Biomonitoring and surveillance with short- and long-read metabarcoding

---

Submitted by Rachel Glover, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Biological Sciences
January 2019

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Acknowledgments

I would like to gratefully acknowledge my main supervisors, Neil Boonham and David Studholme for their guidance, support, encouragement and patience during the writing of this thesis. I would also like to thank my former FERA colleagues Ian Adams, Ian Brittain and Edward Haynes for the sample collection and production of the sequencing datasets used within this thesis. The work was supported by funding from both the Environment Agency and BBSRC project BB/L012251/1. I would particularly like to thank Kerry Walsh at the Environment Agency and Martyn Kelly at the Bowburn Consultancy for their invaluable input and advice.

I would like to add personal thanks to my friends and family for their constant encouragement, love and support while I tried to juggle far too many “big things” at once. I’m not sure there will ever be another time in my life where I will be juggling training for a Commonwealth Games, caring for a terminally ill parent, working full time, moving house, dealing with my own chronic illness, setting up a company and adding a PhD to the mix! The support will never be forgotten!

For Mum and Dad.

Abstract

The aim of this project was to develop applied metabarcoding methods to aid both the monitoring of environmental water quality and surveillance of airborne fungal phytopathogens. An Illumina short-read metabarcoding method was developed which is now in active use by the UK Environment Agency to determine the abundance of diatom species, feeding into the classification of water bodies for the EU Water Framework Directive. Further work was undertaken to future proof this method by comparing the diatom assemblages of three English rivers as determined by light microscopy, the developed Illumina short-read metabarcoding method and long-read nanopore metabarcoding. The river and method comparison study showed that the light microscopy was the outlier and potentially largest source of error as the two separate metabarcoding methods performed very similarly. A study was undertaken to compare the airborne fungal communities in six locations in eastern England over a one month period in 2015 to assess metabarcoding as a potential surveillance tool for the introduction of phytopathogens and the utility of the current UK pollen network for sample acquisition. This study highlighted issues with contamination at stages within the metabarcoding laboratory preparation protocols which made the bioinformatics analysis problematic; however, recommendations are made for procedures to reduce contamination in metabarcoding studies. The final study characterised a weeks' data from the eastern England fungal spore samples with a novel long PCR amplifying the entire ribosomal tandem repeat - including the intergenic spacer - to investigate the regions full utility in the light of nanopore sequencing.

Author's declaration

The work in this thesis depended upon sequencing data generated by laboratory scientists at FERA from samples obtained by the Environment Agency. The author of this thesis designed the amplicon sequencing experiments and carried out the bioinformatics analyses but the laboratory work described in the methods was carried out by collaborators and colleagues at FERA. Preparation of diatoms for microscopy was carried out by either Environment Agency operations staff or Martyn Kelly (Bowburn Consultancy).

Published materials originating from this work:

Kelly, M., Boonham, N., Juggins, S., Kille, P., Mann, D., Pass, D., Sapp, M., Sato, S., **Glover, R** (2018) A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers. *Environment Agency* Report for project SC140024.
https://www.gov.uk/government/publications/a-dna-based-metabarcoding-approach-to-assess-diatom-communities-in-rivers

This work is included as Appendix I.

Mann, D.G., Kelly, M.G., Walsh, K., **Glover, R.**, Juggins, S., Sato, S., Boonham, N., Jones, T (2017) Development and adoption of a next-generation sequencing approach to diatom-based ecological assessments in the UK. *Phycologia* 45 (4), 125.

Conference talk abstract:

Glover, R., Sapp, M., Adams, I., Hany, U., *et al* (2015) Metabarcoding for surveillance and monitoring: meeting policy objectives in the real world. *GENOME* 58 (5), 221. Talk presented at the 6th International Barcode of Life Conference, University of Guelph, Canada, 2015.

Other papers published during the PhD:

Massart, S., Chiumenti, M., De Johnghe, K., **Glover, R.**, Haegeman, A., et al (2018) Virus detection by high-throughput sequencing of small RNAs: Large scale performance testing of sequence analysis strategies. Phytopathology (Accepted) https://doi.org/10.1094/PHYTO-02-18-0067-R

Sabbadin, F., **Glover, R.**, Stafford, R., Rozado-Aguirre, Z., Boonham, N., Adams, I., Mumford, R., and Edwards, R. (2017) Transcriptome sequencing identifies novel persistent viruses in herbicide resistant wild-grasses. Nature Scientific Reports 7, 41987 https://doi.org/10.1038/srep41987 (**Joint first author**)

Pritchard, L., **Glover, R.**, Humphris, S., Elphinstone, J.G., Toth, I.K. (2016) Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods* 8(1), 12-24.

Book chapter published during the PhD:

Glover, R., and Adams, I (2016) Next-generation sequencing. In: *Molecular Methods in Plant Disease Diagnostics: Principles and Protocols*. Published by: CABI.

List of Figures

List of Tables

Contents

Table of Contents

Chapter 1. Introduction

Various definitions exist for the terms ‘monitoring’ and ‘surveillance’ but it is generally accepted that monitoring is the systematic sampling of an environment to assess specific variables that inform its current status (Artiola et al., 2004) and surveillance is the ongoing sampling of an environment whereby action could be taken if the data indicates a new threat (Christensen, 2001). Monitoring and surveillance of air and water environments are important to food security and water quality, respectively; however, the methods currently used can be prohibitively expensive or very resource intensive. Both monitoring and surveillance consist of recording the presence and abundance of species in a sample and when many samples are required in different locations over various time points to characterise changes in an environment the costs can be prohibitive with traditional methods (Targetti et al., 2014). In the UK, there is a requirement for biodiversity monitoring and these obligations are regulated with common agreements and legislation, for example, the Conservation of Habitats and Species Regulations, 2017. There is an important need for methodology which can measure and assess our environment in a cost effective manner, without the loss of critical information. Modern DNA techniques offer the potential for such monitoring methods with the additional resolution for surveillance for new and emerging threats.

Metabarcoding as a method for determining the species present within a sample

DNA barcoding is a technique which can be used to assign unknown individuals to species as well as enhancing the discovery of new species (Hebert et al., 2003a). Creation of the ‘barcode’ involves the PCR amplification and sequencing of a standardised region and the subsequent comparison to a database of known DNA barcodes produces the identification. Many projects in recent years were centred around the production of reference DNA barcodes from morphologically verified specimens (Hollingsworth et al., 2009; Schoch et al., 2012; Ward et al., 2009). Yet despite the push to collect specimens and produce DNA barcodes there were few applied routine uses of standalone DNA barcoding. With the advent of high-throughput sequencing DNA barcoding can now be used in a massively parallel way to identify the complement of species in a sample. This

method is known by many different names (amplicon metagenomics, meta-genetics, targeted metagenomics) but for the purposes of clarity it will be referred to here as metabarcoding. The applied uses for metabarcoding far exceed those of single-specimen DNA barcoding as many samples contain complex mixtures of species rather than single specimens.

Taxonomic challenges in fungal identification

Ribosomal DNA (rDNA) genes are known to accumulate mutations slowly and are present in multiple copies within the genome. These copies are subject to concerted evolution where mutations which appear in one repeat within that species are maintained and copied to other repeats in the region, yet between species the sequences can be very different (Ganley and Kobayashi, 2007). The Internal Transcribed Spacer (ITS) region is frequently used in studies where discrimination of fungal species is required due to its hypervariability. The region is part of a larger tandem repeat (Figure 1.1) comprising the small ribosomal subunit (SSU/18S), ITS1, 5.8S ribosomal submit, ITS2, large ribosomal subunit (LSU/28S), 5S ribosomal subunit and the two intergenic spacers (IGS1 and IGS2). The 5S ribosomal region can vary in position (Iwen et al., 2002) and the length of the ITS1, ITS2, IGS1 and IGS2 spacers can vary between species (Hausner and Wang, 2005) . The ITS1-5.8S-ITS2 region is capable not only of species discrimination but also sub-species and individual identification due to the hypervariability of the ITS spacers.

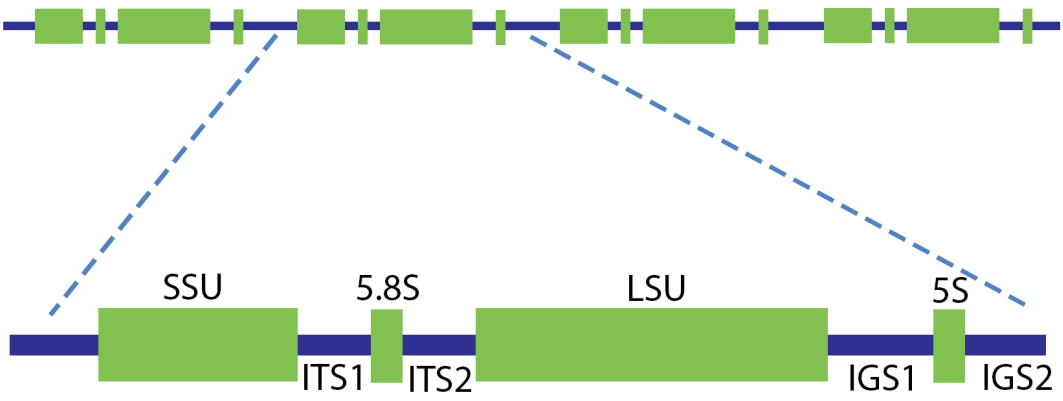


Figure 1.1: The tandemly repeated structure of fungal rDNA, consisting of the small subunit (SSU), 5.8S, large subunit (LSU) and 5S, separated by transcribed spacers ITS1 and ITS2, and the two intergenic spacers IGS1 and IGS2

The ribosomal tandem repeat regions have been used extensively in the past for species identification, with different regions being used for different taxa prior to the adoption of a more standardised approach to fungal identification using ITS (Schoch et al., 2012). It is estimated that there are between 2.2 to 3.8 million fungal species with approximately 3-8% having been named (Hawksworth and Lücking, 2017).

