied species except *A. vortex*) was used for sequencing of reference mito-genomes. For each sample, a library with insert size of 200bp was constructed following manufacturer's instruction (Illumina, Nextera), while 100bp PE reads of a whole Illumina HiSeq2000 lane were produced for 12 independent genomic reference libraries at Beijing Genome Institute (BGI)-Shenzhen. Library construction and assembly of reference mito-genomes and bulk samples was performed by collaborators in BGI-Shenzhen. Raw data from each species were filtered as previously described in Zhou et al. (2013), Tang et al. (2014) and Tang et al. (2015), removing reads with low quality or adaptor contamination. Clean data was assembled using SOAPdenovo-Trans (-K 71) (Xie *et al.* 2014) and IDBA-UD (Peng *et al.* 2012). Assembled sequences were annotated following Tang et al. (2015), to identify candidate mitogenome sequences, which were used for mitogenome reference construction, and then manual correction and checking were done as described by Tang et al. (2014). Thirteen protein-coding genes (PCG) were extracted from all mitogenomes, and each of them were aligned with corresponding reference protein-coding genes from 4 arthropod species (*Macrogyrus oblongus, Gammarus duebeni, Ligia oceanica* and *Siphlonurus immanis*) and 3 mollusc species (*Biomphalaria tenagophila, Physella acuta* and *Oncomelania hupensis*) using CLUSTALW 2.1 (Thompson *et al.* 1994). The translation frame was checked in MEGA6 (Tamura *et al.* 2007), to correct gap length generated inside protein-coding genes by the assembly program when constructing scaffolds based on paired-end reads. In addition, the original read-mapping was done and monitored by using BWA 0.6.2 (Li & Durbin 2009) and SAMTOOLS 0.1.19 (Li *et al.* 2009) respectively following (Tang *et al.* 2014, 2015).

4.3.10 Bioinformatics analysis of shotgun data (bulk communities)

Genomic DNA from the butterfly *Mycalesis mineus*, whose mitogenome was assembled by Tang et al. (2014), was added into each bulk community DNA with a DNA concentration of 1% of the total DNA. Each bulk DNA sample was then used for construction of 200bp insert-size library and sequenced at 2-3 GB depth and 100bp PE on two lanes of a HiSeq2000 at BGI-Shenzhen. Filtered data were aligned onto the 12 previously constructed reference

mitogenomes by BWA and reads that uniquely mapped onto the references with 100% read coverage and at least 99% identity were considered as reads from the focal species.

4.3.11 Statistical analysis

To account for variations in sequencing efficiency, all samples were normalised prior to downstream analysis. The amplicon data were normalised by estimating the proportion of reads (OTU reads) from the total number of reads for each amplicon (target_species_reads/total_community_reads). For the shotgun data, normalization was performed following Tang *et al.* (2015), by mitogenome length (achieved_mitogenome_length / 15000bp) and mito-ratio (MitoNorm), as well as proportion of reads on total reads (pShotgun).

To select the best model explaining the relationship between number of reads and biomass (log transformed), linear and exponential models were explored for each species and sequencing methods. The best model was selected using Akaike information criterion (AIC) (Hu 1987). All statistical analyses, including calculation of model parameters, were performed using the program R (Team 2015).

4.3.12 Community analysis

To visualise community variation resulting for each sequencing treatment for the amplicon data, nonmetric multidimensional scaling was performed (nMDS), using the metaMDS function in the *vegan* package in R (version 3.3.0). Multi-dimensional scaling analysis uses the rank order of species abundances to represent communities in multidimensional space. For this analysis, the Bray-Curtis dissimilarity index was calculated, which is a relative abundance measure. The function “ordispider” in package vegan was used to connect the same communities (resulting from different sequencing treatments) on the ordination plot. The software PRIMER-E v6 (Clarke & Gorley 2006) was also used to examine differences in community composition between sequencing methods (nMDS, Bray-Curtis).

4.4 Results

4.4.1 Amplicon sequencing read results

The total number of amplicon sequencing reads obtained after quality control was 1,430,531, sequenced on a fraction of an Illumina MiSeq lane. More specifically, each amplicon produced the following total number of reads (Mean ± SD), FF130R: 1,004,530 (100,453 ± 87,366), FFFR1: 248,776 (24,878 ± 16,815), B1FR: 177,225 (17,722.5 ± 24,418). Coverage was higher for the 130bp fragment and lower for the two longer fragments (Supplementary Figure S 4.3).

After OTU clustering, the initial number of OTUs obtained exceeded the number of target taxa, which was probably related to the extraction of whole specimens (contaminant OTUs derived from gut contents etc.). For each amplicon, only the following number of OTUs were used in downstream analysis: 49 (FF130R), 20 (FFFR) and 14 (B1FR) after BLAST against our barcode reference database and phylogenetic analysis. Collapsing of multiple OTUs per species was used to account for intraspecific diversity in our data and the observed intraspecific diversity amongst same species OTUs was generally low (Supplementary Table S4.5).

4.4.2 DNA extraction and amplification success

DNA extracted from individual samples for reference genome sequencing was of good concentration but potentially fragmented. DNA extracted from bulk communities was also of good quality with concentration between 49-99 ng/µl, in 4ml elution buffer (ds Qubit) (Supplementary Figure S4.2). Analysis of DNA quality (BGI-Shenzen standard protocols) categorised the quality of DNA samples in category D (based on fragmentation and overall quality of DNA extract) and the quality of samples was deemed appropriate for shotgun sequencing work as in Tang et al. (2014).

4.4.3 Shotgun sequencing results

Twelve out of 13 species were successfully sequenced for their reference mito-genome, while species *A. vortex* was not included in the run due to low quality of extracted DNA. The

remainder species achieved total lengths between 13,627 - 16,159bp, with two species also achieving circular genomes (*N. glauca* and *G. marinus*) (Table 4.4). The average mitochondrial genome length was 14,760bp. The amount of data attributed to mito-reads compared to the total reads per species (mito-ratio) varied largely between species, ranging between 0.011% (*R. balthica*) and 0.664% (*A. aquaticus*), with average mito-ratio at 0.184%. The average sequencing depth was 177.45 (min depth: 6.4 – *G. pulex*, max depth: 670.4 – *E. danica*). See Table 4.4 for detailed information on individual species reference mito-genomes. Shotgun sequencing of the bulk invertebrate samples (mock communities) returned an average number of reads of (±SD) 23,984,200 (±2,248,209.861) per community, and 25,823,450,400 reads overall (Supplementary Figure S4.4).

Table 4.4: **Reference mito-genome sequencing summary results.**
All species achieved assembly of 13 Protein Coding Genes (PCG). Species with (*) achieved circular genomes.

Number	Species	Scaffold number	Total length	Average depth	Mito-ratio (%)
1	*Bathyomphalus contortus*	1	13627	65.5	0.472
2	*Planorbis*	3	13607	30.5	0.033
3	*Bithynia leachi*	1	15624	39.2	0.029
4	*Bithynia tentaculata*	2	15691	36.5	0.049
5	*Physa fontinalis*	1	13792	56	0.626
6	*Potamopyrgus antipodarum*	1	15504	43.3	0.069
7	*Radix balthica*	1	14483	50.4	0.011
8	*Notonecta glauca* *	1	15152	453.1	0.059
9	*Asellus aquaticus*	1	14808	92.8	0.664
10	*Gammarus pulex*	7	13326	6.4	0.015
11	*Ephemera danica*	1	15351	670.4	0.080
12	*Gyrinus marinus* *	1	16159	585.6	0.098

4.4.4 Positive controls

For the two positive controls used to assess shotgun sequencing quality, *D. melanogaster* returned an average of 344.2 (± 51.3) reads, and for *M. mineus* an average of 787.3 (± 125.2) (Supplementary Figure S4.5) (Read number for mitochondrial genomes only). The later was significantly correlated with the number of reads achieved per sample ($R^2 = 0.717$, $p = 0.002$),

while no significant relationship was found for the *D. melanogaster* read number vs. total number. For amplicon sequencing, only the *D. melanogaster* positive control was used. Significant relationships between the positive control sample and the total number of reads were found for two of the amplicons (B1FR: $R^2 = 0.939$, p = 0) (FF130R: $R^2 = 0.610$, p = 0.008), but not for the whole COI region amplicon (FFFR1).

4.4.5 Detection rates per species

A number of false negatives and false positives were found. The proportion presence of false negatives is reported here based on number of expected (known) incidences (cases) for each species in the communities. Incidences are calculated normally as 10 per species (10 communities), except for species *P. fontinalis* (9 incidences) and *N. glauca* (6 incidences) [(11sp. x 10) + (1sp. x 6) + (1sp. x 9) = 125 total incidences/cases] (Table 4.2).

The shotgun approach failed to detect the presence of species in the bulk samples in 7 out of 125 cases (5.6%), for 5 species. Generally, the false negatives with this method occurred only for the lowest and second lowest amount of biomass present for the species in question. For the amplicons, false negatives occurred in 7 cases (5.2% in 5 species) for B1FR, 6 cases (4.5% in 3 species) for FF130R, and 3 cases (2.2% in 3 species) for FFFR (total across all amplicons was 16 out of 405 cases or 4%). Here false negatives appeared not only for the lowest biomass of species but also when up to 10 (FF130R, FFFR), 13 (FFFR) or 17 (B1FR) specimens were known to be present in that community. Overall, false negatives mostly came from gastropod species except *G. pulex* (2 cases) and *E. danica* (1 case).

False positives were detected for *N. glauca* in two cases for the FFFR amplicon, where 111 and 34,511 reads where found (communities 7 and 9 respectively), in communities where that species was known to be absent from the original bulk pool (Table 4.2). Additionally for this species, a lower number of false positive reads was found (<30 reads, amplicons FF130R & FFFR). One more false positive was detected for species *P. fontinalis* with 1204 reads (community 10, amplicon FFFR). These false positives detected here, could be the result of cross-contamination between communities during sample handling or extraction.

4.4.6 Biomass – number of reads regression analysis

Model investigation suggested that exponential and linear models were appropriate for characterising the number of reads to biomass relationships, the model type generally linked to species across the different sequencing methods (Table 4.5). The relationship of reads with biomass was examined individually for each sequencing treatment (three COI amplicons, sum of amplicon data and shotgun data) and each species (13 species, except for the shotgun data where *A. vortex* was not included, see reference mito-genome sequencing), and plotted with the appropriate best-fit model (Figure 4.4, Supplementary Figures 4.6-4.10). Shotgun sequencing results showed positive and mostly significant relationships (11 out of 12 species with 1 trending towards significance; $p = 0.08$). Comparably, PCR-based methods varied across amplicons with sequencing reads from 5-8 species being significantly correlated with biomass (Table 4.5). All species, presented positive reads - biomass relationships, except *E. danica,* which presented negative relationship for the FFFR amplicon (Supplementary Figure S4.8). Sum of the amplicon data improved the relationship obtained for some of the species, mainly in relation to the FFFR amplicon.

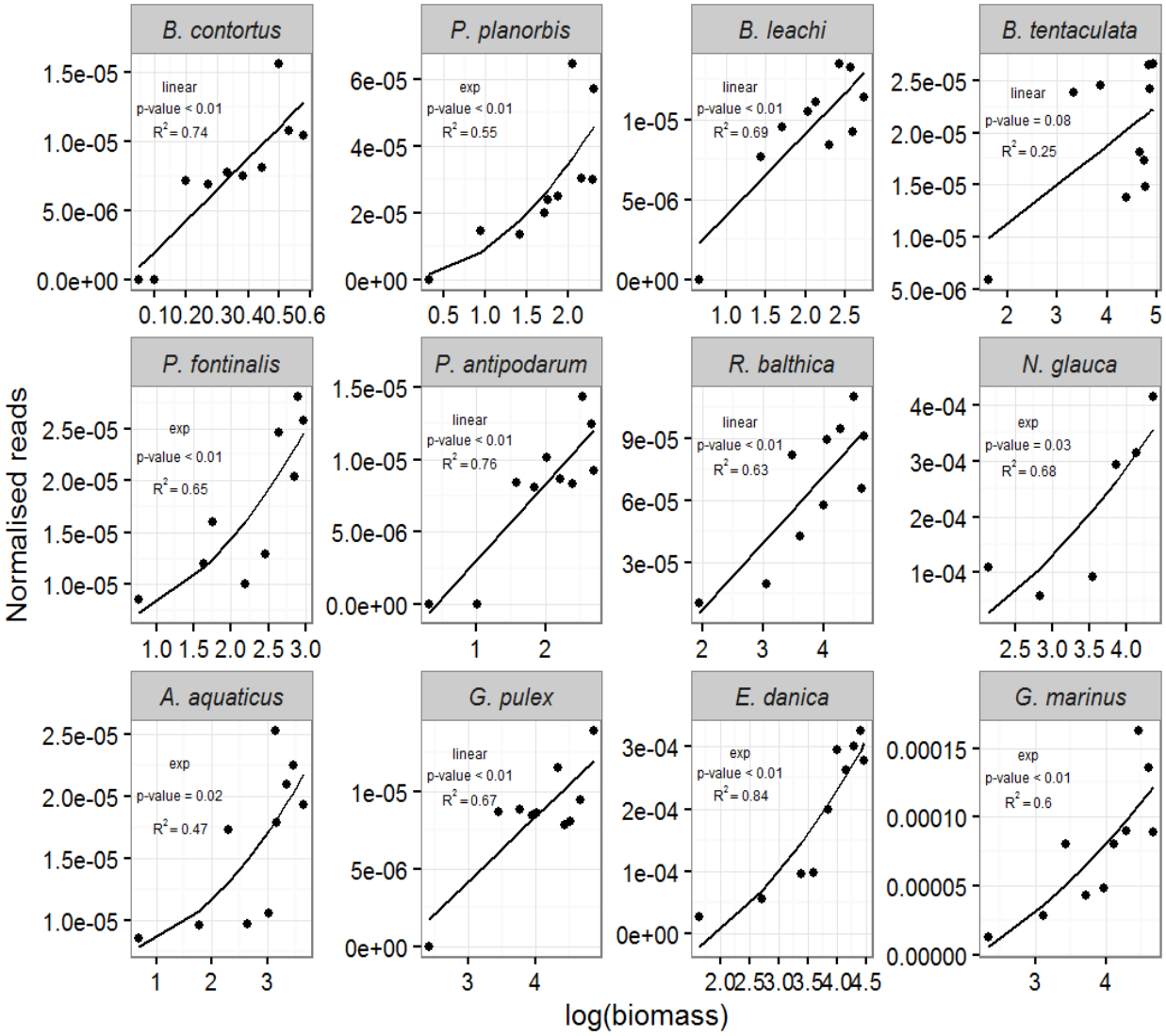


Figure 4.4: Shotgun sequencing regression analysis plots.
Plotted as sequencing reads vs. biomass (x-axis: log Biomass, y-axis: normalized reads). Each box shows data for an individual species, Lines show the fits for each model.

4.4.7 Community analysis results

Comparison between the three COI amplicons on the MDS showed grouping of the same communities along the vertical axis with the exception of communities 9 and 10 (Figure 4.5). This possibly suggests a qualitatively similar community composition in the results obtained by the different amplicons. Simultaneous plotting of amplicon and shotgun data (Figure 4.6a) shows each sequencing treatment separated along the horizontal axis but same axis similarity (same communities) is not as clear for the shotgun data (pink) as in the amplicons. Finally, when the amplicon data were plotted as a sum (SumAmplicon) (Figure 4.6b) against the shotgun reads we could again only observe vertical separation of the groups, although in this case, much clearer than when the individual amplicons were plotted. Moreover, the similarity ranking of communities was almost identical for the two types of sequencing (see order of communities as B, C, G, I etc.).