Metabarcoding of 16S (bacteria) and ITS (fungi) have become common techniques in microbial ecology (Arfi et al., 2012; Caporaso et al., 2011). Many bioinformatics tools have been developed to aid metabarcoding data analysis, for example, QIIME (Caporaso et al., 2010) and mothur (Schloss et al., 2009). When comparisons are being made between samples, species-level resolution of the composition of each sample is not required and most tools default to genus-level. Additionally, almost all metabarcoding bioinformatics methods rely on clustering the amplicon sequences into Operational Taxonomic Units (OTUs) based upon a relatively arbitrary similarity percentage, usually 97%, prior to downstream analysis, which can also decrease the resolution available.

The identification of metabarcoding OTUs to species-level is not required for most applications of the technique. When microbial ecologists are comparing communities with 16S or ITS metabarcoding, genus level taxonomic identifications are more than adequate (Somervuo et al., 2017). Equally, absolute accuracy in the taxonomic assignment of OTUs is not a prerequisite as many studies can be carried out without taxonomic information at all, allowing comparisons between samples, locations or environments based purely on the presence of an OTU and its relative abundance in each sample (Cordier et al., 2017; Cordier et al., 2018). However, when the aim is to use metabarcoding for detection, monitoring or surveillance studies, species-level resolution is very important and, furthermore, accurate identifications are critical (Staats et al., 2016; Bell et al., 2016). As such, the sequencing technology used to produce the metabarcode and the length of the amplicon (and how phylogenetically informative it is) becomes critical to a project's success.

Over the last ten years, the read lengths of next-generation sequencing technologies have matured from ~25bp (Solexa) through to the now-redundant

454 FLX+ at ~800bp, with the predominant sequencing technology being Illumina, with a capability of ~300-500bp amplicons (Figure 1.2). More recently the maturation of very long read nanopore sequencing with the Oxford Nanopore MinION suggests a future where metabarcoding is redundant and metagenomics (the sequencing of all genomic DNA in a sample) replaces it (Juul et al., 2015).

However, beyond the hype is the reality that full reference genomes are not yet available for the vast majority of species to allow accurate identification of species for statutory or regulatory purposes. For example, the Barcode of Life initiative has been working towards the Sanger sequencing and curation of DNA barcodes from all species for the past 12 years and has sequenced 259,955 species from 5,200,520 specimens (www.boldsystems.org). Replacement, even in part, of DNA barcodes with reference genome sequences will take many years to complete and will certainly present significant data management challenges. While a whole genome approach is likely to be the future, metabarcoding will likely be the interim method for many years to come.

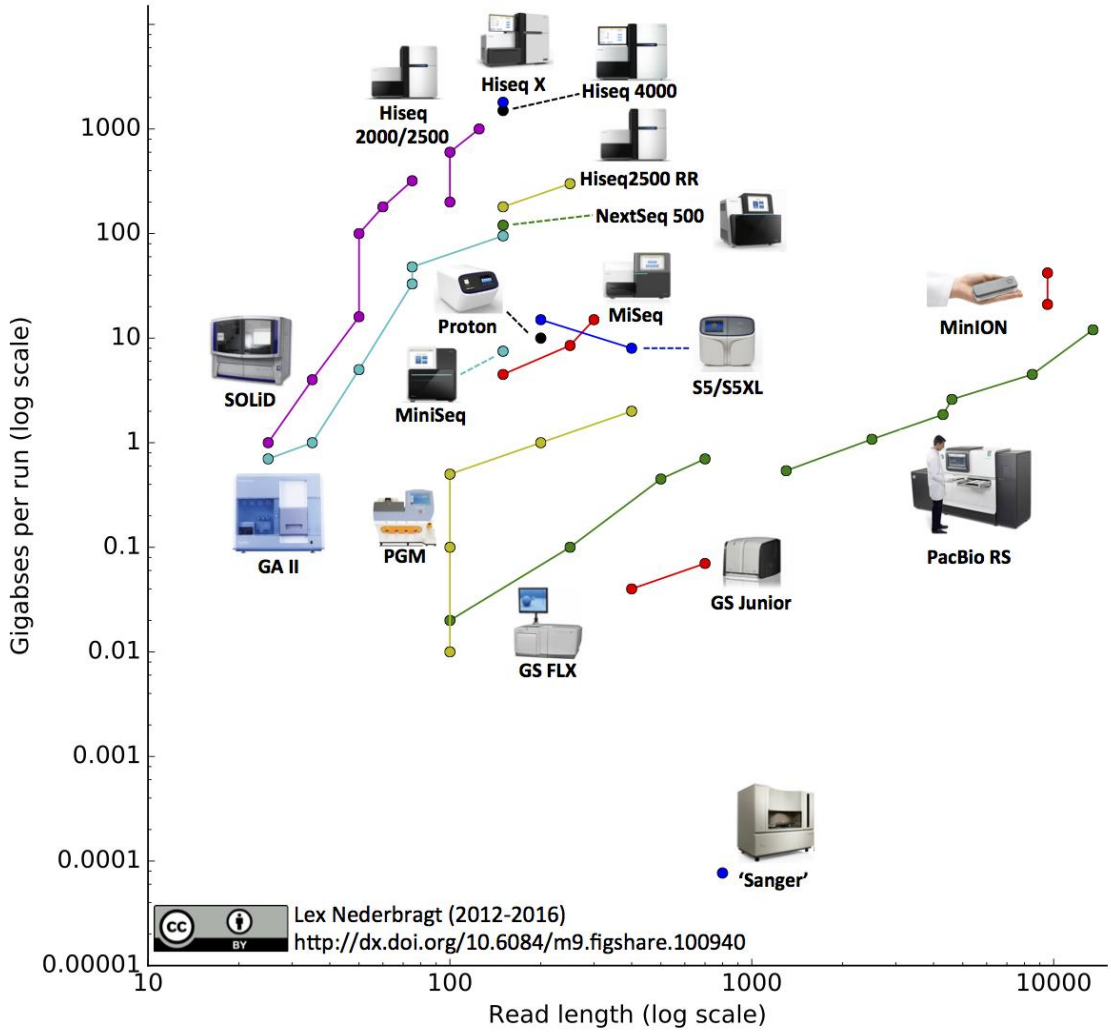


Figure 1.2 The evolution of various sequencing platforms with regards to their read length and throughput in gigabases. Each iteration and release of the technology is shown as an additional point in the same colour. The read lengths and throughput in gigabases has, in most cases, increased as further iterations of the technology have been released. (Credit: Lex Nederbragt https://flxlexblog.wordpress.com http://dx.doi.org/10.6084/m9.figshare.100940)

Short-read metabarcoding

Short-read metabarcoding can be currently defined as metabarcoding carried out with Illumina or Ion Torrent sequencing systems. The now decommissioned 454 sequencer (Roche, USA). was capable of read lengths up to 800bp but in light of newer long-read sequencing technologies it would also be classed as a short-read sequencing method (Wicker et al, 2006; Goodwin et al., 2016). There are multiple sequencing systems produced by Illumina, with each having their own qualities making them appropriate for different experiments, from amplicon and small genome sequencing with the iSeq, MiniSeq and MiSeq series through to whole genome sequencing and metagenomics with the NextSeq and HiSeq series. The short-read metabarcoding experiments in this project were carried out

with the Illumina MiSeq. The MiSeq platform is capable of small whole-genome sequencing (e.g. prokaryotes, viruses) and amplicon sequencing (metabarcoding, SNP genotyping). It is capable of sequencing up to 15Gb of DNA with 25 million sequence reads and 2x300bp read lengths in 55 hours, with a maximum amplicon size of 600bp (www.illumina.com). The MiSeq has a low error rate but the errors do not always occur randomly, they occur in a more predictable fashion with adenine and cysteine being more prone to substitution errors, often with a guanine, and library preparation has a major effect on the distribution of errors (Schirmer et al., 2015). The sequence reads deteriorate in quality from 5' to 3' which can lead to problems during the production of a consensus sequence if it is more than 400bp long, as the "middle" of the sequence will be comprised of the error-prone ends of both reads. This has been shown to lead to issues where the species richness within a sample is overestimated due to sequence errors rather than true sequence diversity (and thus species diversity) within a sample (Flynn et al., 2015). Sequencing platform and PCR primer biases have also been reported in bacterial 16S community sequencing (Tremblay et al., 2015). Stringent sequence quality practices prior to read merging and downstream analysis have been shown to produce more robust datasets for community ecology studies (Eckert et al., 2018).

Long-read metabarcoding

The clear advantage of long-read sequencing technologies to biodiversity studies is the ability to use longer, more informative regions for species identification. There are currently two predominant methods for long read sequencing. PacBio sequencing can generate up to 20Gb per Single Molecule Real Time (SMRT) sequencing cell with high accuracy read lengths up to 30kb. The SMRT cell technology relies upon fluorescently labelled nucleotides being incorporated into an extending DNA molecule. As each labelled nucleotide is incorporated, the fluorophore is cleaved from the nucleotide and a pulse of light is emitted which is detected by the sequencing system (www.pacb.com). The PacBio system is not dissimilar to other next-generation sequencing systems in that it determines the bases in each DNA strand from the detection of bases being incorporated during synthesis of another strand. The second long-read sequencing technology - nanopore sequencing - differs substantially from the other methods and has been

developed by Oxford Nanopore Technologies (ONT). Nanopore sequencing does not require optics or amplification and is carried out through a protein nanopores which are situated on an electrically resistant membrane within a flowcell. As DNA is ratcheted through each nanopore, the different base combinations produce different disruptions in current through the membrane. The current disruption is measured and the bases are identified with base calling software. There are currently three platforms produced by ONT: MinION (single flowcell device), GridION (five flowcell device) and PromethION (48 flowcell device). The most popular platform by far is the MinION due to its extreme portability with dimensions 10.5cm x 2.5cm x 3cm. The error rate for nanopore sequencing is higher than that for PacBio, yet incremental improvements are being made by ONT with regards to the pores and basecallers (Figure 1.3). The portability of the MinION sequencer is more closely aligned to the aspirations of biodiversity scientists of real-time sequencing in the field.