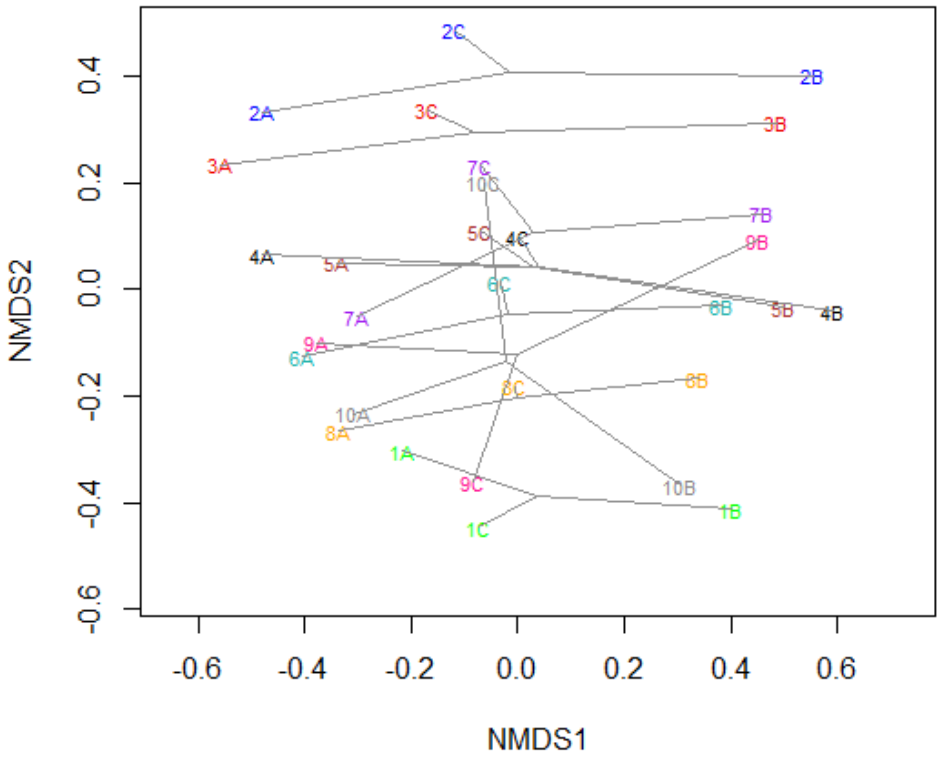


Figure 4.5 nMDS analysis, for amplicon data community composition (Bray-Curtis dissimilarity index). *Samples are coloured according to community (1-10) and named according to amplicon (A: B1FR, B: FF130R, C: FFFR1).*

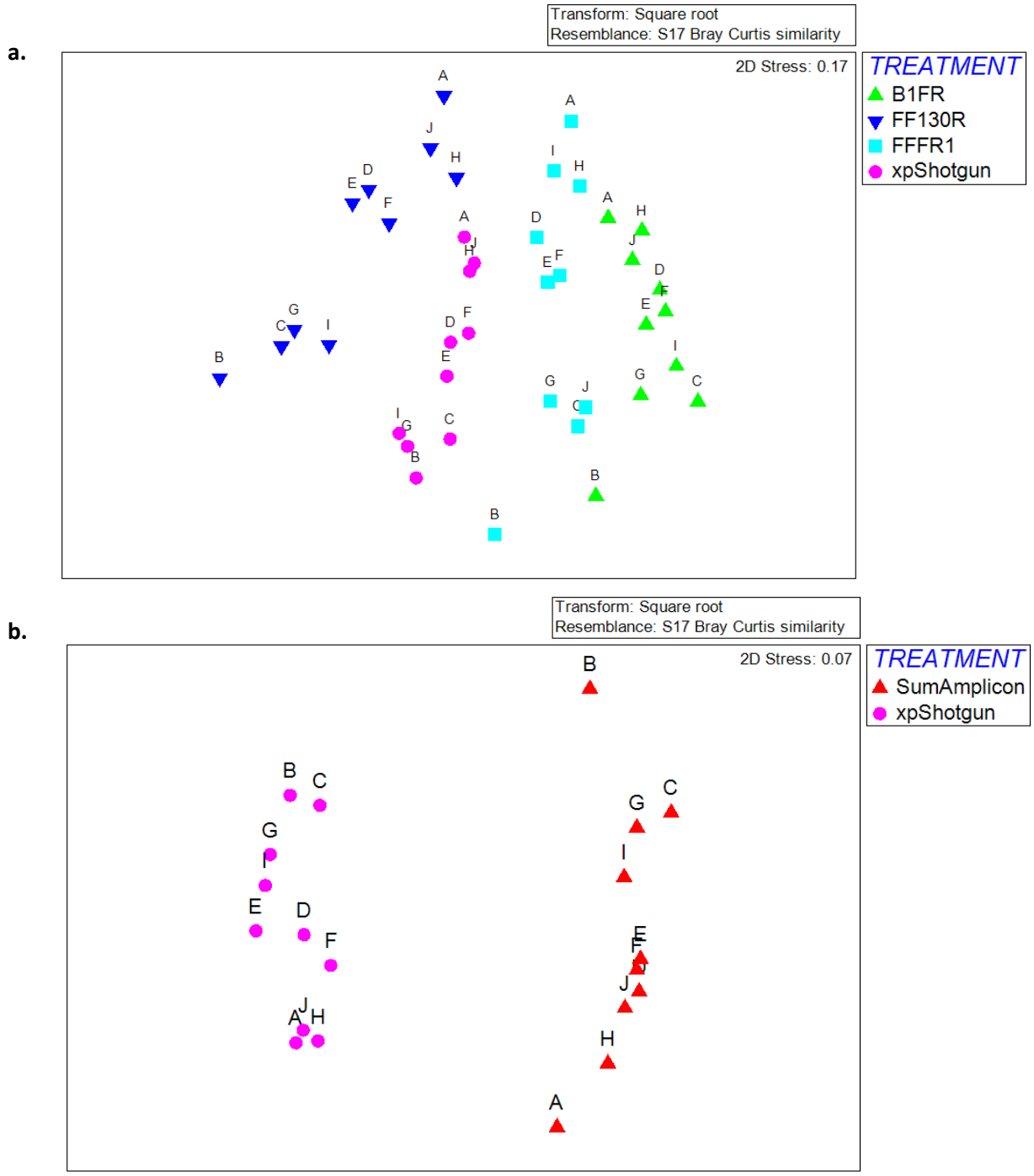


Figure 4.6: nMDS plots for amplicon and shotgun sequencing.
Representation of (**a**.) individual amplicon (B1FR, FF130R, FFFR1) and shotgun (pShotgun) community composition, and (**b**.) summed amplicon data (red) and shotgun data (pink) (Bray-Curtis dissimilarity index).

Table 4.5: Summary table of significance of correlations with biomass for each sequencing treatment.
Amplicon data ("Amplicon", B1FR:450bp, FF130R: 130bp, FFFR: 658bp, SumAmplicon: sum of all amplicon data per species), and shotgun data ("Shotgun", pShotgun: proportion of reads, MitoNorm: mito-ratio normalised). Colours indicate the type of model used (yellow: linear, green: exponential). For the species *A. vortex,* shotgun data were not available (NA). For species *E. danica* - amplicon FFFR, a negative reads- biomass correlation was found (-).

Number	Taxa		Amplicon				Shotgun	
	Family	**Species**	**B1FR**	**FF130R**	**FFFR**	**SumAmpl**	**pShotgun**	**MitoNorm**
1	Planorbidae	*Anisus vortex*	**0.01***	**<0.01***	0.07	**0.02***	NA	NA
2	Planorbidae	*Bathyomphalus contortus*	**<0.01***	0.06	0.07	**0.02***	**<0.01***	**<0.1***
3	Planorbidae	*Planorbis planorbis*	**0.03***	0.06	**0.1***	0.07	**0.01***	**0.01***
4	Bithyniidae	*Bithynia leachi*	0.11	**0.01***	**0.04***	**0.03***	**<0.01***	**<0.01***
5	Bithyniidae	*Bithynia tentaculata*	0.58	0.27	0.37	0.46	0.08	0.09
6	Physidae	*Physa fontinalis*	0.25	**<0.01***	0.57	**0.04***	**0.01***	**0.01***
7	Hydrobiidae	*Potamopyrgus antipodarum*	0.06	**<0.01***	**0.02***	**0.02***	**<0.01***	**<0.01***
8	Lymnaeidae	*Radix balthica*	**0.03***	**0.01***	**0.03***	**0.01***	**<0.01***	**0.01***
9	Notonectidae	*Notonecta glauca*	**0.05***	**0.02***	0.56	0.14	**0.03***	**0.04***
10	Asellidae	*Asellus aquaticus*	**0.05***	0.06	0.15	0.06	**0.02***	**0.03***
11	Gammaridae	*Gammarus pulex*	0.52	0.06	0.32	0.46	**<0.01***	**<0.01***
12	Ephemeridae	*Ephemera danica*	**0.04***	**0.02***	-0.06	**0.02***	**<0.01***	**<0.01***
13	Gyrinidae	*Gyrinus marinus*	**<0.01***	**<0.01***	**<0.01***	**<0.01***	**<0.01***	**0.01***

4.5 Discussion

Here we applied two of the currently most pronounced HTS approaches (metabarcoding vs. shotgun mito-metagenomics) to characterize diversity and species abundance in bulk invertebrate samples and evaluated their performance in accurately estimating biomass and relative abundance content, through the analysis of a structured design of mock communities. Our results confirm that using shotgun mito-metagenomic sequencing provides a more accurate representation of reads to biomass relationships from bulk macroinvertebrate samples, compared to amplicon metabarcoding of the COI gene. Amplicon data did not provide accurate quantitative information on the biomass composition of samples for a large proportion of the species when single amplicon data were analysed and the accuracy of the method slightly improved when results from all three amplicons were combined. Furthermore, cases of rare taxa proved challenging for both methods, which failed to detect low abundance species in several cases, while metabarcoding also misrepresented higher abundance species as well.

4.5.1 Sequencing performance and sample coverage (both methods)

For our reference mito-genome assembly, the depth of sequencing varied between 30 and 670X coverage (**Table 4.4**), with the exception of species *G. pulex*, which achieved the lowest coverage at 6.4 X, but was still assembled to 13,326bp length (using multiple contigs). The length of mito-genome of a congener species to *G. pulex* (*Gammarus duebeni*) has been found to be up to 15,651bp (Krebes & Bastrop 2012). Zhou et al. (2013) report a 10X coverage as sufficient for shotgun mito-genome assembly. For assembling reference mito-genomes in the present work, existing barcode sequences were used as “baits” for mapping, which also allowed lower sequencing coverage to be sufficient compared to *de novo* assembly (read based approach) (Crampton-Platt *et al.* 2016). Generally, the depth required for genome assembly depends on the assembly strategy used and the presence of reference genomes or barcodes; as a rule of thumb assembly based on reference genomes requires much lower sequencing depth than *de novo*, while using short barcoding reads requires intermediate depth (Crampton-Platt *et al.* 2016).

For the metabarcoding work, sequencing coverage varied per amplicon, with the shorter amplicon resulting in significantly higher number of reads than the other two amplicons (Supplementary Figure S4.3). This variation in the depth of sequencing could be attributed to Illumina MiSeq sequencing preferentially amplifying shorter reads when sequenced in a mix or variable efficiency of primer binding. Normalising library contents during sequencing (according to size of molecules included) should therefore be taken into consideration when multiple amplicons are sequenced in the same run.

4.5.2 Reads – biomass relationships

The majority of species presented positive relationships of biomass with the read data, while only one species showed negative relationship (*E. danica*). This reverse trend was found for the FFFR (658bp) amplicon (Supplementary Figure 4.8), which was sequenced using the universal Folmer primers (Folmer *et al.* 1994). Species *E. danica* also failed to amplify during individual barcoding (Table 4.1), suggesting that the results are likely to be related to primer incompatibility.

In many cases, the use of an exponential model was a better descriptor of the relationship between biomass and read number compared to linear models (Figure 4.4, Supplementary Figures 4.6-6.10). This implies that the model used for interpreting the relationship between reads and biomass might affect the final estimations. Never the less, in most cases the number of reads per amplicon increased exponentially with increasing biomass, suggesting a direct biological link between amplicon read number and sequence biomass. Both linear and exponential models have been used for the representation of reads to biomass relationships in published metabarcoding and mito-metagenomics studies (e.g. Zhou *et al.* 2013; Elbrecht & Leese 2015; Tang *et al.* 2015), but other models could be needed to describe such relationships.

4.5.3 False negatives and detection of rare diversity

The percentage of false negative detections for the shotgun work was up to 5.6% (excluding one species from analysis); while for metabarcoding ranged between 2.2%, 4.5% and 5.2%,

for B1FR, FF130R and FFFR amplicon respectively. This suggests that false negative counts are either comparable or somewhat lower for amplicon-based work. Never the less, metabarcoding was also more inconsistent because false negatives were also found for species with higher abundance in the communities (e.g. 10 specimens of *E. danica*, FFFR amplicon). The shotgun method only missed low abundance species, which could be indicative of a need for higher sequencing depth for detection of rare species. For metabarcoding, primer binding related bias could have caused false negatives or abnormal biomass representation, if species were not very compatible with the primer pair used, as was probably the case for species *E. danica* (see individual barcoding results). Additionally, these results could be attributed to inefficient sequencing depth (Supplementary Figure S4.3). This variation in sequencing depth could also influence the quantitative relationships of reads and species abundance (Hajibabaei *et al.* 2011). Increased sequencing depth or use of multiple primers has been previously suggested in order to assist in the detection of species of smaller biomass or smaller relative abundance in the samples through metabarcoding (Hajibabaei *et al.* 2012). The inability to detect rare species could have significant implications for conservation surveys, as is the case for many endangered species (Zhan & MacIsaac 2015).

4.5.4 Reporting on mito-metagenomic work

In Gómez-Rodríguez et al. (2015) mito-metagenomic sequencing was used for characterising 10 natural assemblages of leaf beetles. Comparing the shotgun approach results with and without a prior reference library of the genomes (*de novo*) the authors suggest that using reference sequenced genomes outperforms the *de novo* approach in accuracy and recovery of diversity. Additionally, when a reference mito-genome is available, it is easier to detect and remove Nuclear Mitochondrial pseudogenes (NUMTs) (Bensasson *et al.* 2001) from shotgun sequencing data (Tang *et al.* 2014) (for discussion on the presence of NUMTs in sequencing data, see also Chapter 1). For this experiment, we have used the optimal suggested option for effective mito-genome sequence analysis, as a set of reference mitochondrial genomes were created at the start of the experiment for the species included in the mock communities (Figure 4.1). One exception in this rule was made for the species *A. vortex,* as it was not sequenced for its reference mitogenome due to low quality of the extracted DNA. The

absence of a reference genome made the assembly step more difficult and so this species was not included in downstream analysis of shotgun data.

Normalization of sequencing data is used to account for different DNA concentrations of species, produced by the variability of the number of individuals and body size in the mix (Gillett *et al.* 2014). This variation has been found to influence the quality of the assembly of mitochondrial genomes (Gillett *et al.* 2014). In Tang *et al.* (2015), the shotgun data were normalised based on mitogenome size and mito-ratio. Even though, significant correlations with biomass content were found for non-normalised reads or reads normalised only based on mito-genome size, the combination of both mito-ratio and mitogenome size explained somewhat more variance in their data. For the core analyses, the shotgun reads were normalised according to proportion of reads, and based on mito-ratio, which accounted for the variability of mitochondrial sequencing effort compared to the total amount of sequencing reads. Our investigation of normalization methods showed similar findings between reads normalised according to mito-ratio and proportion of total reads (Table 4.5).

Mito-metagenomic sequencing currently uses a very small fraction of the total sequencing data, since the genomic DNA represents the largest amount of total DNA in the sample. Depending on the taxon, the genomic to mitochondrial DNA ratio (mito-ratio) might vary, but generally approximately 99% of the reads are attributed to genomic DNA, leaving only 0.5-1% of the data to be used (for insects the mito-ratio is 0.5%). Attempts to generalise the expected genomic to mitochondrial DNA ratio are difficult as further work on a wider variety of taxa is necessary (Crampton-Platt *et al.* 2016).

In order to enhance the contribution of mitochondrial DNA during mito-metagenomic sequencing, Zhou et al. (2013) used mitochondrial enrichment via centrifugation, during extraction of invertebrate community samples. In that case, the enrichment process increased the mitochondrial DNA reads, but not largely, with the eventually obtained sequences still only accounting for about 0.5% of the total data (from an initially expected 0.05%). These results suggest that applying enrichment methods still has large room for improvement and other possible routes should be explored. Furthermore, an additional concern while applying enrichment protocols should be to avoid skewing of species proportions in the bulk samples, which could lead to introduction of error in biomass and

relative abundance estimations. In order to avoid any skewing of the species relative abundance ratios, no enrichment processing was applied to our samples.

An alternative method for increasing the mitochondrial contribution in shotgun sequencing of bulk samples was proposed more recently by Liu et al. (2016). This study tested the use of an oligonucleotide capture array designed based on 379 mitochondrial genomes, as a more effective and precise mitochondrial enrichment method. This approach was reported to increase the mitochondrial ratio by 100 fold compared to previous attempts (mitochondrial reads accounted for up to 42% of the sequencing data). Moreover, the use of a capture array was reported to generally maintain the original ratio of species biomass in the sample, with a few variations depending on the phylogenetic distance of the test sample species composition, compared to the species used for designing the array. Microarrays use hybridization of specific nucleotide probes to bind DNA from target species and they are commonly used in gene expression studies, though their use has also been previously suggested for biodiversity monitoring (Hajibabaei *et al.* 2007). The accuracy of the microarray method could nevertheless be limited by the availability of sequencing information for the target organisms used for designing the probes (Hajibabaei *et al.* 2007). Further testing of array work could be very beneficial providing several advantages for future applications, such as decrease in operational costs, by reducing the overall sequencing volume required (Liu *et al.* 2016).