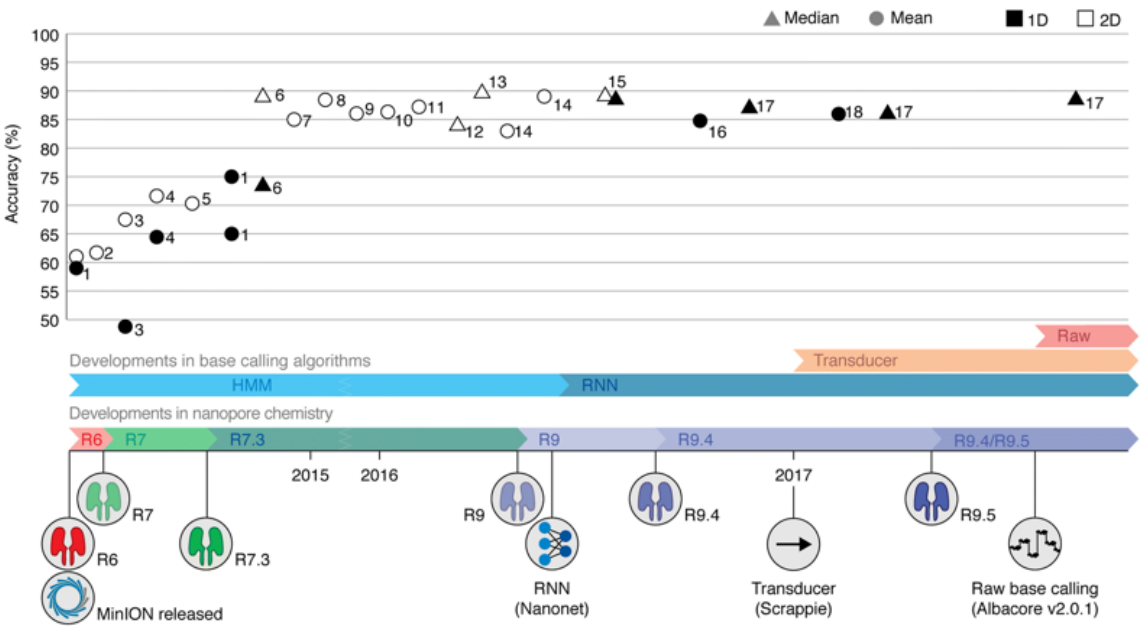


Figure 1.3: Incremental improvements made to nanopore sequencing technology in recent years (from Rang *et al*., 2018). As improvements to pores and base calling algorithms have been made over time the mean and median accuracy of both 1D and 2D reads have improved.

Monitoring and surveillance of aquatic ecosystems with diatoms

Diatoms are tiny single-celled micro algae around 2-200 micrometers in length whose cell wall (frustule) is made of silica. Their taxonomy is primarily based on their cell wall structure and is subject to much discussion and change (Kaczmarska et al., 2007). Despite this, diatoms and their assemblages

(communities) have been found to be useful in environmental studies and as a tool for monitoring water bodies as they respond to different levels of nutrients available, in particular phosphorus and nitrates. In recent years, the chloroplast gene *rbcL* (RuBisCO) has been used to discriminate diatom species due to its key function within respiration, with differing levels of success (Guo et al., 2015; Hamsher et al., 2011; Jones et al., 2005).

Diatoms are currently used in the United Kingdom as part of a suite of biological indicators used to aid decision making associated with the European Union (EU) Water Framework Directive (WFD) in rivers and lakes. Diatoms, along with other algae, which are attached to submerged stones and plant stems are referred to as “phytobenthos” and the EU WFD legislation requires that these are examined to inform the ecological status of the water bodies.

The current method in use by the Environment Agency (EA) uses light microscopy to determine abundance of diatom species, as described in European Standards (CEN, 2014b, 2014a) and the UKTAG partnership (http://www.wfduk.org/resources/rivers-phytobenthos). The abundance of each species is used to calculate the Trophic Diatom Index (TDI) (Kelly and Whitton, 1995). The Water Framework Directive requires that the condition of a water body is expressed as a ratio - the Ecological Quality Ratio (EQR) - using a value expected with no or minimal human impact as the denominator (Kelly et al., 2008). The EQR is ultimately calculated based on observed and predicted reference TDIs and the ratio is subsequently divided into five ecological status classes for reporting: High, Good, Moderate, Poor and Bad.

At the core of the TDI is the determination of the relative abundance of diatoms. However, this is a time consuming process, requiring experienced microscopists and highly skilled individuals to analyse and interpret the data. Given the amount of training involved for individuals, the time involved to analyse each sample, and the number of sites sampled, the microscopy method requires a substantial commitment of resources by the EA. In the current funding climate, alternative solutions which offer a similar level of resolution and precision at lower cost are particularly attractive.

Monitoring and surveillance of fungal species

A small number of plant pathogens have the potential to cause significant economic impact if they were introduced into the United Kingdom and allowed to become established. Such quarantine and regulated pathogens require control and phytosanitary methods to prevent their introduction and spread. A recent example was the 2012 introduction and subsequent spread in the UK of the fungal ash dieback pathogen *Hymenoscyphus fraxinus*. Despite having spread throughout mainland Europe during the previous 20 years from Poland where it was originally reported, the introduction to the UK gained widespread publicity in the national press. A nationwide survey of ash trees in 2012 concluded that the disease had likely been introduced to the UK through both the import of ash seedlings from continental Europe and by wind-borne spores. These conclusions led to the publication of a revised plant biosecurity strategy for Great Britain (DEFRA, 2014) and a wider realisation that surveillance for emerging threats should be a higher priority.

A large number of fungal phytopathogens produce spores with the ability to be spread large distances, for example *Cryphonectria parasitica* (chestnut blight), *Zymospetoria tritici* (wheat septoria blotch) and *Sclerotinia sclerotiorum* (causes various rots). While some fungal pathogens are ubiquitous and prevalent in the United Kingdom, others are not and present a significant risk to plant health if they were to be introduced. A small number of plant pathogens have the potential to cause significant economic impact if they were to be introduced into the United Kingdom and allowed to become established. Between 1970 and 2004, a total of 234 new plant pathogens were described in the UK (Jones and Baker, 2007). More recently, discovery in the UK of *Hymenoscyphus fraxinea* causing dieback in ash trees in the United Kingdom in 2012 demonstrated the importance of routine surveillance for early detection of known plant pathogens (Potter and Urquhart, 2016). Two main routes of entry were suspected in the introduction of ash dieback to the United Kingdom: imports of infected saplings and the airborne transport of spores from the continent (Lawrence and Cheffings, 2014). The outbreak highlights that more surveillance of airborne spores is required in order to improve plant biosecurity in the UK and to enable the early detection of outbreaks of existing or emerging phytopathogens. Phytopathogens requiring regulation that present a risk to plant health within the European Union are

described by the European Plant Protection Organisation (EPPO) on their lists of quarantine pests and pathogens (www.eppo.int).

Surveys of airborne spores have been carried out in the past and methods used to identify the species present have included ELISA (Flückiger et al., 2000), PCR (Calderon et al., 2002; Williams et al., 2001) and real-time PCR (Schweigkofler et al., 2004; Walsh et al., 2005). These methods are reliable and species-specific but have limited use outside of detecting one or two species at a time. Larger-scale methods such as microarrays (Lievens and Thomma, 2005) and denaturing gradient gel electrophoresis (DGGE) (Peccia and Hernandez, 2006) have also been used to examine spore samples but again these have their own methodological limitations. Recent advances in DNA sequencing technology with the introduction of next-generation sequencers have permitted the assessment and comparison of microbial communities by metagenomics amplification and sequencing of the 16S rDNA and the internal transcribed spacer (ITS) regions from bacterial and fungal communities, respectively.

In recent years the applications of fungal metabarcoding have been far reaching and aerosol biosurveillance studies have been common. The ability to collect spores over time with volumetric spore traps (e.g. Burkard traps) enables the spatial and temporal analysis of fungal populations in many different environments. Microbiome studies have been carried out on outdoor fruit and vegetable markets (Ahire and Sangale, 2012), inside living areas (Korpelainen and Pietilainen, 2015) and subway systems (Afshinnekoo et al., 2015). A large project to study the microbiomes of the built environment has been ongoing in recent years (Gilbert and Stephens, 2018) and has characterised homes, offices, hospitals, classrooms, zoos, farms, planes, the International Space Station, and water systems.

There is an important need for methodology which can measure and assess our environment in a cost-effective manner, without the loss of critical information. Modern DNA techniques offer the potential for such monitoring methods with the additional resolution for surveillance for new and emerging threats. This thesis has two main aims. Firstly, to assess Illumina metabarcoding for the replacement of microscopy in the statutory monitoring of water quality in the United Kingdom

and its use in surveillance of fungal plant pathogens. Part of this aim is to develop and implement robust and validated bioinformatics pipelines to enable the operational readiness of these methods. Secondly, to assess the newer long-read nanopore sequencing technology for its potential in enabling higher resolution species identification for more accurate monitoring and surveillance.

Chapter 2. Development of a high-throughput method for assessing the composition of diatom assemblages

Introduction

DNA barcoding is the Sanger sequencing of a standardised region of the nuclear or organellar genome, to catalogue and identify taxa. It was originally proposed for use with mammals and invertebrates (Hebert et al., 2004, 2003a) using the mitochondrial COI gene (Hebert et al., 2003b), but was rapidly adopted for use in fungi with ITS1 (Begerow et al., 2010; Schoch et al., 2012), and plants with a combination of *rbcL* and *matK* (Hollingsworth et al., 2009). The technique can be used to identify specimens independent of their life-stage but requires that only one taxon is present in the sample prior to DNA extraction, amplification and Sanger sequencing. This limitation was circumvented with the introduction of next-generation sequencing technologies, opening up the potential of identifying all the species in a community in a single high-throughput analysis known as metabarcoding (Hajibabaei et al., 2012, 2011; Pierre Taberlet et al., 2012). Further advances in recent years in the isolation of cellular material and DNA from the environment (eDNA) have created the potential for yet more applications of DNA-based species identification using metabarcoding approaches (Deiner et al., 2016; Hänfling et al., 2016; Minamoto et al., 2012; Rees et al., 2014; P. Taberlet et al., 2012). Metabarcoding has been applied to many types of environmental samples, including soil (Schmidt et al., 2013), air (Nicolaisen et al., 2017), and water. In the case of water, methods have been developed for the detection and identification of invasive fish species (Takahara et al., 2013) and invasive aquatic invertebrate species (Klymus et al., 2017).

The method used in the United Kingdom prior to this work used light microscopy to determine the relative abundances of diatom species (Kelly et al., 2008) using methods underpinned by European Standards (CEN, 2014a, 2014b). The Water Framework Directive requires that the condition of a water body is expressed as a ratio - the Ecological Quality Ratio (EQR) - using a value expected with no or minimum human impact as the denominator (Kelly et al., 2008). The EQR is

ultimately calculated based on observed and predicted reference metrics and the ratio is subsequently divided into five ecological status classes for reporting: High, Good, Moderate, Poor and Bad. The Trophic Diatom Index (TDI) (Kelly and Whitton, 1995) is the metric used in the United Kingdom to aid the calculation of the EQR for a waterbody. The TDI gives a score between 0 (very low level of nutrients in the waterbody) and 100 (very high level of nutrients in the waterbody) calculated from the relative abundance of each benthic diatom taxon present in the sample. Each taxon is assigned a weighting based on their nutrient tolerance and this weighting and the abundance of the taxon is used to calculate the final TDI. Until 2017, the calculation of the TDI was based entirely on light microscopy counts of diatom species. The current version of the TDI - referred to as TDI4 - is based upon microscopy alone. An updated version based upon the data produced during this project is referred to as TDI5.

At the core of the TDI is the need to determine the relative abundances of diatom taxa in a sample. However, this is a time-consuming process, requiring experienced microscopists to analyse and interpret the data. Studies have shown considerable variation even amongst experienced analysts (Kahlert et al., 2012). Given the amount of training involved for the microscopists, the time involved to analyse each sample, and the number of sites sampled, the microscopy method requires a substantial commitment of resource. In the current funding climate, alternative solutions which offer a similar level of resolution and precision at lower cost are particularly attractive.

Our aim in this study was to develop a high-throughput DNA metabarcoding method for the determination of diatom relative abundance in river biofilm samples. We also present a comparison of 500 samples assessed using both light microscopy and the metabarcoding method.

Materials and methods

Diatom sample collection

Diatom samples (n=500) were collected from UK rivers using standard Environment Agency sampling techniques for benthic diatoms (http://www.wfduk.org/resources/rivers-phytobenthos). The location for each

sample taken is shown in Figure 2.1. The sampling involved collecting 5 cobbles at each sampling point and placing them in a tray with 50ml of stream water and then brushing the upper surface of each cobble with a toothbrush to remove the biofilm (CEN, 2014a; Kelly et al., 2008). The samples were then transferred to the laboratory in a cool box. An aliquot (5ml) of the suspension of biofilm and water was then transferred using a Pasteur pipette to a sterile 15ml centrifuge tube containing 5ml nucleic acid preservative (3.5M ammonium sulphate, 17mM sodium citrate and 13mM EDTA). The samples were then frozen at -30°C prior to DNA extraction. The remainder of each sample was preserved using Lugol's iodine for morphological analysis by light microscopy.

Construction of a morphologically-verified database of full-length *rbcL* diatom sequences

Samples for traditional DNA barcoding of *rbcL* were collected from 61 locations in England and Scotland, encompassing a wide range of ecological diversity in order to establish a reference database of diatom *rbcL* DNA barcodes. A few drops of diatom suspension were placed in Petri dishes and individual cells of diatoms were isolated using a micropipette or by streaking onto 2-3% agar plates. Selected cells were then transferred into small volumes of freshwater medium in the wells of 96-well plates. After a few days of incubation the health and clonal nature of each culture was confirmed by observation using an inverted microscope. Successfully established clonal cultures were then grown in 90mm petri dishes for DNA extraction and preparation for a voucher slide.

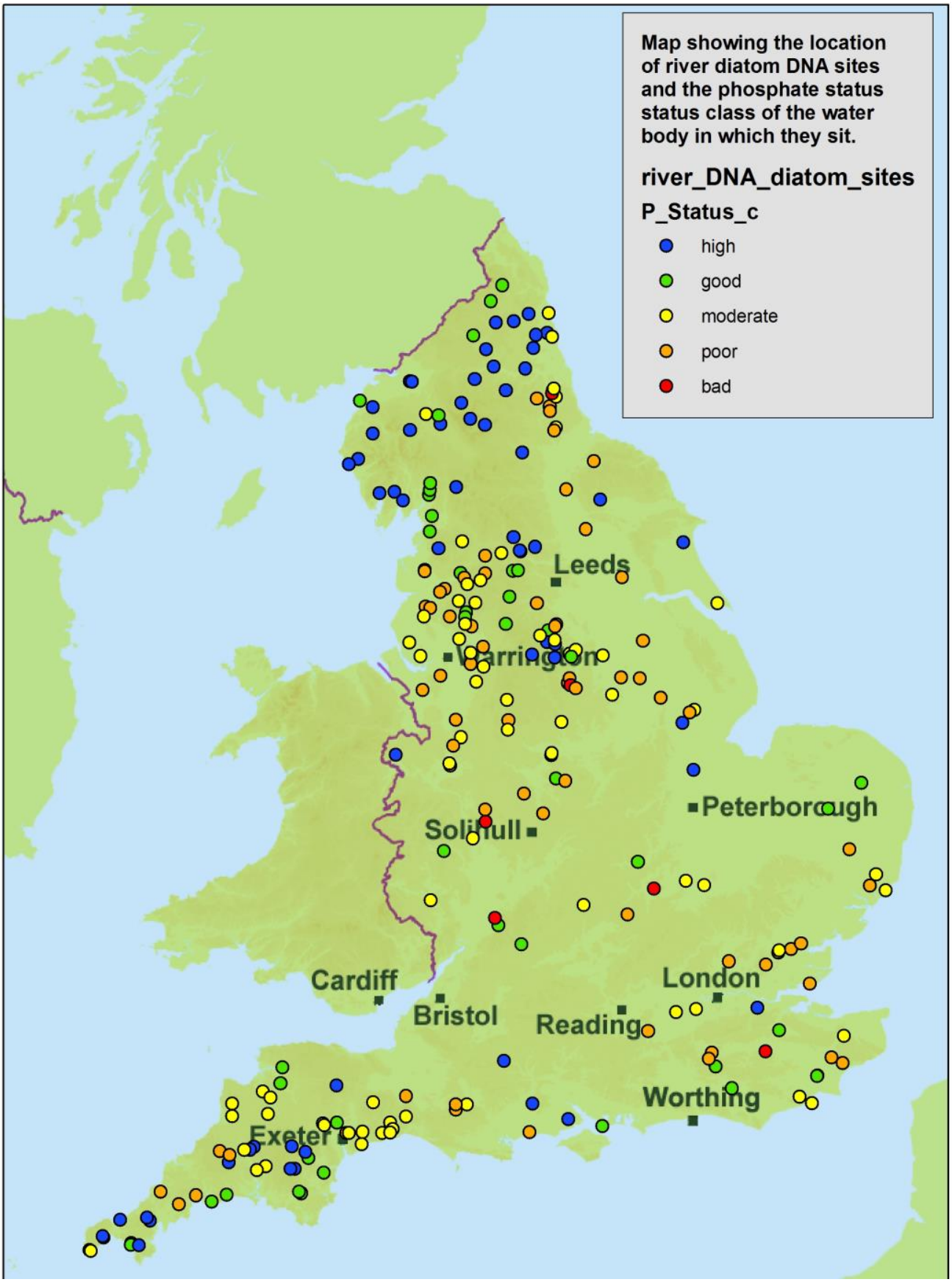


Figure 2.1: Map showing the location of diatom collection sites and the phoshpate status class of the waterbody. Blue = high quality; Green = good quality; Yellow = Moderate quality; Orange = Poor quality; Red = Bad quality

The resulting slurries of cells were collected in 1.5ml test tubes and centrifuged at 2000 x g for 10 mins. Extraction of DNA from each pellet was carried out using a QIAxtractor (Qiagen, UK). Amplification of a 1400bp fragment of the *rbcL* gene was carried out in 25µl reactions containing 10ng diatom DNA, 1mM dNTPs, 1x reaction buffer (Roche Diagnostics, Germany), 1U Taq DNA polymerase (Roche) and 0.5µM each primer (DPrbcL1: AAGGAGAAATHAATGTCT and DPrbcL7: AARCAACCTTGTGTAAGTCTC (Hamsher *et al.*, 2011). The following PCR protocol was followed: 94°C for 3 minutes, followed by 35 cycles at 94°C for 60 seconds, 55°C for 60 seconds and 72°C for 90 seconds, followed by a final extension at 72°C for 5 minutes. The quantity and length of the PCR products were examined by agarose gel electrophoresis by comparison with known size standards. PCR products were purified using ExoSAP-IT (USB Corporation, Ohio, USA). Sequencing was conducted in 10µl volumes using 0.32µM of the PCR primers or the sequencing primers NDrbcL5 (CTCAACCATTYATGCG) and DrbcL11 (CTGTGTAACCCATWAC) described in Jones *et al.*(2005). DNA barcodes were assembled using SeqMan (DNASTAR, Madison, WI).