4.5.5 Reporting on metabarcoding work

Metabarcoding has been mainly used for the recovery of species richness from community samples uncovering in many cases extensive diversity, which would have been difficult to achieve using traditional methods (Leray & Knowlton 2015; Sinniger *et al.* 2016). Additionally, metabarcoding work is increasingly used for ecosystem monitoring, where except for richness counts, accurate estimations of abundance contents of environmental samples are also required (Ji *et al.* 2013; Shokralla *et al.* 2015). It has been suggested that sequencing read abundance could be used as a proxy of relative mass composition of species, where higher proportion of species biomass would reflect higher proportion of sequencing reads (Thomas *et al.* 2016), but this assumption has been questioned (Tang *et al.* 2015). Our results only

partially support this statement but mainly reflect on the larger uncertainty of assumptions on relative abundance of species as they are generated by metabarcoding pipelines. More specifically, the metabarcoding work failed to detect significant relationships between read data and known biomass in our samples in many cases (Table 4.6). The FFFR amplicon data (universal Folmer primers) showed significant read-biomass relationships in only 5 out of 13 species, compared to 8 out of 13 for the other two amplicons. This discrepancy in efficiency between amplicons could be related to primer specificity or sequencing depth. First, because the B1FR and FF130R primers were designed and modified for macroinvertebrate taxa and second because the sequencing coverage achieved for the Folmer region (FFFR) was significantly lower than for the other two amplicons (Supplementary Figure S4.3). Summing of sequencing results from all three amplicons slightly improved the reads/biomass relationships (Table 4.5). Multi-dimensional scaling analysis (nMDS, Bray-Curtis index) revealed similarities in community composition based on the sequencing results for individual amplicons (Figure 4.5). This implies that despite the variations in reads-abundance relationships found in the metabarcoding data for individual species, the community profiles obtained were still comparable, with some exceptions (communities 9-10, Figure 4.5). When assessed against shotgun data, similar patterns were found across treatments (individual amplicons) (Figure 4.6a), but the shotgun data are more condensed across the y-axis (Figures 4.6a-b).

The use of COI as the optimal marker for metabarcoding work has also been questioned on occasion (Deagle *et al.* 2014). Problems could arise if the necessary taxonomic resolution is not available with the COI for the studied taxa. To counteract limitations of the currently most widely used Folmer primers, alternative primers have been designed. Examples of such primers are the so called “Mini-barcodes” (Meusnier *et al.* 2008), and another more recently designed set, covering about 300bp within the COI barcoding region, which appear more successful in recovering a broad range of diversity (Leray *et al.* 2013) and could provide a more viable alternative to the Folmer primers. Furthermore, alternative markers are proposed for use in characterisation of biodiversity through metabarcoding, such as 18S (Zhan *et al.* 2014), or 16S (Epp *et al.* 2012), though the COI still retains its superior value compared to other markers due to the large repositories of reference sequences already available.

Using multiple amplicons to increase accuracy of biodiversity detection in community samples has been discussed by Gibson et al. (2014). In that study a set of 11 primer pairs targeting the COI barcoding region were used and it was shown that combinations of several primers significantly increased the levels of species detection in samples of known content. Our results partially support the idea that the combination of sequencing reads from multiple amplicons can increase the accuracy of metabarcoding, as improvement of the results varied between the different species. (Table 4.5). The results from other multi-marker studies, promoting the simultaneous use of multiple markers or loci as more efficient in biodiversity assessment than single amplicon metabarcoding, are very promising (Dupuis *et al.* 2012; Zhan *et al.* 2014), but the success of these approaches could be influenced by the specific species analysed and the primer pairs used. Furthermore, we should also keep in mind that the use of multiple amplicons or loci also creates additional costs for tagged primers and library preparation as well as handling and data analysis time (Creer *et al.* 2016).

4.5.6 Application on closely related species

Two congener species were used in this study (*B. leachii, B. tentaculata*), which allowed evaluation of the methods' performance when closely related species co-occur in bulk samples. During BLAST identification of the OTUs for these two species, *B. leachii* OTUs were incorrectly identified as *B. tentaculata*, due to the presence of a misidentified sequence in our database (see also results from Chapter 2, Figure 2.5). Phylogenetic analysis revealed the correct annotation of the sequences, but this incident reminds us of the shortcomings of this approach related to incomplete databases or the presence of misidentified sequences us pointed out by Deagle et al. (2014).

The shotgun approach was more successful in differentiating between the two congener species. Annotation of the mito-genomes for the two closely related species was performed by mapping onto the previously generated reference mito-genomes, providing more confidence in the results. In Tang et al. (2014), they also successfully assembled three congeneric species of *Drosophila*, first demonstrating the potential for pooling closely related species, while Tang et al. (2015) further improved the pipeline by pooling and assembling the mitogenomes of 48 species of bees.

4.5.7 Shifting to a mito-genomic multi loci approach - future perspectives

Applications of metagenomic sequencing can be used for biodiversity assessment with multiple possible advantages. Using mito-metagenomics can allow characterisation of multiple species simultaneously, while also allowing spatial replication, since samples from multiple locations can be multiplexed (in comparison to traditional methods). Additionally, acquiring long mitochondrial contigs (called “super barcodes”) could provide better phylogenetic resolution and measurement of intraspecific diversity at a more effective rate than what single COI barcodes could achieve. Shifting towards multi loci approaches will be beneficial for increasing taxonomic resolution and reducing effects of false negatives caused by random drop out of genes due to degradation or insufficient sequencing and multi-loci mito-metagenomics could represent the next phase of currently applied metabarcoding approaches (Tang *et al.* 2014). Furthermore, multi-loci advocates suggest that combinations of multiple markers increases delimitation success for closely related species compared to single marker work (Dupuis *et al.* 2012). Overall, the metagenomic approach could present more effective and accurate detection of biomass and abundance in mixed samples, compared to the more widely used to date COI metabarcoding (Crampton-Platt *et al.* 2016), while multiplexing is meant to reduce analytical cost compared to construction of individual libraries for amplicon sequencing.

4.6 Supplementary Information

Supplementary Table S4.1: Detailed description of equations used for calculation of specimen biomass based on measured body dimensions. For body dimensions see, SW: shell width, SH: shell height, BL: body length. For other measurements, n: number of specimens measured, range: variance of body dimensions of available specimens, avg: average body dimension. In equation: DM, W: dry mass, L: dimension according to species.

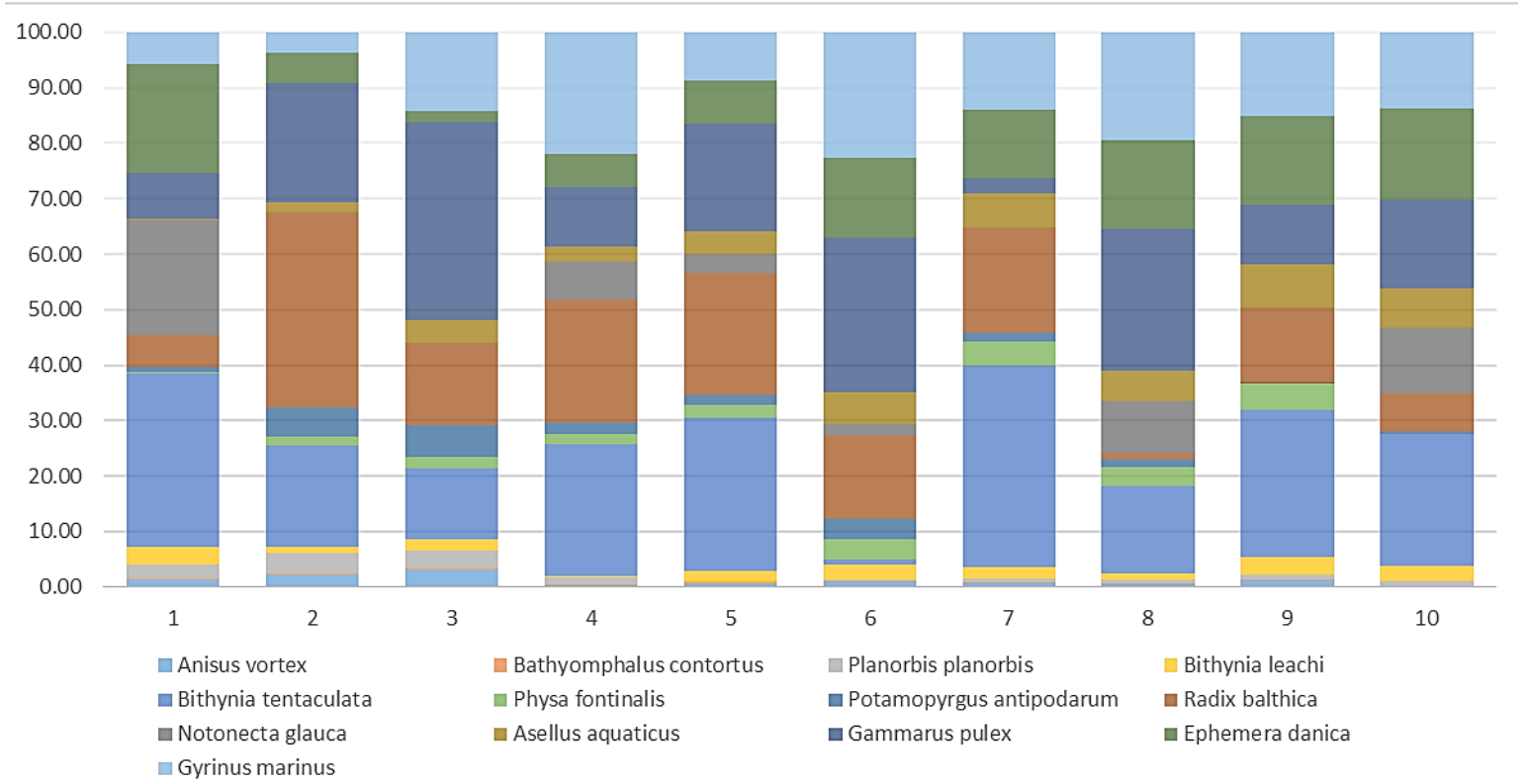
Species	Approximation	Publication	Dimension	n	intercept ± SE	slope ± SE	r2	Range	Average	Equation
Anisus vortex	*Anisus rotundatus*	Caquet 1993	SW	226	2.53	-10.4	0.91	24.61-51.22	45.11	ln W = 2.53 ln L + (-10.4)
Bathyomphalus contortus	*Anisus rotundatus*	Caquet 1993	SW	80	2.53	-10.4	0.91	22.8-41.04	30.65	ln W = 2.53 ln L + (-10.4)
Planorbis planorbis	*Anisus rotundatus*	Caquet 1993	SW	144	2.53	-10.4	0.91	4.56-11.63	6.47	ln W = 2.53 ln L + (-10.4)
Bithynia leachii	*Bithynia tentaculata*	Baumgartner 2003	SH	83	0.010673407	3.23±0.25	0.96	1.62 - 5	3.9	$DM = 0.01067 . L^{3.23}$
Bithynia tentaculata	*Bithynia tentaculata*	Baumgartner 2003	SH	180	0.010673407	3.23±0.25	0.96	2.307-9.57	6.13	$DM = 0.01067 . L^{3.23}$
Physa fontinalis	*Physa fontinalis*	Caquet 1993	SH	70	3.07	-11.4	0.88	26.9 - 53.8	44.41	ln W = 3.07 ln L + (-11.4)
Potamopyrgus antipodarum	*P. antipodarum*	Mahrlein 2015	SH	186	0.0251	2.07±0.06	0.94	2.3-4.7	3.78	$DM = 0.0251 . L^{2.07}$
Radix balthica	*Radix peregra*	Baumgartner 2003	SH	94	0.008565609	3.19	0.94	3.75-10.74	7.6	$DM = 0.008566 . L^{3.19}$
Asellus aquaticus	*Asellus aquaticus*	Baumgartner 2003	BL	142	0.002029431	3.75	0.69	2.92-7.53	5.35	$DM = 0.0020294 . L^{3.75}$
Gyrinus marinus	Gyrinidae	Benke 1999	BL	55	0.0531±0.0031	2.586±0.210	0.67	6.91-8.55	7.7	$DM= 0.0531 . L^{2.586}$
Gammarus pulex	*Gammarus minus*	Benke 1999	BL	53	0.012	2.74	0.95	6.9-18.5	12.2	$DM = 0.012 . L^{2.74}$
Ephemera danica	Ephemera sp.	Benke 1999	BL	108	0.0021±0.0003	2.737±0.079	0.99	12.19-20.57	16.33	$DM = 0.0021 . L^{2.737}$
Notonecta glauca	Hemiptera	Benke 1999	BL	31	0.0031±0.0002	2.904±0.157	0.81	14.2-15.75	14.85	$DM = 0.0031 . L^{2.904}$

Supplementary Table S4.2: Biomass estimates for each species included in the mock communities after conversion using published regressions.
Values are presented in milligrams (mg).

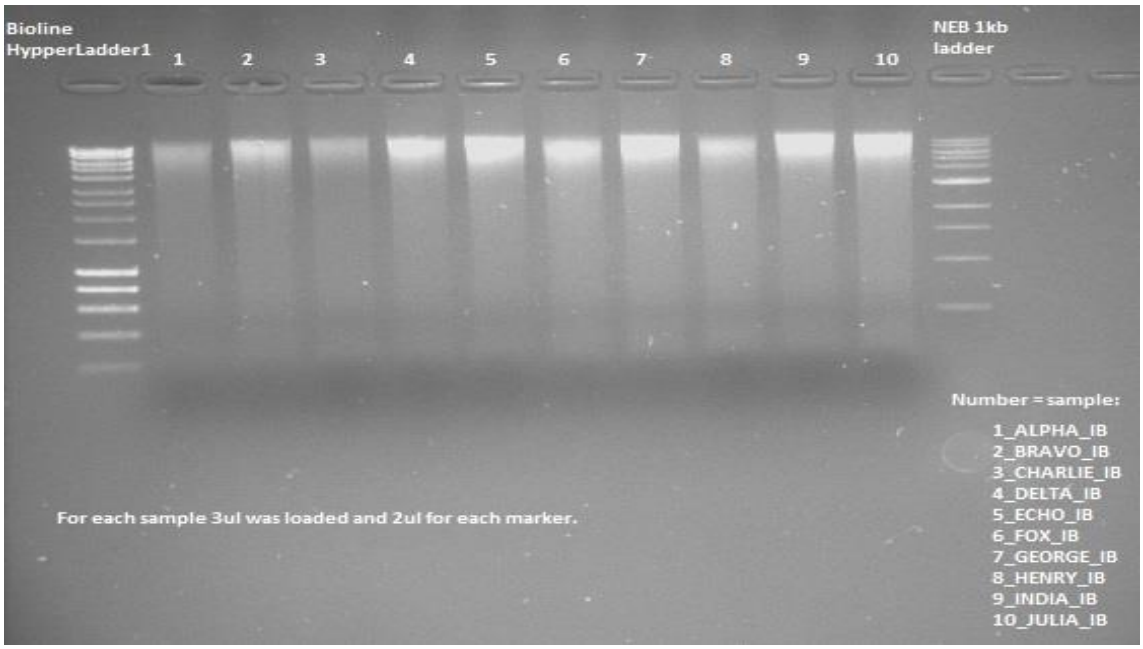
	Community									
Species	**1 Alpha**	**2 Bravo**	**3 Charlie**	**4 Delta**	**5 Echo**	**6 Fox**	**7 George**	**8 Henry**	**9 India**	**10 Julia**
Anisus vortex	5.08	5.75	6.44	0.70	3.41	2.84	2.10	1.42	4.21	0.13
Bathyomphalus contortus	0.79	0.70	0.65	0.56	0.47	0.40	0.31	0.05	0.10	0.22
Planorbis planorbis	8.87	9.04	6.75	7.66	0.37	1.56	3.10	4.53	4.77	5.55
Bithynia leachi	11.97	3.19	4.52	0.94	8.95	10.30	7.45	6.61	12.50	14.54
Bithynia tentaculata	115.60	46.33	26.58	113.80	128.77	4.01	135.45	79.47	104.71	124.64
Physa fontinalis	1.16	4.14	4.78	7.91	10.81	13.10	16.42	17.02	18.70	**0.00**
Potamopyrgus antipodarum	3.82	13.37	11.66	9.89	8.10	13.87	5.27	6.46	0.36	1.74
Radix balthica	20.47	89.72	31.37	106.12	101.69	57.27	70.67	6.08	53.80	35.91
Notonecta glauca	77.38	**0.00**	**0.00**	33.88	16.09	7.47	**0.00**	46.54	**0.00**	62.09
Asellus aquaticus	1.00	4.92	9.00	13.02	19.50	22.01	22.68	27.56	30.73	37.72
Gammarus pulex	30.49	54.70	75.11	51.18	90.38	104.72	10.13	128.80	42.39	83.34
Ephemera danica	71.92	14.24	4.18	28.62	35.93	54.17	46.02	80.35	62.88	86.31
Gyrinus marinus	21.69	9.36	30.00	104.94	40.35	85.77	52.20	98.29	60.07	72.01
Total (mg)	**370.25**	**255.46**	**211.02**	**479.23**	**464.82**	**377.48**	**371.80**	**503.18**	**395.20**	**524.22**

Supplementary Table S4.3: Estimated biomass content per community and species as percentage (%) of the total biomass for each community.
Highlighted in bold the species absent in the community.