Determination of conserved regions and primer design

The diatom DNA barcode database, totalling 1232 full-length *rbcL* diatom sequences comprising 390 species, was used to develop a short *rbcL* barcode suitable for high throughput NGS analysis. The diatom sequences were aligned with MAFFT (Katoh and Standley, 2013) using default settings. The diatom *rbcL* alignments were analysed using a Python script findAlignmentPrimers.py (archived at https://github.com/rachelglover/diatom-analysis) that identified regions having more than 96% sequence similarity where fewer than 4 sequences had a gap at that region; this prevented gaps from being mis-identified as highly conserved regions. To prevent the erroneous calling of sites as variable due to sequence errors or rare variants, single-base sites were called as variable in the highly conserved regions only when more than 5 of the sequences in the alignment were different at that position when compared to the most prevalent base. Highly conserved regions of the alignment were assessed with Primer3 (Untergasser et al., 2012) for primer design, including degenerate bases at the variable sites. When multiple candidate primers were identified for a region, selection of the best individual primer was based upon the lowest number of degenerate nucleotides (to minimise the amplification of non-target sequences)

and the highest percentage of sequence conservation of that primer against the original diatom *rbcL* alignment.

Estimation of the resolving power of the short *rbcL* barcode

Each potential short *rbcL* metabarcoding region was independently assessed for its ability to discriminate diatom species. DNA sequences from the region of the *rbcL* alignment under investigation were trimmed from the alignment. Operational Taxonomic Units (OTUs) were then picked with UCLUST (Edgar, 2010) from the simulated short barcode sequences with the similarity set to 100% to ensure that all unique variants at that location were used for analysis. The OTUs were then assigned to taxa using BLAST (Altschul et al., 1990) against the morphologically verified diatom DNA barcode database. The OTUs, sequence counts and taxonomic assignments were then used to calculate the number of sequences from that specific *rbcL* region which had been correctly assigned to each taxonomic level. This processing step was carried out using a custom script processOTUs.py (archived at https://github.com/rachelglover/diatom-analysis). The counts for each region were plotted using R v3.0.2 (R Core Team, 2017).

Preparation and analysis of diatoms by light microscopy

Samples for light microscopy were digested either with a mixture of sulphuric and oxalic acids, with potassium permanganate or cold hydrogen peroxide. Following digestion, samples were rinsed several times to remove all traces of the oxidising agent. Between rinses samples were centrifuged at between 3000 and 5000 rpm for 4-5 minutes or allowed to stand overnight to ensure that all diatoms had settled to the bottom of the tube. Permanent slides were prepared using Naphrax (Brunel Microscopes, Chippenham) as a mountant (Kelly et al., 2018). At least 300 TDI scoring valves of benthic diatoms on each slide were identified to the highest resolution possible using a Nikon BX40 microscope with 100x oil immersion objectives with phase contrast and their abundance recorded.

DNA extraction, amplification and sequencing of the short *rbcL* barcode

DNA extraction was carried out with the enzymatic lysis method (Eland et al., 2012). This method involved 5 hours of incubation with Proteinase K, followed by column purification using Qiagen DNeasy® Blood and Tissue kit according to the manufacturer's instructions. The quantity of DNA was estimated using a Qubit

fluorometer and dsDNA BR Assay kit following the manufacturer's instructions (Thermo Fisher Scientific, Cat: Q32850). Genomic DNA was stored at -30°C prior to PCR and sequencing.

Amplification of *rbcL* prior to Illumina sequencing was carried out using the following method. PCR reactions of 30µl containing 6µl of HF buffer (NEB, USA), 0.3µM *rbcL*-646F (ATGCGTTGGAGAGARCGTTTC), 0.3µM *rbcL*-998R (GATCACCTTCTAATTTACCWACAACTG), 0.3mM dNTPs, 0.3µl Phusion high-fidelity DNA polymerase (NEB) and 0.5µl of a 1:10 dilution of extracted sample DNA. The final reaction volume was made up with nuclease free water to 30µl. The following PCR protocol was followed: 98°C for 2 minutes, followed by 35 cycles at 98°C for 20 seconds, 55°C for 45 seconds and 72°C for 60 seconds, followed by a final extension at 72°C for 5 minutes. All PCR reactions were carried out on a C1000 thermal cycler (Bio-Rad, UK). Each run contained a number of negative controls including no-template controls, index PCR controls and extraction buffer controls. PCR products were visualised on 1% agarose gels. They were then purified using AMPure Beads and prepared for sequencing following the Illumina 16S Metagenomics Sequencing library preparation protocol. The final library fragments were then quantified using Picogreen (Lifetech, UK), measured on a Fluoroskan Ascent fluorimeter (Thermo Scientific, UK) and mixed in equal quantities to create a 20nM pool. This pool was assessed using a D1000 tapestation tape (Agilent, UK) prior to running in the presence of 10% PhiX (Illumina, UK) on an Illumina MiSeq with a V3 2x300 flow cell (Illumina, UK).

Quality control and sequence analysis

The data from each instrument run was analysed independently to mitigate against any intra- and inter-run variation that may have been introduced during PCR or library preparation. Negative controls (no-template PCR controls, index-PCR controls, extraction buffer controls) were also sequenced. Prior to the downstream bioinformatics analysis the sequence files for these controls were checked to make ensure they were blank.

Sequences from each sample were subjected to a very stringent quality control procedure, consisting of four main steps which are summarised in Figure 2.2.

Firstly, PCR amplification primers were removed from both sequenced strands of DNA using Cutadapt v1.9.1 (Martin, 2011). Secondly, poor quality 3' end of sequences from both strands were trimmed with Sickle v1.33 (Joshi and Fass, 2016) in paired-end mode. Thirdly, trimmed paired-end reads were joined to form a single consensus sequence using PEAR V0.9.6 (Zhang et al., 2014). Finally, a further round of quality assessment to remove any sequences with an overall accuracy of less than 99.9% using Sickle v1.33 (Joshi and Fass, 2016) in single-read mode. The QC procedure was automated in a custom script ampliconQC.py (archived at https://github.com/rachelglover/diatom-analysis). Remaining PhiX sequences were identified by mapping the good quality reads to the PhiX reference sequence (NC_001422) with bowtie2 (Langmead et al., 2009) and were removed from the dataset with bedtools (Quinlan and Hall, 2010).

High-quality sequences were clustered into operational taxonomic units (OTUs) with UCLUST (Edgar, 2010) at 97% similarity and the most abundant sequence in the cluster selected for the representative sequence, using QIIME v1.9.1 (Caporaso et al., 2010). Representative sequences were assigned to taxa following blastn against a reference database of full-length *rbcL* diatom sequences (http://github.com/rachelglover/diatom-analysis/diatoms.sequences.FINAL2017.fasta) with an e-value threshold of 0.01. Once completed, a similarity threshold of 95% for each BLAST identification was applied (http://github.com/rachelglover/diatom-analysis/create_taxonomy_assignments_from_blast.py) with those sequences with hits below 95% similarity being described as having no specific identification. Relative abundance calculations were carried out within QIIME v1.9.1 (Caporaso et al., 2010) and the Trophic Diatom Index (TDI) was calculated (Kelly et al., 2008).

Method reproducibility and repeatability

Four field samples were randomly selected for use in the reproducibility and repeatability experiments (samples 114061 (River Ehen), 114078 (River Derwent), 114092 (River Team) and 114161 (River Wear)). Inter-individual reproducibility was tested by two different staff members carrying out the PCR and clean-up steps of the sequencing protocol on all four samples. Each sample was amplified in triplicate to test the repeatability of amplification from the same

DNA extract. To test for inter-instrument reproducibility, the sequencing was completed with the same library preparation split between two Illumina MiSeq instruments: one at Fera Science Ltd (York, UK) and one at NewGene Ltd (Newcastle, UK). Initial analysis was carried out with the bioinformatics pipeline described above. Following this, beta diversity was calculated with the Bray-Curtis dissimilarity metric. Adonis (Anderson, 2001) was used to assess the variance between the OTU composition of the four field samples for each experiment, totalling 56 sequencing samples.

Comparison of light microscopy and metabarcoding

Non-metric multidimensional scaling (NMDS) (McCune et al., 2002) was used to investigate the structure of the LM and NGS datasets using R (R Core Team, 2017) with the vegan package (Oksanen et al., 2007) for multivariate analyses. The similarity in structure was tested using a Procrustes analysis and the associated permutation test (Peres-Neto and Jackson, 2001) in the vegan package, and also by scatterplots and computation of Pearson's correlation coefficient. Calculation of the Trophic Diatom Index v.4 (TDI4) values was carried out with the DARLEQ2 software (http://www.wfduk.org/resources/category/biological-standard-methods-201).

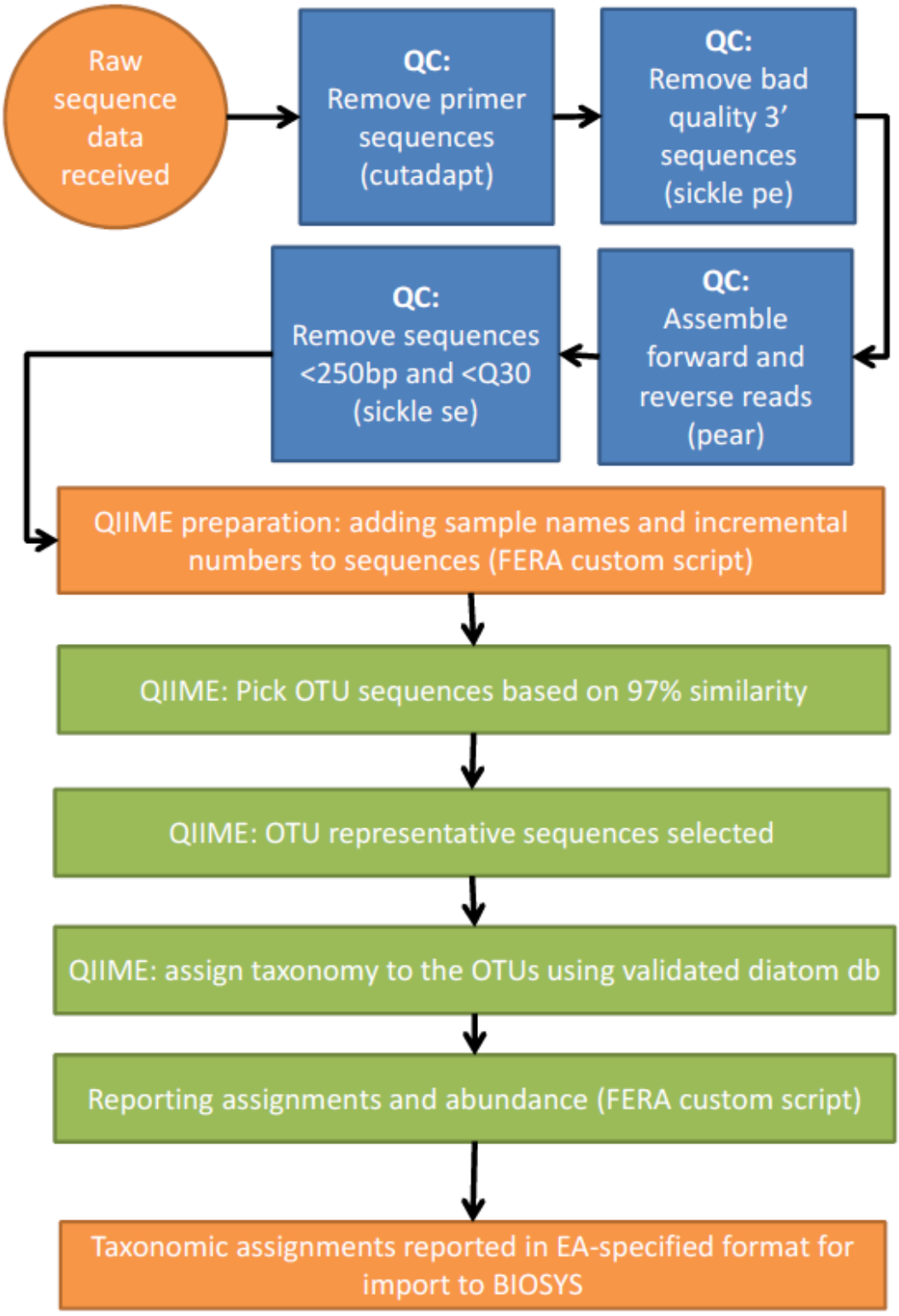


Figure 2.2 Quality control and QIIME pipeline for analysis of diatom NGS data. Notes: db=database; EA=Environment Agency; FERA=Food and Environment Research Agency; pe=paired end; QC=quality control; se=single end; BIOSYS=EA databases for storing, manipulating and reporting data from freshwater and marine biological surveys.

Results

Determination of conserved *rbcL* regions and primer design

In total, 11 regions of *rbcL* were identified as having more than 96% identity across the length of the alignment of full-length *rbcL* sequences (Figure 2.3). As such, these regions were determined to be suitable for conserved primer design.

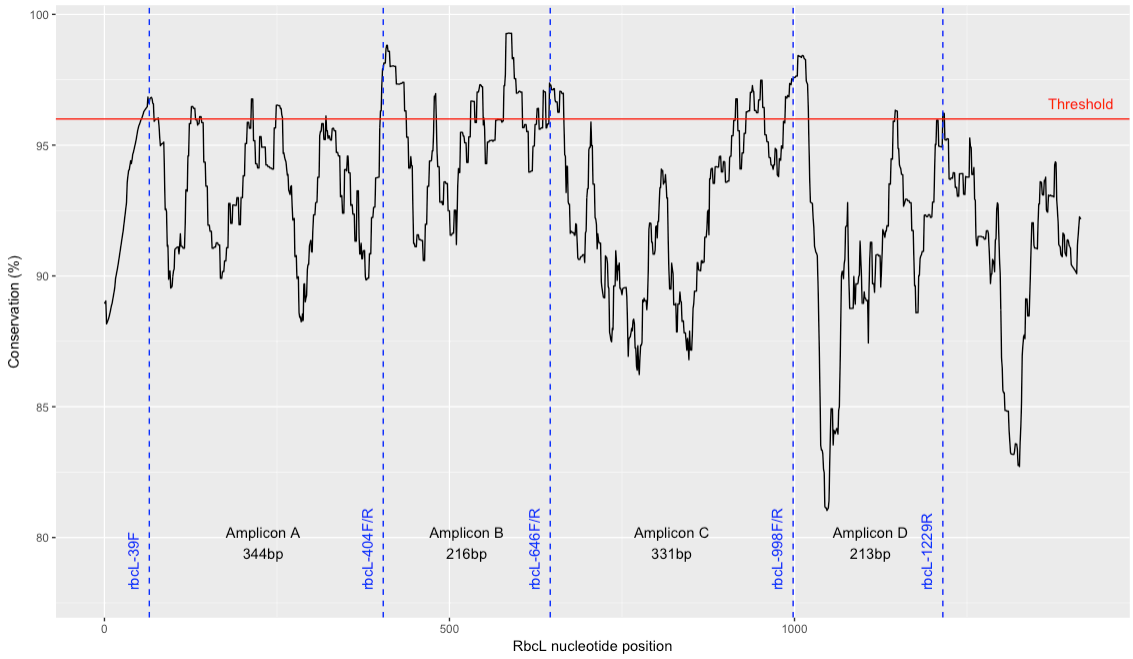


Figure 2.3: The conservation of nucleotides at each position in an alignment of full-length *rbcL* diatom sequences. The red line shows the threshold used to identify conserved regions suitable for primer design. Blue dashed lines show the location of selected primer target sequences. Amplicons suitable for short-read Illumina sequencing are also shown.

Primers were designed to amplify regions along the *rbcL* gene that showed good potential for species discrimination (Figure 2.3). The locations of the four predicted amplicons (A-D) are shown in Figure 4, with the predicted amplicon length varying from 213 bp to 344 bp.

Estimation of the resolving power of the short *rbcL* barcode

The sequences of the four predicted amplicons, A-D, were subjected to an *in silico* analysis to determine the resolving power of each region. The numbers of

sequences that could be correctly assigned by the analysis pipeline to class, family, genus, species and isolate were calculated for each predicted amplicon region (

Table 2.2). For example, a count of 1 for the taxonomic level ‘class’ means that ‘class’ was the lowest taxonomic level at which an accurate taxonomic classification could be made for that sequence using that amplicon. Based on this we used the sum of the ‘species’ and ‘isolate’ counts as an assessment of the efficacy of a particular amplicon to be used for species-level taxonomic assignments.

The taxonomic assignments in

Table 2.2 demonstrate that all amplicon regions could be used to provide an adequate number of species level assignments for the 390 diatom sequences present in the original database. As the diatom metric Trophic Diatom Index (TDI) estimation is based upon species-level discrimination, the numbers of correct species- and isolate-level identifications were plotted against the lengths of the amplicons (Figure 2.4). A small number of sequences in amplicons A, B and D could not be assigned to any taxon at any taxonomic level as the reference database sequences used in the alignment did not cover these regions and so no identification could be made for those amplicon sequences.

Table 2.1: Primers designed to amplify four regions of *rbcL* shown to be suitable for short-read metabarcoding

Primer name	Sequence (5' -> 3')
rbcL-39F	TGW-CCG-TTA-CGA-ATC-TGG-TG
rbcL-404F	CWG-CDT-TAC-GTT-TAG-AAG-ATA-TGC-G
rbcL-404R	CGC-ATA-TCT-TCT-AAA-CGT-AAH-GCW-G
rbcL-646F	ATG-CGT-TGG-AGA-GAR-CGT-TTC
rbcL-646R	GAA-ACG-YTC-TCT-CCA-ACG-CAT
rbcL-998F	CAG-TTG-TWG-GTA-AAT-TAG-AAG-GTG-ATC
rbcL-998R	GAT-CAC-CTT-CTA-ATT-TAC-CWA-CAA-CTG
rbcL-1229R	ATW-GTA-CCA-CCA-CCC-AAC-TGT-A

Table 2.2: Amplicons assessed *in silico* for their ability to assign amplified sequences to species-level taxonomy. The numbers in the green boxes denote the number of sequences which could be correctly assigned by the pipeline to each taxonomic level.

Amplicon	Forward primer	Reverse primer	Amplicon length (bp)	Lowest taxonomic level where the taxonomic assignment was correct					
				Class	Family	Genus	Species	Isolate	No identification
A	rbcL-39F	rbcL-404R	344	2	0	21	204	156	7
B	rbcL-404F	rbcL-646R	216	2	0	37	202	142	7
C	rbcL-646F	rbcL-998R	331	2	0	22	201	165	0
D	rbcL-998F	rbcL-1229R	213	2	0	26	202	151	9

The sequences for amplicon C produced the largest number of species- and isolate-level identifications and lowest number of higher-taxonomy identifications. PCR annealing temperature studies between 50°C and 60°C (data not shown, but available in the EA report, Appendix I) showed that the primers designed to amplify amplicon C gave the best performance, amplifying an intense band of the correct predicted size across the full range of annealing temperatures tested. In contrast primers for amplicons A, B and D generated mis-priming products of the incorrect sizes at temperatures below 58°C and faint bands of the correct predicted size for amplicons B and D. Amplicon C, amplified using primers *rbcL*-646F and *rbcL*-998R, was used for all subsequent sequencing and method assessment.