	Community									
Species	**1 Alpha**	**2 Bravo**	**3 Charlie**	**4 Delta**	**5 Echo**	**6 Fox**	**7 George**	**8 Henry**	**9 India**	**10 Julia**
Anisus vortex	1.37	2.25	3.05	0.15	0.73	0.75	0.56	0.28	1.06	0.03
Bathyomphalus contortus	0.21	0.28	0.31	0.12	0.10	0.11	0.08	0.01	0.03	0.04
Planorbis planorbis	2.39	3.54	3.20	1.60	0.08	0.41	0.83	0.90	1.21	1.06
Bithynia leachi	3.23	1.25	2.14	0.20	1.93	2.73	2.00	1.31	3.16	2.77
Bithynia tentaculata	31.22	18.13	12.60	23.75	27.70	1.06	36.43	15.79	26.50	23.78
Physa fontinalis	0.31	1.62	2.26	1.65	2.33	3.47	4.42	3.38	4.73	**0.00**
Potamopyrgus antipodarum	1.03	5.23	5.53	2.06	1.74	3.67	1.42	1.28	0.09	0.33
Radix balthica	5.53	35.12	14.87	22.14	21.88	15.17	19.01	1.21	13.61	6.85
Notonecta glauca	20.90	**0.00**	**0.00**	7.07	3.46	1.98	**0.00**	9.25	**0.00**	11.84
Asellus aquaticus	0.27	1.93	4.26	2.72	4.20	5.83	6.10	5.48	7.78	7.19
Gammarus pulex	8.24	21.41	35.59	10.68	19.44	27.74	2.73	25.60	10.73	15.90
Ephemera danica	19.43	5.57	1.98	5.97	7.73	14.35	12.38	15.97	15.91	16.47
Gyrinus marinus	5.86	3.67	14.22	21.90	8.68	22.72	14.04	19.53	15.20	13.74



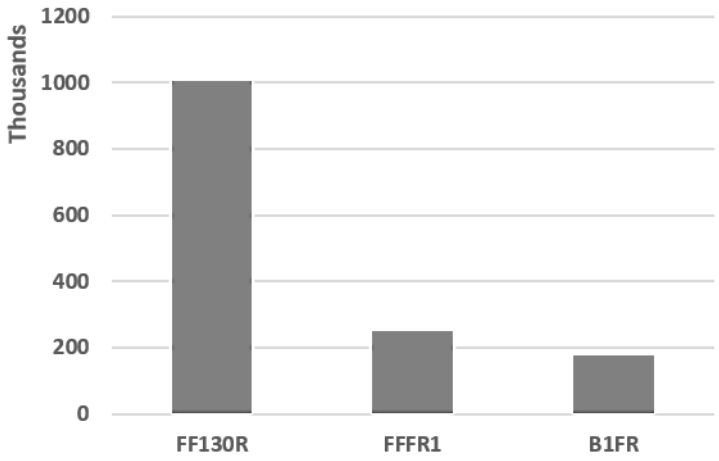
Supplementary Figure S4.1: Graphical representation of estimated percentage biomass composition of mock communities (x-axis: communities 1-10, y-axis: relative abundance %).



Supplementary Figure S4.2: Agarose gel picture of DNA extracts for bulk communities (0.8% agarose gel, 3µl DNA loaded, 2µl Bioline Hypperladder, 2µl NEB 1kb ladder).
Samples are loaded in community order 1-10, see also in-figure index for community codes.

Supplementary Table S4.4: DNA extraction information (bulk communities), including. dsQubit & Nanodrop measurements (Total eluate volume 4ml per community).

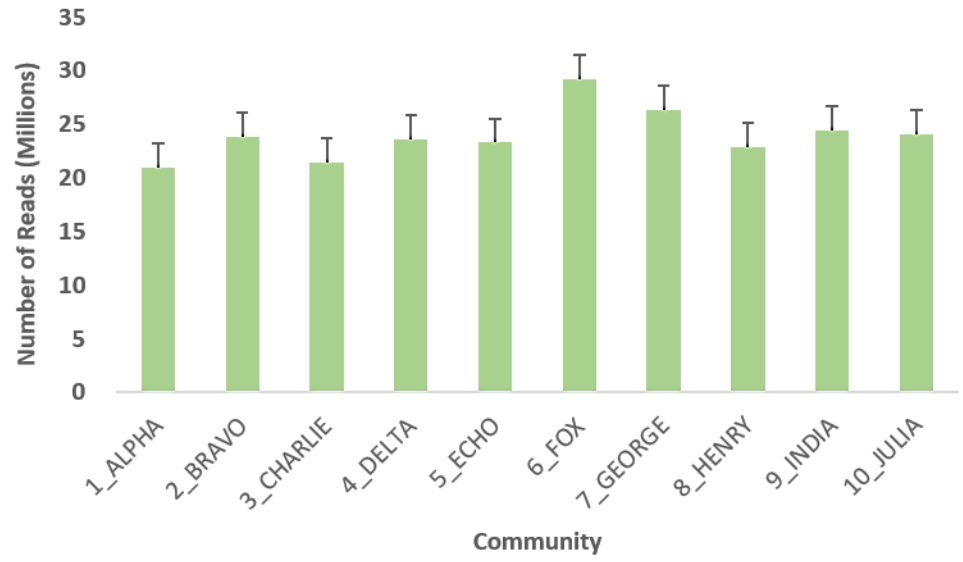
Community	Extraction date	Qubit	Nanodrop	260/280	Number of Species
1_ALPHA_IB	27/02/2015	48.9	54.73	1.79	14
2_BRAVO_IB	02/03/2015	79.8	73.33	1.86	13
3_CHARLIE_IB	02/03/2015	50.2	52.76	1.87	13
4_DELTA_IB	03/03/2015	71.4	78.94	1.86	14
5_ECHO_IB	03/03/2015	98.8	85.19	1.9	14
6_FOX_IB	04/03/2015	80.8	71.99	1.84	14
7_GEORGE_IB	04/03/2015	86.6	79.72	1.86	13
8_HENRY_IB	04/03/2015	59.2	70.55	1.92	14
9_INDIA_IB	05/03/2015	62.4	74.82	1.9	13
10_JULIA_IB	05/03/2015	72.2	81.76	1.89	13



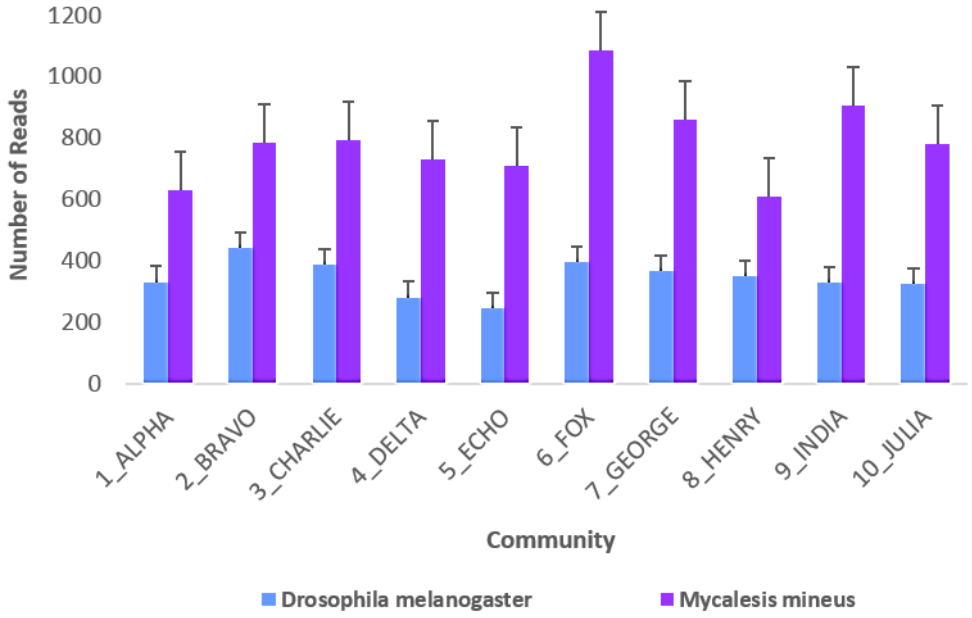
Supplementary Figure S4.3: Total number of generated MiSeq amplicon reads, for each amplicon (x-axis: FF130R, FFFR1, B1FR, y-axis: amplicon reads in thousands).

Supplementary Table S4.5: Within group distance calculation per amplicon.
Only OTUs with >97% BLAST ID were used for distance calculation. (*) the three highest intraspecific distances measured across the three amplicons.

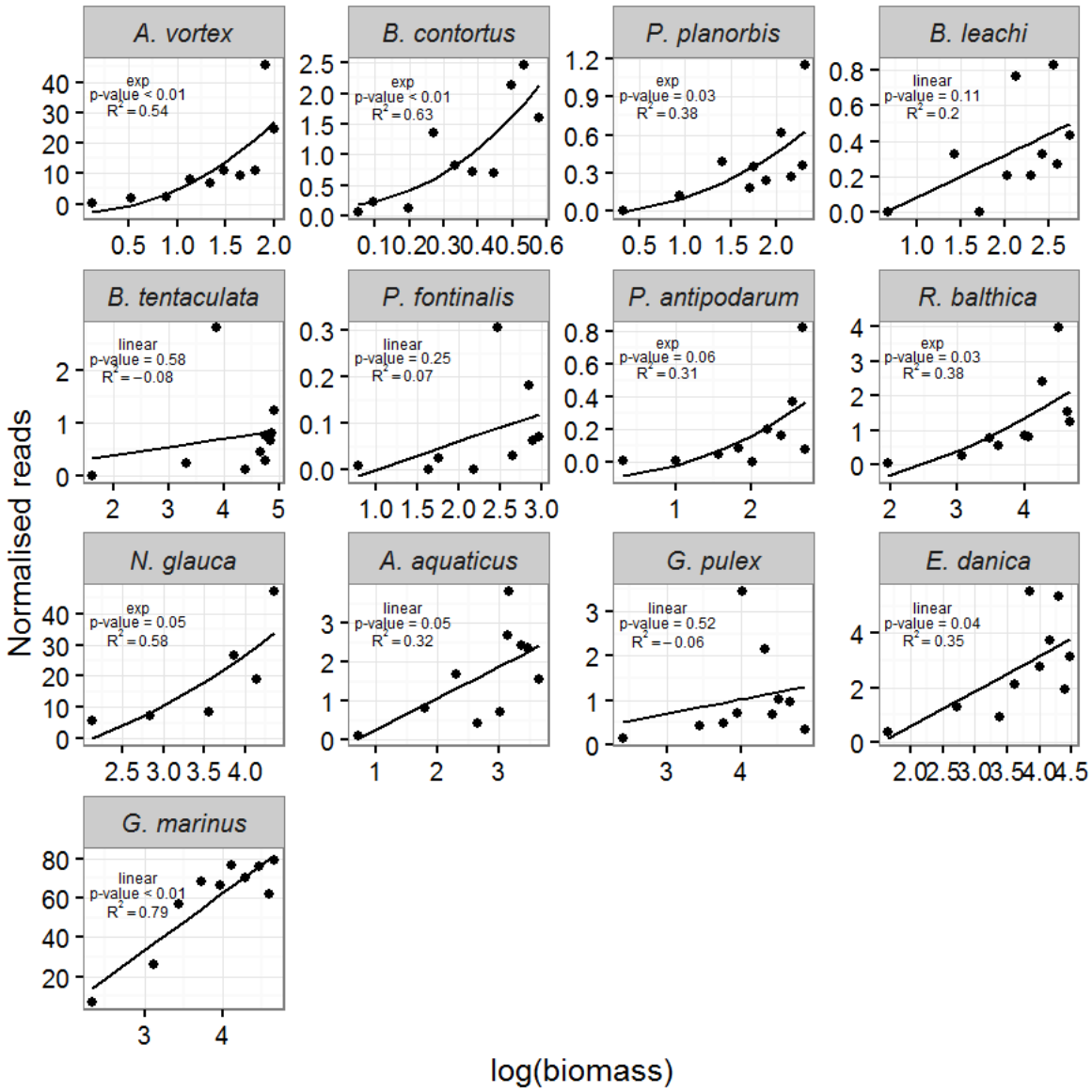
	Within group distance		
Species	**B1FR**	**FF130R**	**FFFR**
Anisus vortex	0.001	0.003	0.003
Bathyomphalus contortus	0.002	0	0
Planorbis planorbis	0	0	0
Bithynia leachi	0	0	0
Bithynia tentaculata	0.015	0.018	0.020
Physa fontinalis	0.001	0.014	0
P. antipodarum	0	0	0
Radix balthica	0.009	0.025*	0.008
Notonecta glauca	0.003	0.037*	0.006
Asellus aquaticus	0.004	0.000	0.006
Gammarus pulex	0	0	0
Ephemera danica	0.010	0.016	0.015
Gyrinus marinus	0	0.034*	0.010
Drosophila melanogaster	0.008	0	0.003
Mean	0.004	0.011	0.005



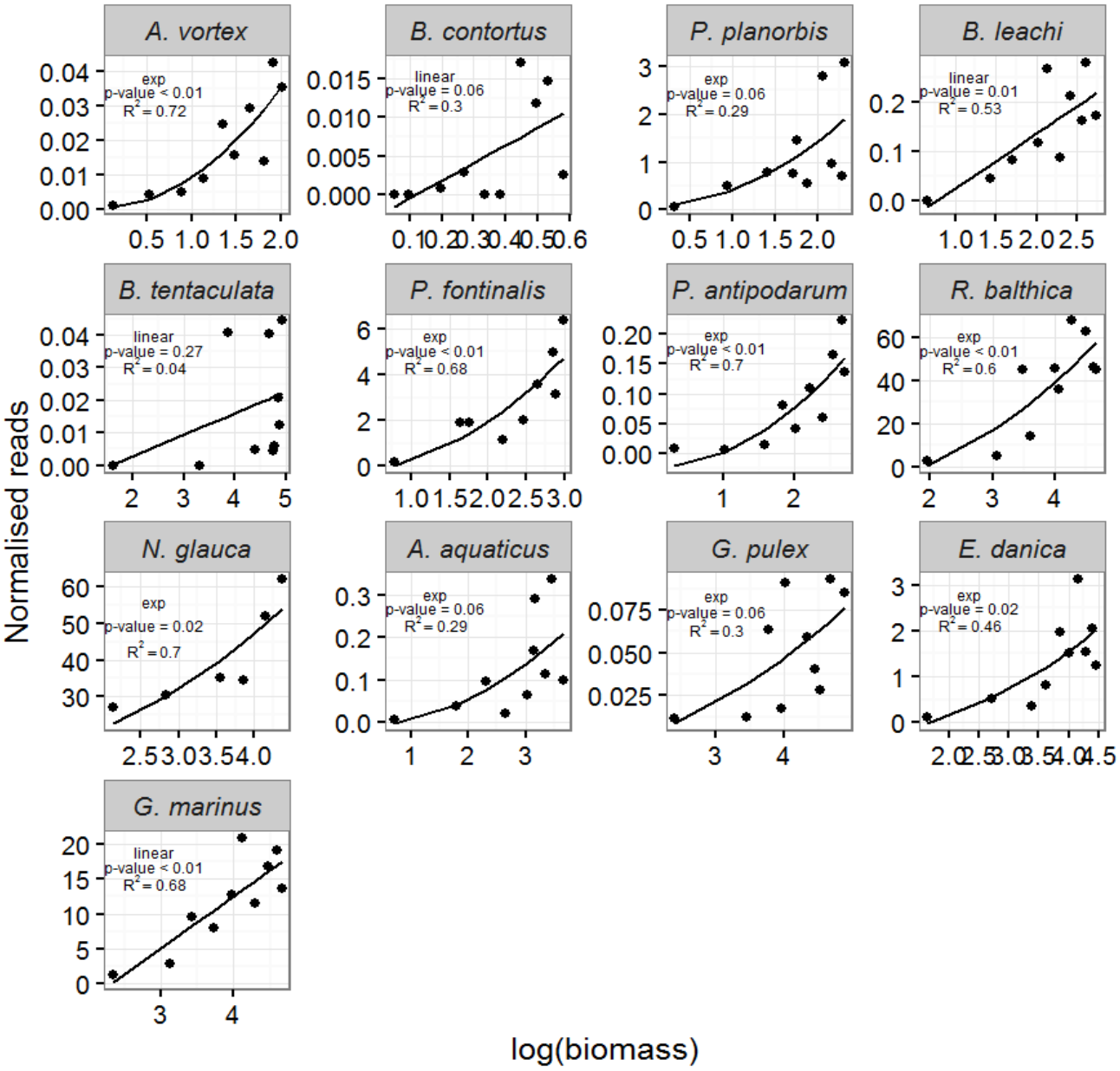
Supplementary Figure S4.4: Number of shotgun reads per bulk sample 1-10 (y-axis in million reads). Error bars represent one Standard Deviation.