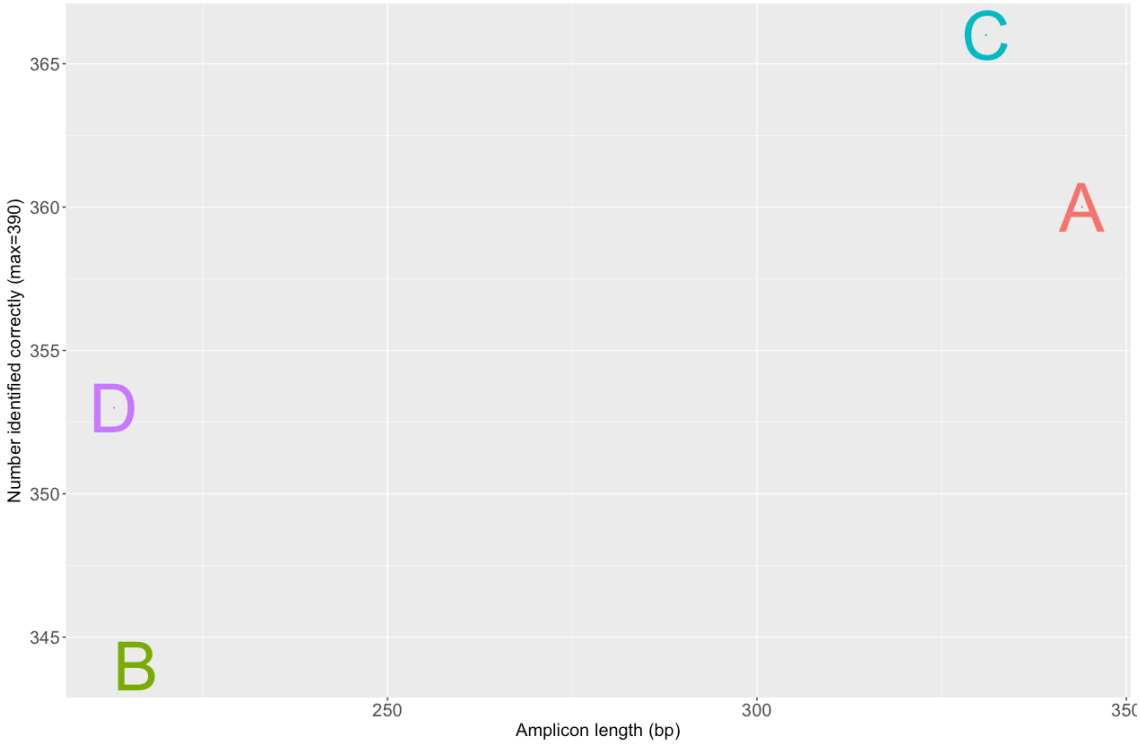


Figure 2.4: Correct species-level taxonomic assignments plotted against the length of the amplicon, showing that amplicon C provides the highest number of sequences correctly identified to species when tested *in silico.*

Method reproducibility and repeatability

The results in Table 2.3 can be used to draw a number of conclusions about the reproducibility of the method. The low R^2 values paired with the very high *p* values lead to the conclusion that there are no significant differences between the samples when split by staff member and by different MiSeq instruments. In

contrast, when the same test is applied to split the samples themselves as a control, the R^2 values are high and the differences are significant ($p < 0001$).

Table 2.3: Inter-individual and inter-machine reproducibility statistics, assessed using adonis (Anderson, 2001)

Experiment	adonis results (R^2)	adonis result (*p* value)
Inter-individual reproducibility	0.00539	0.994
Inter-machine reproducibility	0.00405	0.997
Control (diverse samples)	0.79659	0.001

The low number of PCR replicates (n=3) precluded the use of adonis to assess the reproducibility of replicates by staff member as the number of replicates and diversity between replicates was too low to detect and evaluate the differences. The relative abundance of each species detected is therefore shown visually in Figure 2.5. No significant differences were detected between staff members, PCR replicates or separate sequencing instruments when the same diatom samples were processed.

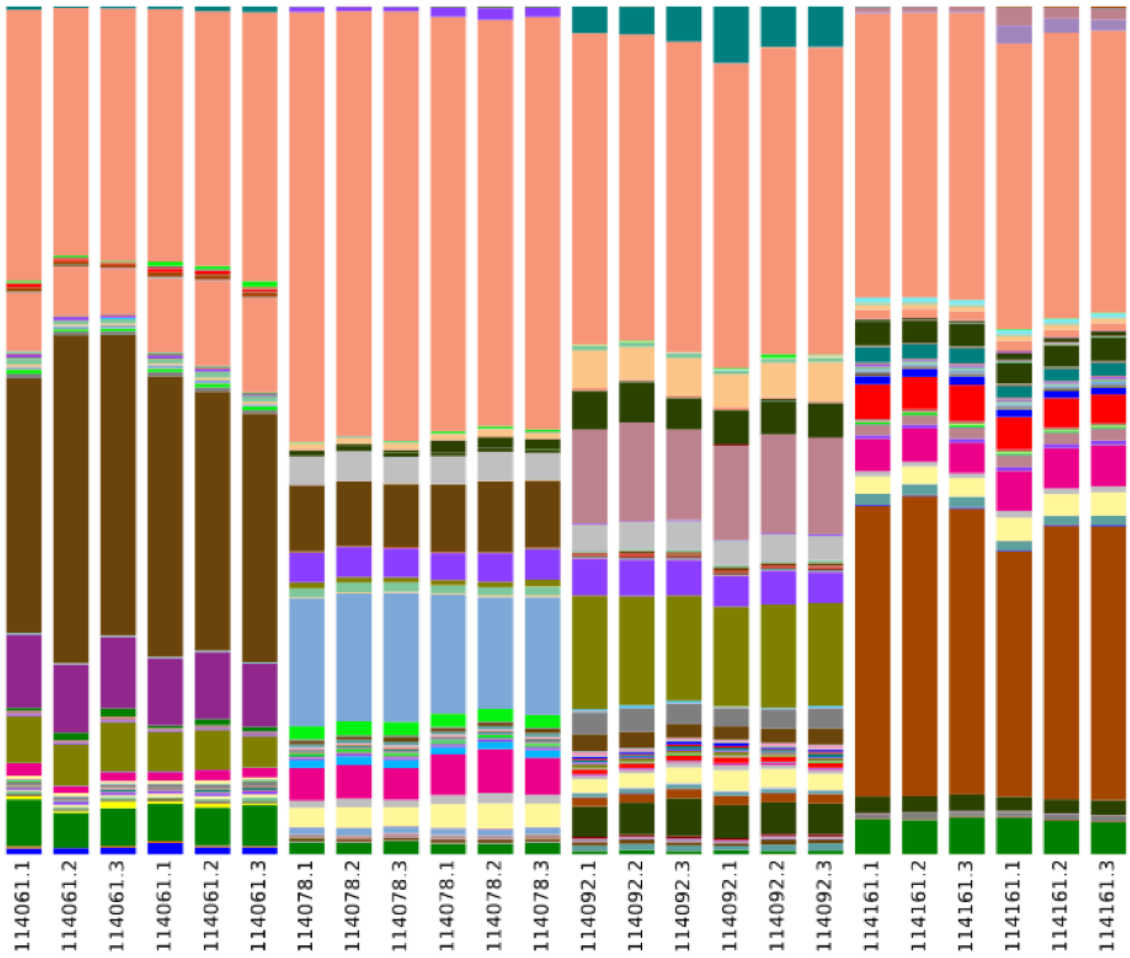


Figure 2.5: Stacked bar chart showing the relative abundances of diatom DNA sequences amplified for each of the four field samples with 3 PCR replicates each (1,2,3) performed by two different people. Each colour represents a different diatom species detected.

Comparison between the TDIs produced from light microscopy and metabarcoding

After taxonomic harmonisation of the species names the light microscopy (LM) and metabarcoding datasets contained a total of 493 and 306 benthic taxa respectively. Figure 2.6 shows the distribution of total number of sequence reads once planktic taxa had been excluded and the distribution of the relative abundances of unassigned reads in the metabarcoding dataset. The average number of reads was 41,048 per sample with over half the total read count in 354 samples not being assigned to a taxa within the *rbcL* reference database.

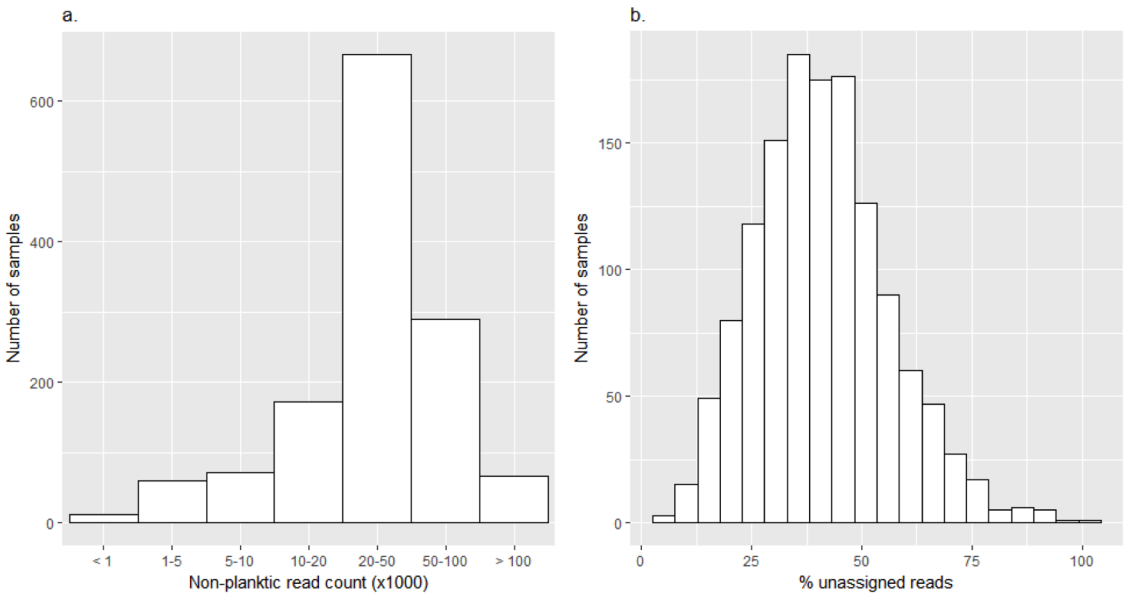


Figure 2.6: a. Number of reads of non-planktic taxa in the NGS dataset; and b. Proportion of reads that could not be assigned to taxa present in the *rbcL* reference database