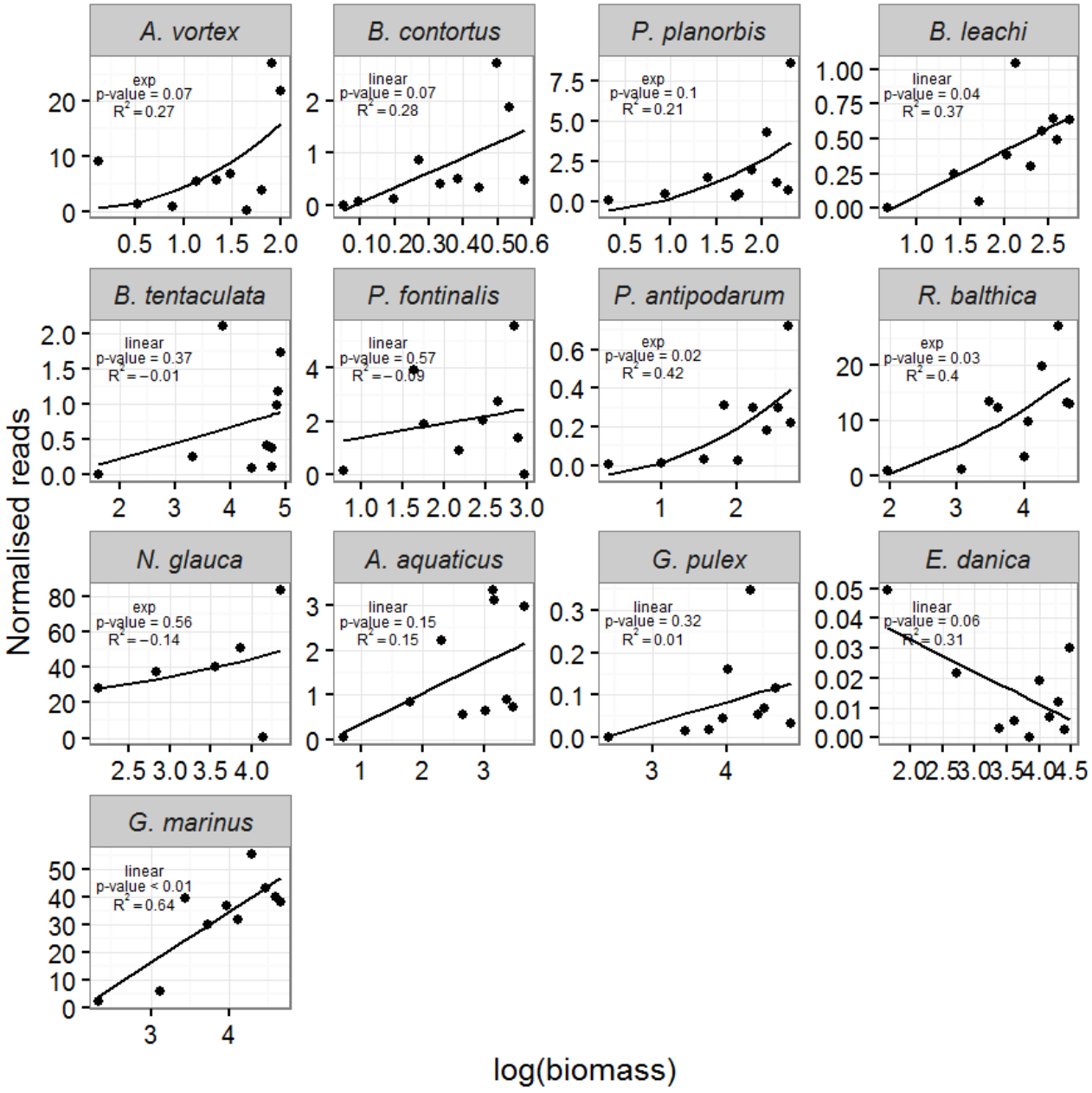
Supplementary Figure S4.5: Number of clean reads for positive control species, derived from shotgun sequencing of bulk communities.
Error bars represent one Standard Deviation. Blue: *D. melanogaster* (whole body positive control), purple: *M. mineus* (DNA extract positive control). Read number for mitochondrial genome sequences only.



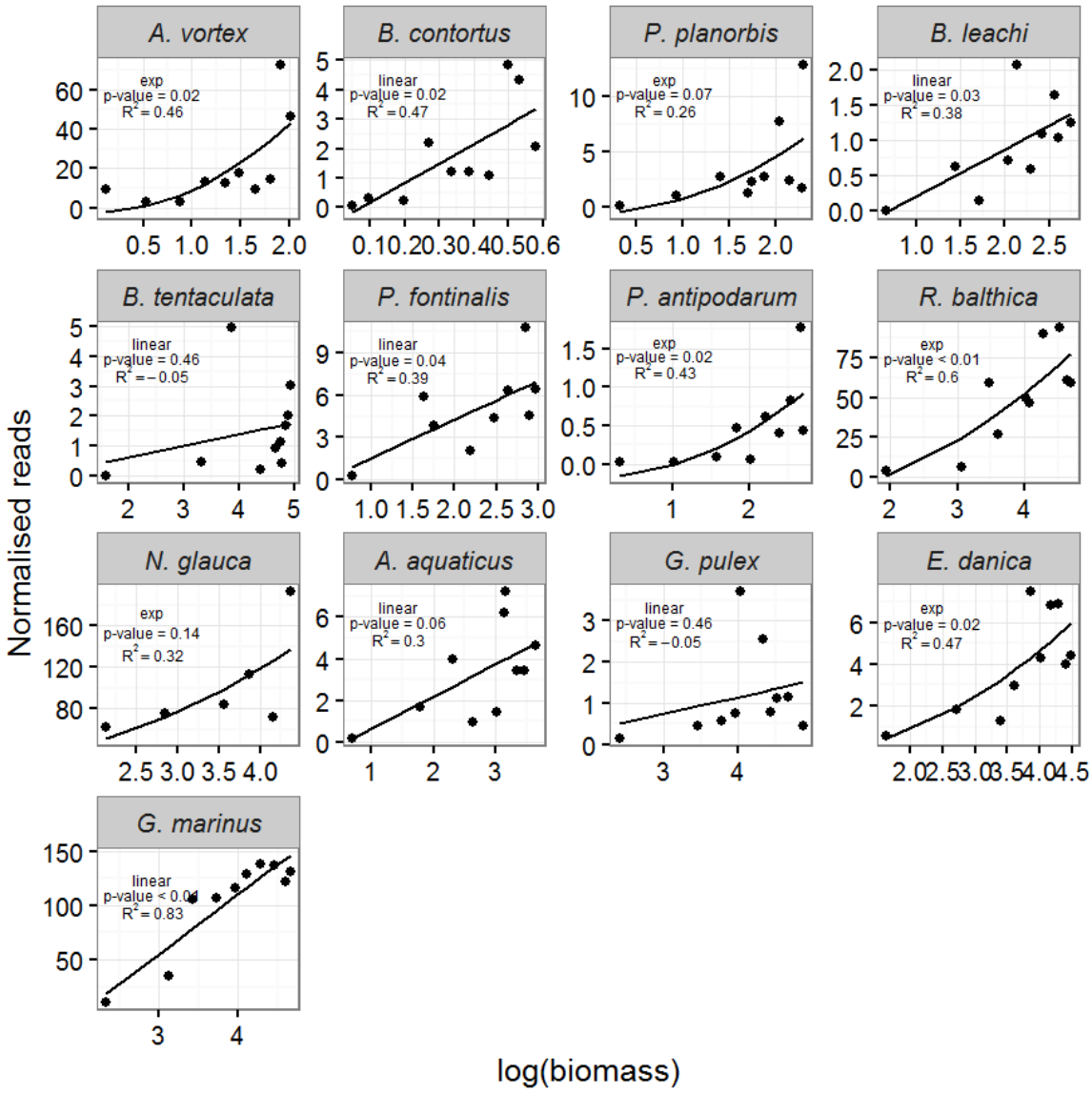
Supplementary Figure S4.6: Amplicon B1FR regression analysis, plotted as sequencing reads vs. biomass (x-axis: log Biomass, y-axis: normalized reads). *Each box shows data for an individual species, Lines show the fits for each model.*



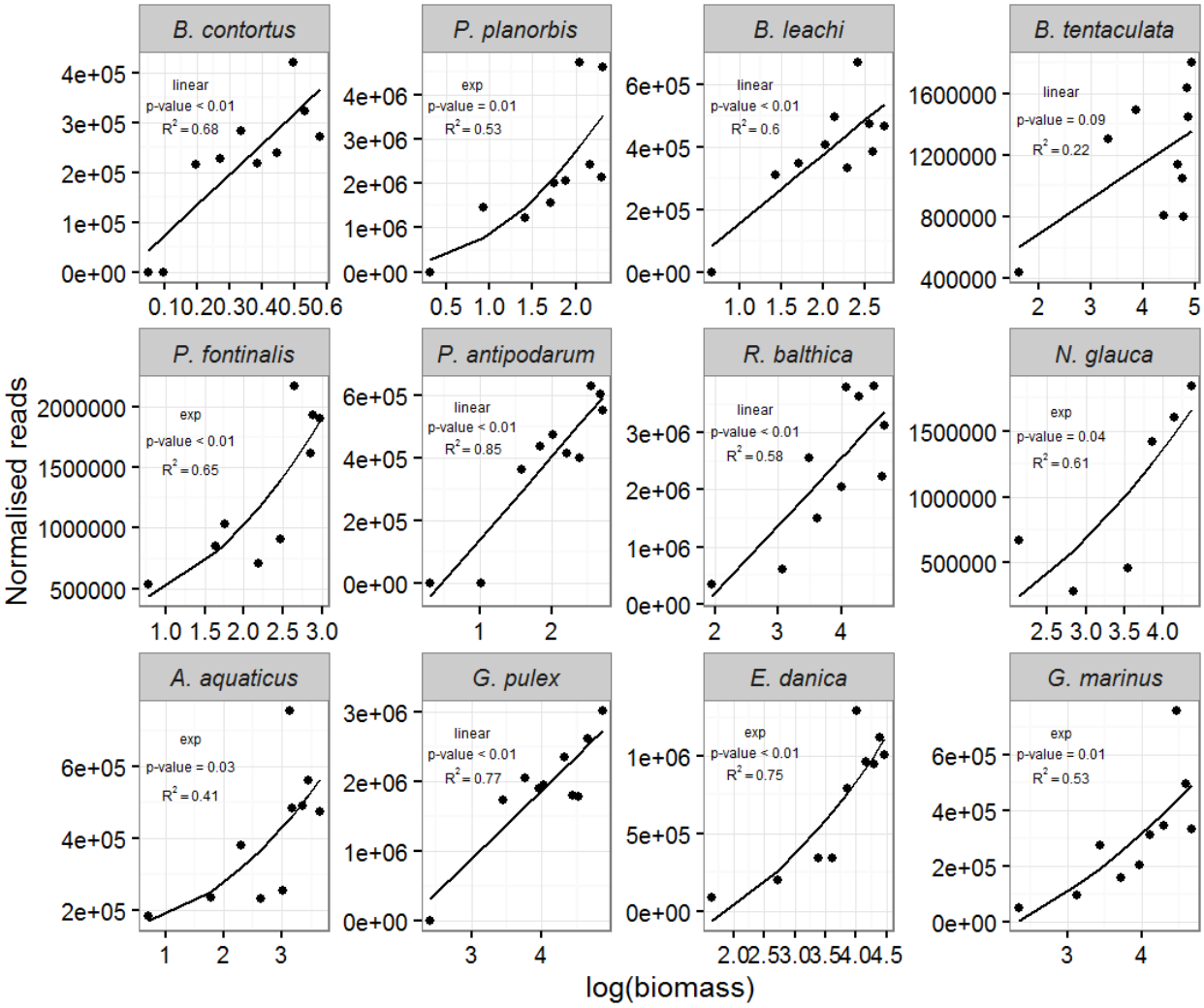
Supplementary Figure S4.7: Amplicon FF130Rregression analysis, plotted as sequencing of reads vs. biomass (x-axis: log Biomass, y-axis: normalized reads). *Each box shows data for an individual species, Lines show the fits for each model.*



Supplementary Figure S4.8: Amplicon FFFR regression analysis, plotted as sequencing of reads vs. biomass (x-axis: log Biomass, y-axis: normalized reads). *Each box shows data for an individual species, Lines show the fits for each model.*



Supplementary Figure S4.9: Sum of metabarcoding reads across amplicons regression analysis, plotted as sequencing reads vs. biomass (x-axis: log Biomass, y-axis: normalized reads). *Each box shows data for an individual species, Lines show the fits for each model.*



Supplementary Figure S4.10: Shotgun regression analysis (mito-ratio normalised data), plotted as proportion of reads vs. biomass (x-axis: log Biomass, y-axis: normalised reads). Each box shows data for an individual species, Lines show the fits for each model.

References

Baird, D.J. & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.

Baumgärtner, D. & Rothhaupt, K.O. (2003). Predictive Length-Dry Mass Regressions for Freshwater Invertebrates in a Pre-Alpine Lake Littoral. *International Review of Hydrobiology*, **88**, 453–463.

Benke, A.C., Huryn, A.D., Smock, L. a & Wallace, J.B. (1999). Length-mass relationships for freshwater macroinvertebrates in North America with particular reference to the southeastern United States. *Journal of the North American Benthological Society*, **18**, 308–343.

Bensasson, D., Zhang, D., Hartl, D.L. & Hewitt, G.M. (2001). Mitochondrial pseudogenes : evolution ' s misplaced witnesses. *TRENDS in Ecology & Evolution*, **16**, 314–321.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009). BLAST plus: architecture and applications. *BMC Bioinformatics*, **10**, 1.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., Mcdonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. & Knight, R. (2010). QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nature methods*, **7**, 335–336.

Caquet, T. (1993). Comparative life-cycle, biomass and secondary production of three sympatric freshwater gastropod species. *Journal of Molluscan Studies*, **59**, 43–50.

Clarke, K.R. & Gorley, R.N. (2006). Primer v6: User Manual/Tutorial. 192.

Clarke, L.J., Soubrier, J., Weyrich, L.S. & Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, **14**, 1160–1170.

Collins, A., Ohandja, D.G., Hoare, D. & Voulvoulis, N. (2012). Implementing the Water Framework Directive: A transition from established monitoring networks in England and Wales. *Environmental Science and Policy*, **17**, 49–61.

Crampton-Platt, A., Yu, D.W., Zhou, X. & Vogler, A.P. (2016). Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience*, **5**, 15.

Cranston, P.S. (1990). Biomonitoring and invertebrate taxonomy. *Environmental Monitoring and Assessment*, **14**, 265–273.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W.K., Potter, C. & Bik, H.M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, **56**, 68–74.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., Taberlet, P., Taberlet, P., Coissac, E., Hajibabaei, M., Rieseberg, L., Yu, D., Ji, Y., Emerson, B., Wang, X., Ye, C., Yang, C., Ding, Z., Ficetola, G., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessiere, J., Taberlet, P., Pompanon, F., Geller, J., Meyer, C., Parker, M., Hawk, H., Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glockner, F., Bru, D., Martin-Laurent, F., Philippot, L., Schloss, P., Gevers, D., Westcott, S., Clarke, L., Soubrier, J., Weyrich, L., Cooper, A., Ji, Y., Barba, M. De, Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., Taberlet, P., Leray, M., Yang, J., Meyer, C., Mills, S., Agudelo, N., Ranwez, V., Boehm, J., Machida, R., Little, D., Deagle, B., Kirkwood, R., Jarman, S., Zhou, X., Shokralla, S., Gibson, J., Nikbakht, H., Janzen, D., Hallwachs, W. & Hajibabaei, M.

(2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology letters*, **10**, 1789–1793.

Dupuis, J.R., Roe, A.D. & Sperling, F.A.H. (2012). Multi-locus species delimitation in closely related animals and fungi: One marker is not enough. *Molecular Ecology*, **21**, 4422–4436.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Elbrecht, V. & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol (M. Hajibabaei, Ed.). *PLoS ONE*, **10**, e0130324.

Epp, L.S., Boessenkool, S., Bellemain, E.P., Haile, J., Esposito, A., Riaz, T., Erséus, C., Gusarov, V.I., Edwards, M.E., Johnsen, A., Stenøien, H.K., Hassel, K., Kauserud, H., Yoccoz, N.G., Bråthen, K.A., Willerslev, E., Taberlet, P., Coissac, E. & Brochmann, C. (2012). New environmental metabarcodes for analysing soil DNA: Potential for studying past and present ecosystems. *Molecular Ecology*, **21**, 1821–1833.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

Gibson, J.F., Shokralla, S., Curry, C., Baird, D.J., Monk, W.A., King, I. & Hajibabaei, M. (2015). Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS ONE*, **10**, 1–15.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 8007–12.

Gillett, C.P.D.T., Crampton-Platt, A., Timmermans, M.J.T.N., Jordal, B.H., Emerson, B.C. & Vogler, A.P. (2014). Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, **31**, 2223–2237.

Goecks, J., Nekrutenko, A. & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, **11**, R86.

Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M.J.T.N., Baselga, A. & Vogler, A.P. (2015). Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, **6**, 883–894.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C. & Baird, D.J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, **6**, e17497.

Hajibabaei, M., Singer, G. a C., Clare, E.L. & Hebert, P.D.N. (2007). Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC biology*, **5**, 24.

Hajibabaei, M., Spall, J.L., Shokralla, S. & van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, **12**, 28.

Hänfling, B., Lawson Handley, L., Read, D.S., Hahn, C., Li, J., Nichols, P., Blackman, R.C., Oliver, A. & Winfield, I.J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-

term data from established survey methods. *Molecular Ecology*, **25**, 3101–3119.

Hiiesalu, I., Pärtel, M., Davison, J., Gerhold, P., Metsis, M., Moora, M., Öpik, M., Vasar, M., Zobel, M. & Wilson, S.D. (2014). Species richness of arbuscular mycorrhizal fungi: Associations with grassland plant richness and biomass. *New Phytologist*, **203**, 233–244.

Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L., Fazekas, A.J., Graham, S.W., James, K.E., Kim, K.-J., Kress, W.J., Schneider, H., van AlphenStahl, J., Barrett, S.C.H., van den Berg, C., Bogarin, D., Burgess, K.S., Cameron, K.M., Carine, M., ChacÃ³n, J., Clark, A., Clarkson, J.J., Conrad, F., Devey, D.S., Ford, C.S., Hedderson, T.A.J., Hollingsworth, M.L., Husband, B.C., Kelly, L.J., Kesanakurti, P.R., Kim, J.S., Kim, Y.-D., Lahaye, R., Lee, H.-L., Long, D.G., MadriÃ±Ã¡n, S., Maurin, O., Meusnier, I., Newmaster, S.G., Park, C.-W., Percy, D.M., Petersen, G., Richardson, J.E., Salazar, G.A., Savolainen, V., Seberg, O., Wilkinson, M.J., Yi, D.-K. & Little, D.P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, **106**.

Hu, S. (1987). *Akaike information criterion statistics*. KTK Scientific Publishers, Tokyo.

Jackson, J.K., Battle, J.M., White, B.P., Pilgrim, E.M., Stein, E.D., Miller, P.E. & Sweeney, B.W. (2014). Cryptic biodiversity in streams: a comparison of macroinvertebrate communities based on morphological and DNA barcode identifications. *Freshwater Science*, **33**, 312–324.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P., Edwards, F.A., Larsen, T.H., Hsu, W.W., Benedick, S., Hamer, K.C., Wilcove, D.S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B.C. & Yu, D.W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Joshi, N. & Fass, J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. *Available at https://github.com/najoshi/sickle.*, 2011.

Kelly, R.P., Port, J.A., Yamahara, K.M. & Crowder, L.B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS ONE*, **9**, e86175.

Krebes, L. & Bastrop, R. (2012). The mitogenome of Gammarus duebeni (Crustacea Amphipoda): A new gene order and non-neutral sequence evolution of tandem repeats in the control region. *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, **7**, 201–211.

Leray, M. & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*, **2014**, 201424997.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. & Machida, R.J. (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**, 34.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., Zhang, H., Misof, B., Kjer, K.M., Tang, M., Niehuis, O., Jiang, H. & Zhou, X. (2016). Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, **16**, 470–479.

Loreau, M. & de Mazancourt, C. (2013). Biodiversity and ecosystem stability: A synthesis of

underlying mechanisms. *Ecology Letters*, **16**, 106–115.

Mächler, E., Deiner, K., Steinmann, P. & Altermatt, F. (2014). Utility of environmental DNA for monitoring rare and indicator macroinvertebrate species. *Freshwater Science*, **33**, 1174–1183.

Mährlein, M., Pätzig, M., Brauns, M. & Dolman, A.M. (2016). Lengt-mass relationships for lake macroinvertebrates corrected for back-transformation and preservation effects. *Hydrobiologia*, **768**, 37–50.

Martin, M. (2011). Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011. Date of access 05/08/2015. *EMBnet*, **17**, 10–12.

Meusnier, I., Singer, G.A.C., Landry, J.-F., Hickey, D.A., Hebert, P.D.N. & Hajibabaei, M. (2008). A Universal DNA Mini-barcode for Biodiversity Analysis. *BMC Genomics*, **9**, 214.

O'Donnell, J.L., Kelly, R.P., Lowell, N.C. & Port, J.A. (2016). Indexed PCR primers induce template-Specific bias in Large-Scale DNA sequencing studies (A.R. Mahon, Ed.). *PLoS ONE*, **11**, e0148698.

Peng, Y., Leung, H.C.M., Yiu, S.M. & Chin, F.Y.L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.

Pfrender, M., Hawkins, C., Bagley, M., Courtney, G., Creutzburg, B., Epler, J., Fend, S., Ferrington, L., Hartzell, P., Jackson, S., Larsen, D., Lvesque, C.A., Morse, J., Petersen, M., Ruiter, D., Schindel, D. & Whiting, M. (2010). Assessing Macroinvertebrate Biodiversity in Freshwater Ecosystems: Advances and Challenges in DNA-based Approaches The Quarterly Review of Biology. *Source: The Quarterly Review of Biology*, **85**, 319–340.

Piñol, J., Mir, G., Gomez-Polo, P. & 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**, 819–830.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, **4**, 406–25.

Shaw, J.L.A., Clarke, L.J., Wedderburn, S.D., Barnes, T.C., Weyrich, L.S. & Cooper, A. (2016). Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biological Conservation*, **197**, 131–138.

Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., Golding, G.B. & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports*, **5**, 9687.

Sinniger, F., Pawlowski, J., Harii, S., Gooday, A.J., Yamamoto, H., Chevaldonné, P., Cedhagen, T., Carvalho, G. & Creer, S. (2016). Worldwide analysis of sedimentary DNA reveals major gaps in taxonomic knowledge of deep-sea benthos. *Frontiers in Marine Science*, **3**, 92.

Sweeney, B.W., Battle, J.M., Jackson, J.K. & Dapkey, T. (2011). Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society*, **30**, 195–216.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Tang, M., Hardman, C.J., Ji, Y., Meng, G., Liu, S., Tan, M., Yang, S., Moss, E.D., Wang, J., Yang, C., Bruce, C., Nevard, T., Potts, S.G., Zhou, X. & Yu, D.W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, **6**, 1034–1043.

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A. & Zhou, X. (2014). Multiplex sequencing of pooled mitochondrial genomes - A crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, **42**.

Team, R.C. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. R Foundation for Statistical Computing.

Thomas, A.C., Deagle, B.E., Eveson, J.P., Harsch, C.H. & Trites, A.W. (2016). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, **16**, 714–726.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P.F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G.H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.M., Peroux, T., Crivelli, A.J., Olivier, A., Acqueberge, M., Le Brun, M., M??ller, P.R., Willerslev, E. & Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, **25**, 929–942.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.W., Li, Y., Xu, X., Wong, G.K.S. & Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30**, 1660–1666.

Yang, C., Wang, X., Miller, J.A., De Blécourt, M., Ji, Y., Yang, C., Harrison, R.D. & Yu, D.W. (2014). Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators*, **46**, 379–389.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhan, A., Bailey, S.A., Heath, D.D. & Macisaac, H.J. (2014). Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities. *Molecular Ecology Resources*, **14**, 1049–1059.

Zhan, A. & MacIsaac, H.J. (2015). Rare biosphere exploration using high-throughput sequencing: research progress and perspectives. *Conservation Genetics*, **16**, 513–522.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J. & Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 4.

Chapter 5
General Discussion

Chapter 5: General Discussion

5.1 Overview of experimental chapters

For the first chapter, I constructed a DNA barcode reference library for selected species of three groups of freshwater macroinvertebrates. This task was challenging mainly due to the logistics of sample acquisition and sequencing success rates for some species. Overall, 94 species were sequenced, out of which 55 Trichoptera, 17 Gastropoda and 22 Chironomidae species. Analysis of COI barcode sequences found a varying fit of the marker per group regarding accuracy of species level delimitations, as was expected due to known within-group divergences from the literature. Member species from all groups were generally well defined by DNA barcodes, with few exceptions, which could be related to cases of misidentification, incomplete sampling, cryptic species or hybridization. Barcode sequencing of chironomid specimens was challenging due to the nature of the pupal exuviae material used, and eventually led to low success rates of Sanger sequencing for this group (25%). Low levels of *Wolbachia* infections were found in Trichoptera (one species), while rotifer (in Trichoptera) and annelid (in Gastropoda) infestations were found in other species. The presence of such infestations could possibly reduce the efficiency of barcoding due to co-amplification of the parasites along with the target taxa. The amount of probable taxonomic misidentifications found ranged between 5.4% - 8% depending on the group.

In the second chapter, the use of water extracted environmental DNA (eDNA) was tested for biodiversity detection across an annual scale, in a temperate freshwater lake in N. Wales. Both water extracted DNA (eDNA) and invertebrate community samples (chironomid exuviae) were high-throughput sequenced for two varying length amplicons (COIF: 658bp, COIS: 235bp) of the Cytochrome Oxidase subunit I (COI). Our findings show that eDNA can successfully be used for tracking richness patterns of the target taxon (Chironomidae) from both eDNA and community DNA samples. Furthermore, eDNA was able to uncover a wider diversity of organisms from the lake ecosystem, including aquatic and semiaquatic taxa (fish and amphibians) and a large variety of arthropod taxa. Both

amplicons presented seasonal patterns of β-diversity for animal taxa as well as total diversity detected, which follow seasonal expected sub-groupings between summer and winter months. Environmental DNA results from the longer COI fragment presented annual richness patterns more closely resembling the seasonal variation known to occur for chironomids (Armitage *et al.* 2012), compared to the shorter fragment. Nevertheless, the reduced sequencing depth for the long fragment meant it could not be retained for all analysis. The shorter fragment detected higher diversity and with a better depth, but failed to characterise temporal patterns from eDNA, as it did for community samples. These findings indicate that eDNA metabarcoding can be used for detection of invertebrate diversity, and that longer fragments could be more effective in presenting contemporary diversity, but might require increased sequencing effort because longer fragments are expected to be less abundant. Finally, comparison with taxonomically identified chironomid exuviae, which were simultaneously collected, presented comparable levels of diversity with the COIS fragment both at the richness and abundance level, and the abundance levels were found to be significantly correlated with expected species frequencies for the UK. Overall, I show that the application of eDNA is very promising for enhancing biomonitoring and ecosystem level patterns of biodiversity.

For the final experimental chapter, I used metabarcoding of three different amplicons of a single marker (COI barcoding region) and mito-metagenomic sequencing (shotgun sequencing of mitochondrial genomes) to characterise the relative abundance composition of 10 mock communities of macroinvertebrates. I have found evidence of PCR related biases in the metabarcoding work which might have been the cause of random misrepresentation of species in some of the mock communities. The mito-metagenomic approach was also found to miss certain taxa, but only for those present at very low relative abundance. Importantly, mito-metagenomic sequencing was found to present highly significant correlations with biomass content of the communities, when metabarcoding failed to show significant relationships for a large number of species. The accuracy of read-biomass relationships obtained from metabarcoding varied between amplicons, and the combination of sequencing reads across amplicons only slightly improved the correlations (for some species). These last findings only partially support

studies, which propose the use of multiple amplicons for increased diversity detection in community samples. Furthermore, shotgun mito-genome sequencing performed best when a reference genome was already obtained for the particular species. Overall, the superiority of mito-metagenomic sequencing for more accurate characterisation of community composition is supported, whilst considering that the sequencing depth and the presence (or absence) of a reference mito-genome could be important limiting factors of the accuracy of the method.

The present chapter discusses the main findings of the thesis, placed in a wider context and in relation to future applications in biomonitoring. The limitations of the work are also presented whilst alternative paths are suggested based on upcoming developments in the field.

5.2 Summary of main findings per chapter

Chapter 2

- In total, 94 indicator species of macroinvertebrates were barcoded for the purposes of a UK Barcode Reference Library, including 55 Trichoptera, 17 Gastropoda and 22 Chironomidae.
- The COI barcoding region can successfully be applied for the identification of Trichoptera, Gastropoda and Chironomidae to the species level.
- Individual sequencing of Chironomidae exuviae presented a challenging source of DNA for individual specimen DNA Barcoding, due to the presence of environmental contaminants. This finding suggests that other tissue sources (e.g. larvae, adults) might be more successful for future DNA barcoding efforts of Chironomidae species.
- Low levels of misidentification were detected amongst Trichoptera and Gastropoda studied species, ranging between 5.4-5.5%.
- Collection of DNA barcoding data can aid in providing a connection between identification of different life stages, flagging the presence of possible cryptic

species for further analysis, aid re-classification of taxonomic keys where needed, and the overall gathering of genetic information.

Chapter 3

- The collection of temporal eDNA data as performed here, is a novel aspect for the field, as no other study has currently addressed the variation of macrobial eDNA through an annual cycle of collection, allowing detection of seasonal variations of eDNA presence in the wild.
- The present study offers one of the first cases where metabarcoding of eDNA is being used for detection of invertebrate species in the wild, as well as for an important indicator group.
- Comparison of the performance of two lengths of COI amplicons suggests that longer fragments could more accurately present the contemporary diversity. Nevertheless, sequencing of longer fragments requires increased sequencing effort due to their lower availability related to faster degradation rates.
- Chironomidae richness was successfully detected through the year from both sample types and with substantial overlap with taxonomically identified samples.
- Seasonal patterns of beta diversity were found for both total and animal taxa that were detected through eDNA samples.
- Metabarcoding of chironomid exuviae collected with the CPET technique, used for the first time in this study, can be employed for characterisation of the chironomid community in lakes, with potential for future biomonitoring applications.

Chapter 4

- Comparison of metabarcoding and metagenomics pipelines suggests that the use of PCR-free sequencing of mitochondrial genomes can more accurately represent species biomass in bilk invertebrate samples.
- Species richness estimations were comparable between the two sequencing methods.

- Detection of rare or low biomass species was influenced by the depth of sequencing used. Cases of undetected species for the metabarcoding pipeline were not always due to low biomass, but could also be related to primer specificity.
- Our results suggest that the combination of reads from all three amplicons used, improved the accuracy of biomass estimations by a small degree, but was more successful in removing false negatives from the metabarcoding data.
- The unique design of the present experiment allowed us to perform precise comparisons between methods, due to individual measurements of specimens and known content per community. Additionally, use of whole bodies for extraction of DNA resembles possible real-life applications with reduced handling time compared to specimen subsampling protocols.

5.3 The barcode reference library paradox - to build or not to build?

The construction of a barcode reference library for UK macroinvertebrates for Chapter 2, was an exercise in perseverance and patience. Some of the difficulties involved in that process included recruitment of qualified taxonomists, sample preservation and transportation, and difficulties with extracting some taxonomic groups/life stages. Designing a sampling strategy for a moderate sized barcode reference library requires considerable effort, especially when endeavouring to describe geographic variation as well. Summing the costs of commissioning taxonomist experts for species collection and identification, and sample transportation in ethanol significantly increases the cost for construction of a reference library, in addition to the extraction, amplification and sequencing costs. With all that in mind, it is not a surprise that many studies, which use metabarcoding, do not embark in reference library construction, alongside the HTS work. Nevertheless, metabarcoding largely relies on the existence of a reference library for accurate taxonomic assignment of OTUs (Taberlet *et al.* 2012b), and that constitutes the "to build or not to build" paradox of constructing barcode reference libraries.

An alternative collection strategy that could help with library construction would be a "bio-blitz" type of specimen collection (e.g. Baird & Sweeney 2011). In this case a concentrated effort is performed, with the participation of taxonomy experts who collect and identify a multitude of species in a short time and space. This approach is usually not targeting some specific species but rather as many as can be found and processed in that short period, but it is the collaboration between taxonomists and molecular scientists that provides the advantage here. To demonstrate the potential of rapid barcoding data collection, a large inventory of a temperate ecosystem species was completed and published over the 6th International Barcode of Life Conference in Guelph (August 2016). Moreover, collaborations between the UK Environment Agency and Bangor University have also yielded sizable collections (ca. 200 species) resulting from bio-blitz type sampling days following the empirical work completed for this PhD. Following the acquisition of further resources, such collections will further augment the work started here.

One of the weaknesses of the presently constructed barcode reference library (Chapter 2) is the limited number of specimens that were sequenced in some species, due mainly to low availability of specimens. This probably limits our ability to estimate levels of intra specific divergence for these species (Joly *et al.* 2014). Nevertheless, the species information collected here, which are in most cases commonly used indicator species, will be useful for development of future work. Additionally, I was able to utilise immediately the barcodes collected for this library, for taxonomic assignment of chironomid OTUs from eDNA samples (Chapter 3), taxonomic assignment of OTUs for metabarcoding of mock communities and for assembly of reference mito-genomes and metagenomes (Chapter 4), which demonstrates the utility of this effort in practise. On an international level, extended consortia have been formed in order to pursue the construction of barcoding libraries on a large scale. Such are for example the Norwegian, German and Brazilian Barcode of Life projects (NorBoL, GBOL and BrBOL). These projects comprise collaborators with taxonomic or molecular expertise and infrastructure and funding from government or EU sources.

5.4 Next-generation barcoding future developments.

The new era of DNA barcoding could embrace two new developments in the future. First, multiplexing of individual species barcoding for next generation sequencing instead of individual Sanger sequencing (Shokralla *et al.* 2015) and second, sequencing of mitochondrial genomes in the form of "super-barcodes" (Crampton-Platt *et al.* 2016).

Sequencing of individual species barcodes on a HTS platform (Illumina MiSeq) was tested by Shokralla *et al.* (2015), demonstrating not only that this approach is feasible, but also that it can produce a larger number of individual barcodes in comparison to Sanger sequencing. A calculation of the costs related to generating barcodes with this method, showed an estimated $7 per specimen for the Sanger method versus a $1.5 per specimen for HTS. This is an almost 5 times reduction in cost, with an associated ~5 times reduction in hands on processing time. The overall cost of individual Sanger sequencing is a factor of severe limitation for many studies and generally discourages upscaling of efforts (Shokralla *et al.* 2015), but this development could encourage an increase in barcoding endeavours in the future. Nevertheless, taking into consideration the operational costs of performing HTS platform runs (costs per run, not per sample), these approaches will be practically applicable only when many species are included in the run.

On the other hand, sequencing the entire mitochondrial genomes of species, instead of a single marker, could soon become a reality due to the current advances of mito-metagenomics (Joly *et al.* 2014). Shotgun sequencing of mixed assemblages has already shown its potential for assembling large numbers of partial mitochondrial genomes through "genome skimming" (Linard *et al.* 2015), or assembly of a multitude of complete mito-genomes from bulk samples (Tang *et al.* 2014, 2015; Gómez-Rodríguez *et al.* 2015). However, the potential of this method could also be applied for providing the complete information of mitochondrial genomes from sample mixtures, which could be used as "super barcodes" comprising multiple markers instead of just one (Crampton-Platt *et al.* 2016).

5.5 Environmental DNA from concept to practise

The utility of environmental DNA has been promoted as: the new bright future of ecological monitoring; one of the most important tools in tackling the identification of species of conservation importance (Sutherland *et al.* 2013) and as a "game changer" in biodiversity monitoring (Lawson Handley 2015). Nevertheless, the distance that remains to be covered between method testing and end line stakeholder applications is still significant (but see Biggs et al. 2015 for recent breakthroughs in the use of eDNA for detection of Great Crested Newts in the UK). Several important considerations regarding the nature of eDNA remain to be resolved for eDNA surveys to be practically applicable. The most important of these considerations include, (1) the relationship of eDNA with abundance as a way to estimate population size, (2) the determination of the lowest cut-off of abundance that would allow positive eDNA detection, (3) the role of inhibitors in the detection of eDNA from target taxa, (4) the persistence time of eDNA after it is released in the environment, (5) the way eDNA is distributed in the environment (Biggs *et al.* 2015). For use in real life surveys practical considerations also include the ease of sample collection, number of replicates to be collected and immediate preservation.

Some results presented so far suggest that abundance estimation through eDNA samples is possible (Doi *et al.* 2015; Klymus *et al.* 2015) though these results have been criticised, especially regarding the ability of eDNA to distinguish between total biomass and relative abundance or variation of results due to patchy distribution of organisms in the wild (Iversen *et al.* 2015). Moreover, the lowest abundance that allows species detection would have to be determined with more accuracy to reduce the amount of false negatives. Since it has been reported that the life stage, metabolic rate or temperature can alter the shedding rates of eDNA (Maruyama *et al.* 2014; Klymus *et al.* 2015), species-specific studies will possibly have to be undertaken including observations on life-stage composition of the population and seasonal sampling to account for temperature variations.

Biggs et al. (2015) performed a "citizen science" survey and demonstrated the value of recruiting volunteers to assist with sample collection on wide scale surveys. These types of surveys are made feasible with the type of sampling used for eDNA, since the simple

collection of water samples requires less time and expertise than traditional surveys would. In Chapter 3, I used filtration and freezing of filter membranes for sample collection and preservation, which was applicable here due to local sampling and possibility for immediate laboratory processing and storage at -80°C. Nevertheless, this approach would not be feasible for sampling remote locations without access to freezer storage. An alternative method for sample preservation which has been suggested includes storage of filter membranes in CTAB or Longmire's buffer, which would allow sufficient preservation and ease of transportation in the field (Renshaw *et al.* 2015).

For eDNA surveys to be accurate, it is imperative that we can be confident regarding the contemporary nature of the diversity detected (Thomsen & Willerslev 2015). At first, this statement suggests that accurate knowledge of how long eDNA can persist in the environment is needed. Since it has been shown that DNA found in sediments can persist for longer periods that in water (Barnes *et al.* 2014), we should also control for sediment contamination in our samples to ensure contemporary representation. Another way to control for analysing contemporary DNA could be to increase the length of the DNA fragments analysed. Previous studies have shown that longer DNA fragments degrade more rapidly (Lindahl 1993), which implies that any long fragments found are likely to be contemporary in nature. The downfall of working with long fragments though is that they are also less abundant (Deagle *et al.* 2006), making them harder to detect. In Chapter 3, I tested this hypothesis by comparing two fragment lengths of the COI barcoding region. Indeed, the diversity patterns observed over time for the longer fragment where more closely matching the community DNA metabarcoding analysis and literature based expected patterns. Nevertheless, the sequencing depth achieved for this marker was lower than that normally required, which is probably a result of the expected lower abundance for longer fragments. Future testing of this hypothesis should either increase the sequencing depth or employ group specific/ group blocking primers, which would provide enhanced detection of particular groups of interest.

Most of the eDNA applications so far have focused on specific species of interest; that mainly includes animals of conservation importance (e.g. great crested newt) and invasive species, which constitute major threats by their introduction to non-native ecosystems

(e.g. American Bullfrog and Asian carp). The practical application of eDNA has already been extensively used for the detection of Asian carp species in the Great Lakes and other freshwater systems (Klymus *et al.* 2015). When it comes to invasive species detection, eDNA could also prove valuable as an early detection system due to its increased sensitivity compared to traditional methodologies (Dejean *et al.* 2011; Jerde *et al.* 2011). If early detection and prevention of the establishment of invasive species is made possible through eDNA detection, the benefits could extend beyond biodiversity conservation purposes to societal benefits as well, such as ecosystem services and the economy (Taberlet *et al.* 2012a). As an example, the estimated costs for controlling freshwater invasive species in Great Britain range between £26.5 - 43.5 million per year depending on the extent of management efforts undertaken (Oreska & Aldridge 2011). It is therefore understandable that any development for early detection of invasive species would be economically beneficial.

The future of eDNA applications will involve the study of a wide range of organisms whilst looking at ecological interactions, food webs and ecosystem structure (Goldberg *et al.* 2015). In Chapter 3 of the present work, I used eDNA to detect diversity at the ecosystem level, using universal primers, which is a fairly novel approach for the field. Here I also used eDNA for the detection of invertebrate species which is also a rare thing in eDNA work, as only few papers have undertaken invertebrate detection so far (e.g. Thomsen *et al.* 2012; Deiner *et al.* 2015). It is interesting that invertebrates, even though they are extensively used for ecosystem monitoring, have been generally overlooked in conservation research (Donaldson *et al.* 2016). It could be possible that this is related to their higher diversity, which makes them more difficult to identify, especially using qPCR approaches that have dominated the eDNA field so far.

Overall, identifying the weaknesses in eDNA analysis can only promote the accuracy of surveys and even though we should strive for higher quality, when comparing eDNA applications with already established methods, we should remember that even existing methodologies do not come without flaws. The full adoption of eDNA for applied monitoring will require time, but with the rate of the increasing advances this will

hopefully reduce the time periods involved, augmented by effective collaboration between stakeholders and researchers.

5.6 The potential of eDNA for enhancing studies of temporal turnover

Another important aspect of biodiversity studies involves the estimation of species assemblages' variation over time (Magurran 2011), also known as temporal turnover (Korhonen *et al.* 2010). The collection of temporal data is essential for monitoring changes is biodiversity, while long-term data can assist in deciphering the underlying causes of the change (Magurran *et al.* 2010). However, it is important to distinguish between changes that are attributed to natural phenomena such as temporal turnover and natural drivers (Lallias *et al.* 2015), or those that are due to anthropogenic influences (Magurran *et al.* 2010). Nevertheless, most ecological studies use spatial replication while the temporal aspects of biodiversity tend to be neglected. To address the lack of temporal data in similar studies, I used an annual cycle of collection in Chapter 3, to gain an understanding of seasonal variations and ecological relationships of species presence and community composition overtime.

Temporal turnover has been found to vary in aquatic ecosystems depending on several factors, such as the size and type of ecosystem (Korhonen *et al.* 2010). For example, larger ecosystems exhibit faster turnover than smaller ecosystems, as do lakes compared to rivers. Latitude also affects turnover rates, as yearly species turnover is faster in the tropics (Korhonen *et al.* 2010). Temporal turnover effects were detected for chironomids in Chapter 3, following variation that is expected for this group in temperate latitudes (Armitage *et al.* 2012). Further studies could extent this work to study different types of ecosystems or sites from different latitudes. Using traditional methodology such studies might be very difficult or impossible due to the high workload required for conventional ecological assessments. Nevertheless, the multiplexing options available for metabarcoding, the ease of sample collection for eDNA analysis and possibility for detection of a wide range of taxonomic groups, could make such research possible in the future.

5.7 Bioinformatics challenges for HTS monitoring applications

A possible pitfall when working with HTS data is the implementation of an "accurate" bioinformatics pipeline, as the type of tools and analysis approach used for processing metabarcoding data can strongly influence the results obtained (Thomsen & Willerslev 2015). The selection of taxonomic assignment method varies between studies but generally, the accuracy of the taxonomic assignment process relies largely on the presence of a formatted and curated reference database (Taberlet *et al.* 2012b). For the present work, taxonomic assignment was performed through BLAST identification and subsequent phylogenetic analysis. Even though parameters for best-hit selection can be chosen in BLAST (e.g. e-value), relying solely on the top hit can be risky due to the presence of errors in public databases. Verification of the best hit by a combination of low e-value, high maximum identity, selection from a number of top hits (e.g. top 10 hits) and phylogenetic reconstruction was used in metabarcoding analysis for the present work (Chapters 3 & 4), in order to minimize BLAST related errors. Additionally, alternative approaches for taxonomic assignment exist, such as the RDP classifier (Wang *et al.* 2007), SAP (Munch *et al.* 2008) which provides Bayesian based taxonomic assignment, and phylogenetic placement methods such as pplacer (Matsen *et al.* 2010).

Quality filtering of sequencing reads should always be employed in diversity studies in order to remove errors introduced into the dataset due to sample degradation, contamination, PCR amplification artefacts and sequencing errors (Coissac *et al.* 2012). The baseline of quality control for sequencing reads should include some minimum steps for trimming of sequencing reads based on Phred quality scores (provided by Illumina). The removal of singletons and chimeras has generally been established in most biodiversity studies using methodologies which are common ground from the more developed field of HTS microbial diversity (Bik *et al.* 2012). The removal of chimeras in particular, which are by-products of the amplification process due to the merging of multiple sequences (Edgar *et al.* 2011) is a crucial step, as it has been shown that their presence can inflate diversity estimates (Kunin *et al.* 2010). This step can be performed either *de novo*, or with the use of a reference database. Using a reference database provides more accuracy, but its use is limited by the absence of appropriately curated

databases, especially for whole ecosystem diversity studies (such as the present), though it is more feasible for taxonomic group specific studies (see Hänfling *et al.* 2016).

A more controversial aspect of the filtering pipeline involves abundance based filtering for removal of low abundance reads (Bokulich *et al.* 2013). For this step, a lowest abundance level of filtration is selected, but the criteria of selection tend to vary between studies, and no specific consensus currently exists (Murray *et al.* 2015). Strict abundance filtering can be beneficial for removal of low level contamination, but we have to be mindful that real rare diversity could be discarded at the same time (Zhan & MacIsaac 2015), so a careful selection of a filtering threshold is advised depending on the study (Bokulich *et al.* 2013). Here we have selected a dual strategy for abundance filtering, using a level defined by the proportion of non-target reads found in positive control samples (see Port *et al.* 2016) and comparison against the expected levels of diversity for that particular ecosystem based on historical data (see Valentini *et al.* 2016).

Metagenomic analysis can also suffer from bioinformatics related errors. These could be for example related to the ability of the assembler to detect chimeric contigs or even form viable contigs in highly diverse samples, especially when closely related species are present in the mix (Gómez-Rodríguez *et al.* 2015). In Chapter 4 I used two congener species (*B. tentaculata* and *B. leachii*) to test the ability of the assembly process to handle closely related species, with satisfactory results. Nevertheless, the success of this step was probably assisted by the previously sequenced reference genomes for these species.

5.8 Perspectives on the utility of the COI marker for biodiversity assessment studies

The COI barcoding marker has been very valuable for detecting diversity in community analysis using metabarcoding (e.g. Hajibabaei *et al.* 2011; Ji *et al.* 2013), nevertheless its use does not come without criticism both in regards to the universality of the COI marker and its suitability for HTS (Deagle *et al.* 2014). These criticisms suggest that the standard markers used for barcoding might not be compatible with the needs of HTS of environmental samples (Coissac *et al.* 2012). For example, the length of the amplicons produced by classic barcoding primers (~650bp) could be too long for metabarcoding, due to the current limitations of the Illumina chemistry (maximum length of reads 2x300bp,

including primer and adaptor sequences). Also the fact that for metabarcoding we have to work with fragmented or degraded samples in many cases, requires targeting of shorter fragments (e.g. diet analysis) (Coissac *et al.* 2012). Furthermore, it is suggested that universal primers (or even the amplification process itself) might produce taxonomic bias, by uneven representation of species presence and relative abundance in community samples (Yu *et al.* 2012).

Even though efforts are made to detect new markers (e.g. 16S, Epp *et al.* 2012; Clarke *et al.* 2014), the reality at this point is that there is currently no perfect marker for metabarcoding, but instead the marker selection should be study specific (Deagle *et al.* 2014). With that in mind, I have used the COI barcoding region with universal primers (Folmer *et al.* 1994; Hebert *et al.* 2003) throughout this work due to its potential advantages for this particular study, as it covered multiple study requirements, such as amplicon length, universality and availability of data in public databases (NCBI and BOLD). Moreover, the COI has been found to perform well for macroinvertebrate metabarcoding studies (Hajibabaei *et al.* 2011), while several studies have also demonstrated the utility of the COI for species level identification for a variety of aquatic taxonomic groups such as Trichoptera, Diptera, Gastropoda (which are used in this study) as well as other macrobial taxa such as fish or Amphibia.

Furthermore, for the present work I have also used other shorter amplicons of the barcoding region, with a particular interest on the taxa under investigation (chironomid targeting primers in Chapter 3 and two primer pairs modified to amplify our target species, in Chapter 4). The combination of multiple primer pairs has been found to increase accuracy for biodiversity detection (Gibson *et al.* 2014), including both augmentation in the detection of species richness as well as relative abundance. The latter was explored in Chapter 4, where it was shown that combining the results of all three amplicons used, increased the accuracy of the recovery of species relative abundance in the mock communities, but not considerably. Furthermore, other universal primers have more recently been designed for metazoan diversity studies (Leray *et al.* 2013), but in this case the amplicon produced is shorter (~300bp) than the Folmer region amplicon. The

longer amplicon was chosen here in order to assess the effects of amplicon length on eDNA analysis (Chapter 3).

5.9 Additional work

In addition to the work described in detail in the three preceding experimental chapters, further data were acquired which do not feature in this thesis. In relation to Chapter 3, the initial experimental design also involved sequencing of three additional markers. One more COI marker with an intermediate length (~450bp) was sequenced both from eDNA and chironomid community samples. Furthermore, a ribosomal RNA (16S) fragment targeting bacteria, and the RbcL marker, targeting diatoms, were also sequenced from water extracted eDNA samples. Due mainly to time constrains these additional data were not included in the analysis presented here, but remain to be analysed soon after in order to provide further insights of between group dynamics of the lake ecosystem. In relation to the barcode reference library (Chapter 1), a number of additional species was collected mainly from Trichoptera and Coleoptera as well as some members of the Ephemeroptera, Plecoptera, Gastropoda, and Isopoda groups were also collected. These species were not sequenced in the course of the present project due to budgetary and time limitations. The extensive diversity contained in these collected specimens, which have been collected from different types of ecosystems (streams, ponds, lakes) and wide geographic range spanning from Cornwall to Scotland, constitute a valuable resource and they will hopefully be incorporated into future projects. Furthermore, the co-authoring of a book chapter relating to the conservation and monitoring of freshwater ecosystems with community structure and ecosystem function was undertaken (Gray *et al.* 2015).

5.10 Implications of the work for the stakeholder community and future suggestions

Even though many methodological advances have been made in the DNA based side of biomonitoring work (e.g. Hajibabaei *et al.* 2011; Ji *et al.* 2013), the traditional biomonitoring community has been slow in applying these new advances in the study area, while some of the methods still rely on obsolete ecological notions (Woodward *et al.* 2013). However, instead of forcing old methodologies to provide answers to new questions, we should aim to adopt newly developed approaches into the future (Jackson *et al.* 2016).

In recent years, the stakeholder community (in the UK and other countries) has been increasingly involved in research and development of DNA based approaches for ecological monitoring, policy makers have begun to be influenced by modern eDNA applications, and are open to investigating opportunities for integrating them to their monitoring regimes (Kelly 2016). One such example is the work presented in this thesis, which was facilitated by the Environment Agency. Other stakeholder parties are also currently investing in method development for freshwater and marine ecosystems, for a variety of target organisms (see UK eDNA Working Group proceedings). Additionally, the National Environment Research Council (NERC) has recently recognised the potential of using molecular tools, such as eDNA, in research and ecosystem management by funding three projects as part of a Highlight topic on "eDNA: a tool for 21st century ecology".

The cases where this work has actually been legally recognised and implemented as an efficient tool are still scarce, as the data acquired should be sufficiently reliable to satisfy legal standards (Kelly *et al.* 2014). One particular legally recognised case has been made so far in the UK for the detection of great crested newts (e.g. Biggs *et al.* 2015). Through this work, it was demonstrated that use of eDNA detection was more effective than traditional methods, with an estimated eDNA detection from one sample visit being equivalent to 5 survey visits (3 survey methods) (Biggs *et al.* 2015). This application advantage, due to increased sensitivity of eDNA, could reduce operating costs for stakeholders, through the requirement of fewer survey visits. Similar findings supporting the cost effectiveness of DNA-based monitoring have been reported by Ji *et al.* (2013), who also suggest that

metabarcoding for biodiversity surveys could allow direct measurements of total diversity instead of the commonly used indicator groups.

A possible hurdle for the adoption of eDNA work in practise by stakeholders is related to the recovery of accurate species abundance information, which is still not fully resolved, due mainly to primer bias related issues as discussed previously (Chapter 4). However, the level of quantitative information required by the various policy applications varies (Kelly *et al.* 2014). For example, detection of invasive species surveys relies on presence-absence data (e.g. Schmidt *et al.* 2013; Klymus *et al.* 2015), while some of the Water Framework Directive (WFD) measures rely on relative abundance counts (Hatton-Ellis 2008). Nevertheless, progress has already been made in standardising relative abundance results, as was done for example in Evans *et al.* (2016). This attempt nevertheless refers to the specific studied taxa, for which the use of multiple primer pairs has proven beneficial (Kelly 2016). The unresolved question here however is, whether these conclusions are transferable to other organisms or different life stages. Other considerations involved in the full adoption of molecular approaches for ecosystem monitoring include, costs of establishing infrastructure, which is not currently available, as well as availability of appropriately trained personnel to undertake this work. As an example, use of eDNA sampling as in Biggs *et al.* (2015), required minimum training and experience of the participating samplers, which indicates that potential adoption of this sampling approach by the stakeholders might have fewer training requirements than previously feared.

The future calls for an urgent need for innovative approaches, which would allow large scale monitoring and a move from targeted single-species essays towards community wide meta-analysis (Thomsen & Willerslev 2015). A horizon scanning exercise identified three types of directions that the policy makers could move towards improving biomonitoring of freshwaters (Jackson *et al.* 2016). Particularly it was suggested that the use of new technological advances such as molecular tools and remote sensing, while enhancing citizen science networks could represent a significant advance in tackling logistical issues in large scale ecological surveys (Jackson *et al.* 2016).

5.11 Concluding remarks

Overall, this work has attempted to provide linkages between individual barcoding, metabarcoding of communities and mitochondrial genomes leading to the enhancement of ecological assessment as a means of preserving biodiversity and monitoring health of freshwater ecosystems. I have shown here that detection of biodiversity on an ecosystem wide scale is possible using metabarcoding of eDNA, and that optimization of this methodology could enhance our certainty of accurately characterising contemporary diversity in freshwater ecosystem. Furthermore, I provide evidence that the impediment of PCR-based methods could be overcome by the incorporation of PCR-free whole mito-genome sequencing into routine assessment, which would increase our confidence on community composition estimates. Finally, the advances in HTS monitoring will greatly benefit by the continuous efforts to populate reference library records from the single marker to the whole mito-genome level.

References

Armitage, P.D., Pinder, L.C. & Cranston, P. (2012). *The Chironomidae: biology and ecology of non-biting midges*. Chapman and Hall, London.

Baird, D.J. & Sweeney, B.W. (2011). Applying DNA barcoding in benthology: the state of the science. *Journal of the North American Benthological Society*, **30**, 122–124.

Barnes, M.A., Turner, C.R., Jerde, C.L., Renshaw, M.A., Chadderton, W.L. & Lodge, D.M. (2014). Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science and Technology*, **48**, 1819–1827.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R.A., Foster, J., Wilkinson, J.W., Arnell, A., Brotherton, P., Williams, P. & Dunn, F. (2015). Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). *Biological Conservation*, **183**, 19–28.

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R. & Thomas, W.K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution*, **27**, 233–243.

Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, I., Knight, R., Mills, D.A. & Caporaso, J.G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, **10**, 57–59.

Clarke, L.J., Soubrier, J., Weyrich, L.S. & Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, **14**, 1160–1170.

Coissac, E., Riaz, T. & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, **21**, 1834–1847.

Crampton-Platt, A., Yu, D.W., Zhou, X. & Vogler, A.P. (2016). Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience*, **5**, 15.

Deagle, B.E., Eveson, J.P. & Jarman, S.N. (2006). Quantification of damage in DNA recovered from highly degraded samples--a case study on DNA in faeces. *Frontiers in zoology*, **3**, 11.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., Taberlet, P., Taberlet, P., Coissac, E.,

Hajibabaei, M., Rieseberg, L., Yu, D., Ji, Y., Emerson, B., Wang, X., Ye, C., Yang, C., Ding, Z., Ficetola, G., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessiere, J., Taberlet, P., Pompanon, F., Geller, J., Meyer, C., Parker, M., Hawk, H., Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glockner, F., Bru, D., Martin-Laurent, F., Philippot, L., Schloss, P., Gevers, D., Westcott, S., Clarke, L., Soubrier, J., Weyrich, L., Cooper, A., Ji, Y., Barba, M. De, Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., Taberlet, P., Leray, M., Yang, J., Meyer, C., Mills, S., Agudelo, N., Ranwez, V., Boehm, J., Machida, R., Little, D., Deagle, B., Kirkwood, R., Jarman, S., Zhou, X., Shokralla, S., Gibson, J., Nikbakht, H., Janzen, D., Hallwachs, W. & Hajibabaei, M. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology letters*, **10**, 1789–1793.

Deiner, K., Walser, J.C., Mächler, E. & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, **183**, 53–63.

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P. & Miaud, C. (2011). Persistence of environmental DNA in freshwater ecosystems. *PLoS ONE*, **6**, e23398.

Doi, H., Uchii, K., Takahara, T., Matsuhashi, S., Yamanaka, H. & Minamoto, T. (2015). Use of droplet digital PCR for estimation of fish abundance and biomass in environmental DNA surveys. *PLoS ONE*, **10**, e0122763.

Donaldson, M.R., Burnett, N.J., Braun, D.C., Suski, C.D., Hinch, S.G., Cooke, S.J. & Kerr, J.T. (2016). Taxonomic bias and international biodiversity conservation research. *Facets*, **1**, 105–113.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Epp, L.S., Boessenkool, S., Bellemain, E.P., Haile, J., Esposito, A., Riaz, T., Erséus, C., Gusarov, V.I., Edwards, M.E., Johnsen, A., Stenøien, H.K., Hassel, K., Kauserud, H., Yoccoz, N.G., Bråthen, K.A., Willerslev, E., Taberlet, P., Coissac, E. & Brochmann, C. (2012). New environmental metabarcodes for analysing soil DNA: Potential for studying past and present ecosystems. *Molecular Ecology*, **21**, 1821–1833.

Evans, N.T., Olds, B.P., Renshaw, M.A., Turner, C.R., Li, Y., Jerde, C.L., Mahon, A.R.,

Pfrender, M.E., Lamberti, G.A. & Lodge, D.M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, **16**, 29–41.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 8007–12.

Goldberg, C.S., Strickler, K.M. & Pilliod, D.S. (2015). Moving environmental DNA methods from concept to practice for monitoring aquatic macroorganisms. *Biological Conservation*, **183**, 1–3.

Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M.J.T.N., Baselga, A. & Vogler, A.P. (2015). Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, **6**, 883–894.

Gray, C., Bista, I., Creer, S., Demars, B.O.L., Falciani, F., Don, T.M., Sun, X. & Woodward, G. (2015). Freshwater conservation and biomonitoring of structure and function: Genes to ecosystems. *Aquatic Functional Biodiversity: An Ecological and Evolutionary Perspective* (eds A. Belgrano, G. Woodward & U. Jacob), pp. 241–271. Elsevier.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C. & Baird, D.J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, **6**, e17497.

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

Hatton-Ellis, T. (2008). The Hitchhiker's guide to the Water Framework Directive. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **18**, 111–116.

Hebert, P.D.N., Ratnasingham, S. & Waard, J. (2003). Barcoding animal life : cytochrome c oxidase subunit 1 divergences among closely related species Barcoding animal life : cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B*, **270**, S96–S99.

Iversen, L.L., Kielgast, J. & Sand-Jensen, K. (2015). Monitoring of animal abundance by environmental DNA - An increasingly obscure perspective: A reply to Klymus et al., 2015. *Biological Conservation*, **192**, 479–480.

Jackson, M.C., Weyl, O.L.F., Altermatt, F., Durance, I., Friberg, N., Dumbrell, A.J., Piggott, J.J., Tiegs, S.D., Tockner, K., Krug, C.B., Leadley, P.W. & Woodward, G. (2016). Recommendations for the Next Generation of Global Freshwater Biological Monitoring Tools. *Large-Scale Ecology: Model Systems to Global Perspectives* (eds R. Kordas, A.J. Dumbrell & G. Woodward), pp. 615–636. Elsevier Ltd.

Jerde, C.L., Mahon, A.R., Chadderton, W.L. & Lodge, D.M. (2011). 'Sight-unseen' detection of rare aquatic species using environmental DNA. *Conservation Letters*, **4**, 150–157.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P., Edwards, F.A., Larsen, T.H., Hsu, W.W., Benedick, S., Hamer, K.C., Wilcove, D.S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B.C. & Yu, D.W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Joly, S., Davies, T.J., Archambault, A., Bruneau, A., Derry, A., Kembel, S.W., Peres-Neto, P., Vamosi, J. & Wheeler, T.A. (2014). Ecology in the age of DNA barcoding: The resource, the promise and the challenges ahead. *Molecular Ecology Resources*, **14**, 221–232.

Kelly, R.P. (2016). Making environmental DNA count. *Molecular Ecology Resources*, **16**, 10–12.

Kelly, R.P., Port, J. a., Yamahara, K.M., Martone, R.G., Lowell, N., Thomsen, P.F., Mach, M.E., Bennett, M., Prahler, E., Caldwell, M.R. & Crowder, L.B. (2014). Harnessing DNA to improve environmental management. *Science*, **344**, 1455–1456.

Klymus, K.E., Richter, C.A., Chapman, D.C. & Paukert, C. (2015). Quantification of eDNA shedding rates from invasive bighead carp Hypophthalmichthys nobilis and silver carp Hypophthalmichthys molitrix. *Biological Conservation*, **183**, 77–84.

Korhonen, J.J., Soininen, J. & Hillebrand, H. (2010). A quantitative analysis of temporal

turnover in aquatic species assemblages across ecosystems. *Ecology*, **91**, 508–517.

Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. (2010). Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, **12**, 118–123.

Lallias, D., Hiddink, J.G., Fonseca, V.G., Gaspar, J.M., Sung, W., Neill, S.P., Barnes, N., Ferrero, T., Hall, N., Lambshead, P.J.D., Packer, M., Thomas, W.K. & Creer, S. (2015). Environmental metabarcoding reveals heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine ecosystems. *The ISME journal*, **9**, 1208–21.

Lawson Handley, L. (2015). How will the 'molecular revolution' contribute to biological recording? *Biological Journal of the Linnean Society*, **115**, 750–766.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. & Machida, R.J. (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**, 34.

Linard, B., Crampton-Platt, A., Gillett, C.P.D.T., Timmermans, M.J.T.N. & Vogler, A.P. (2015). Metagenome skimming of insect specimen pools: Potential for comparative genomics. *Genome Biology and Evolution*, **7**, 1474–1489.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, **362**, 709–715.

Magurran, A.E. (2011). *Measuring biological diversity in time (and space)* (A.E. Magurran & B.J. McGill, Eds.). Oxford University Press.

Magurran, A.E., Baillie, S.R., Buckland, S.T., Dick, J.M., Elston, D.A., Scott, E.M., Smith, R.I., Somerfield, P.J. & Watt, A.D. (2010). Long-term datasets in biodiversity research and monitoring: Assessing change in ecological communities through time. *Trends in Ecology and Evolution*, **25**, 574–582.

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M. & Minamoto, T. (2014). The release rate of environmental DNA from juvenile and adult fish. *PLoS ONE*, **9**, e114639.

Matsen, F. a, Kodner, R.B. & Armbrust, E.V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, **11**, 538.

Munch, K., Boomsma, W., Huelsenbeck, J.P., Willerslev, E. & Nielsen, R. (2008). Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, **57**, 750–757.

Murray, D.C., Coghlan, M.L. & Bunce, M. (2015). From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. *PLoS ONE*, **10**, e0124671.

Oreska, M.P.J. & Aldridge, D.C. (2011). Estimating the financial costs of freshwater invasive species in Great Britain: A standardized approach to invasive species costing. *Biological Invasions*, **13**, 305–319.

Port, J.A., O'Donnell, J.L., Romero-Maraccini, O.C., Leary, P.R., Litvin, S.Y., Nickols, K.J., Yamahara, K.M. & Kelly, R.P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, **25**, 527–541.

Renshaw, M.A., Olds, B.P., Jerde, C.L., Mcveigh, M.M. & Lodge, D.M. (2015). The room temperature preservation of filtered environmental DNA samples and assimilation into a phenol-chloroform-isoamyl alcohol DNA extraction. *Molecular Ecology Resources*, **15**, 168–176.

Schmidt, B.R., Kéry, M., Ursenbacher, S., Hyman, O.J. & Collins, J.P. (2013). Site occupancy models in the analysis of environmental DNA presence/absence surveys: A case study of an emerging amphibian pathogen. *Methods in Ecology and Evolution*, **4**, 646–653.

Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., Golding, G.B. & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports*, **5**, 9687.

Sutherland, W.J., Bardsley, S., Clout, M., Depledge, M.H., Dicks, L. V., Fellman, L., Fleishman, E., Gibbons, D.W., Keim, B., Lickorish, F., Margerison, C., Monk, K.A., Norris, K., Peck, L.S., Prior, S. V., Scharlemann, J.P.W., Spalding, M.D. & Watkinson, A.R. (2013). A horizon scan of global conservation issues for 2013. *Trends in Ecology and Evolution*, **28**, 16–22.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012a). Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012b). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular*

Ecology, **21**, 2045–2050.

Tang, M., Hardman, C.J., Ji, Y., Meng, G., Liu, S., Tan, M., Yang, S., Moss, E.D., Wang, J., Yang, C., Bruce, C., Nevard, T., Potts, S.G., Zhou, X. & Yu, D.W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, **6**, 1034–1043.

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A. & Zhou, X. (2014). Multiplex sequencing of pooled mitochondrial genomes - A crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, **42**.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Wiuf, C., Rasmunsen, M., Gilbert, M.T.P., Orlando, L. & Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, **21**, 2565–2573.

Thomsen, P.F. & Willerslev, E. (2015). Environmental DNA - An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, **183**, 4–18.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P.F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G.H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.M., Peroux, T., Crivelli, A.J., Olivier, A., Acqueberge, M., Le Brun, M., M??ller, P.R., Willerslev, E. & Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, **25**, 929–942.

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Woodward, G., Gray, C. & Baird, D.J. (2013). Biomonitoring for the 21st Century: new perspectives in an age of globalisation and emerging environmental threats. *Limnetica*, **32**, 159–174.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhan, A. & MacIsaac, H.J. (2015). Rare biosphere exploration using high-throughput

sequencing: research progress and perspectives. *Conservation Genetics*, **16**, 513–522.