The compositions of the LM and metabarcoding datasets were broadly similar with *Achnanthidium minutissimum* type having the highest maximum relative abundance in both methods and being the most frequently recorded. There were, however, a number of differences in details: *Melosira varians,* for example, was both more frequently recorded and occurred at a higher relative abundance in the metabarcoding dataset compared to LM samples, whilst the opposite was true for *Platessa conspicua*. *Luticola ventricosa* and *Lemnicola hungarica* occasionally occurred in high numbers in the metabarcoding results but are unlikely to be missed by LM analysts. A discrepancy also occurred for the genera *Fistulifera* and *Mayamaea;* in both cases the maximum abundance recorded was higher in LM than in metabarcoding.

Following the initial comparisons of the distribution of species within the LM and metabarcoding datasets, both were then subject to NMDS ordinations in order to examine the consequences of any differences on the structure of the datasets. The success of NMDS is given by the stress, which quantifies the agreement between the 2D representation and original dissimilarities where values less than 0.1 represent a good ordination that inferences can be drawn from, where values 0.1-0.2 represent ordination that is useable with caution and values >0.3 indicating that the ordination may be misleading (McCune and Grace, 2002). NMDS yielded ordinations with low levels of stress (LM: 0.17; metabarcoding:

0.18) that faithfully represented the original inter-sample dissimilarities. The two ordinations showed similar structure in terms of the first axes of each being strongly correlated (Pearson correlation coefficient, r=0.87, Figure 2.7a) and for the correlation between the first two axes assessed by a procrustes analysis ($p<0.001$, 999 permutations). Moreover, the first axis of the NMDS based on LM was strongly (negatively) correlated with TDI4 (Pearson correlation coefficient, r=-0.94, Figure 2.7b).

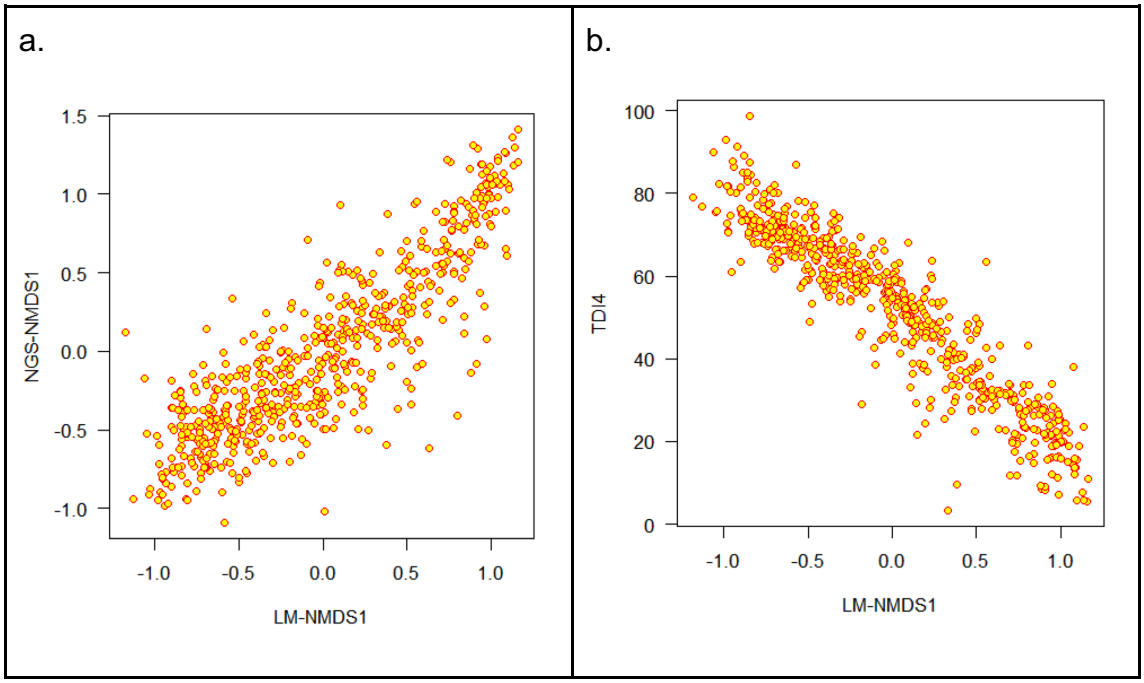


Figure 2.7: a. Comparison of the first axes of NMDS ordinations performed using LM and metabarcoding data. Pearson correlation coefficient, r=0.87. b. Axis 1 of NMDS of LM data versus TDI4 (r=-0.94)

TDI4, when calculated using the metabarcoding data was strongly correlated with the TDI4 calculated using light microscopy data (Figure 2.8) but the line deviated from 1:1 (Lin's concordance correlation coefficient: 0.81), with many metabarcoding analyses returning higher values from the same sample than LM when the TDI was low and moderate. This may reflect the generally high numbers of *Achnanthidium minutissimum,* which has a high LM:metabarcoding ratio in low nutrient (low TDI) sites, and higher numbers of taxa such as *Navicula lanceolata* and, in particular, *Melosira varians,* which have much lower LM:metabarcoding ratios.

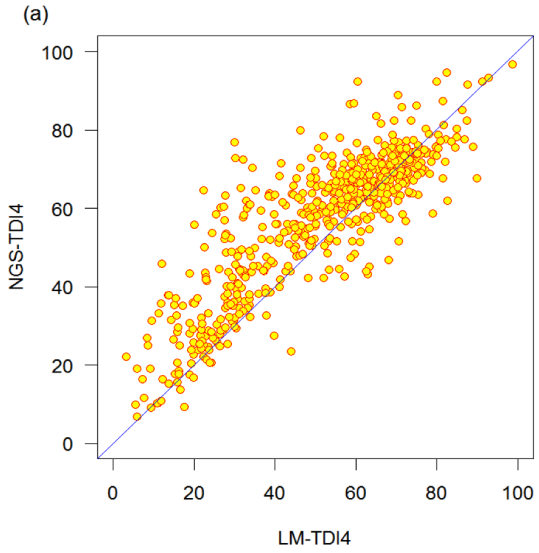


Figure 2.8: Comparison between the TDI calculated with light microscopy and metabarcoding data (r=0.86, Lin's r=0.81)

Discussion

We demonstrate that DNA metabarcoding can produce similar results to those obtained from traditional light microscopy for diatom assemblage analysis, offering the possibility of cheaper and more rapid assessment of water quality.

Application of metabarcoding approaches to diatom community analysis in the past has been limited (Kermarrec et al., 2014; Visco et al., 2015; Zimmermann et al., 2015). Kermarrec *et al.,* (2014) investigated the use of 18S rDNA, *rbcL* and cytochrome oxidase I (COI) using 454 pyrosequencing to assess the ecological quality of rivers. They found that *rbcL* was the most useful of the three markers for molecular identification of diatoms, a finding previously noted in DNA barcoding studies (Hamsher et al., 2011). However, the 454 pyrosequencing technology used in these studies is no longer available. Both Visco et al. (2015) and Zimmerman *et al.* (2015), utilised 18S rDNA for their comparisons between light microscopy and metabarcoding, albeit with smaller numbers of samples. Based on a much larger dataset of 500 samples covering water bodies from across the United Kingdom (Figure 2.1), we have demonstrated that a short amplicon of 331bp can be used to accurately determine relative abundances of diatom species present in a sample and that estimates of assemblage structure follow similar trends in both light microscopy and metabarcoding. However, there

were certain species-specific biases that will require further investigation and elucidation. In some cases, differences between light microscopy and metabarcoding may be easily explained by the presence of gaps in the barcode reference database; for example, we were not able to obtain and sequence *Achnanthidium pyrenaicum* or *Gomphonema clacifugum,* which were both present in the light microscopy. Unfortunately, the chemical processes required to prepare the diatom frustules for light microscopy are harsh and DNA isolation from the slides would have been a futile effort. Other differences between the light microscopy and metabarcoding are more difficult to explain. For example, *Luticola ventricosa* and *Lemnicola hungaria* occasionally were estimated in higher numbers by metabarcoding than by light microscopy, or the opposite situation where *Fistulifera* species and *Mayamaea* species showed higher abundance in the light microscopy results. Zimmermann *et al*. (2015) also noted differences where their metabarcoding method almost always led to a higher number of identified taxa. However, subsequent reanalysis of their light microscopy results confirmed the metabarcoding outputs. This demonstrated that the light microscopy methods require considerable expertise in order to accurate identify diatoms to species. Because our method targets an *rbcL* amplicon located in the chloroplast genome, results could be influenced by the diatom size (Vasselon et al., 2017b) as well as by the number of chloroplasts, and by the number of copies of *rbcL* per chloroplast, all of which may vary between species or genera. This could be negated by the use of species- or genus-specific weightings within the calculation of the Trophic Diatom Index.

Our results demonstrate that longer amplicons produce a greater number of accurate species-level diatom identifications. We applied a 97% identity threshold when clustering OTUs as it provided adequate clustering of diatom sequences but also because it reduced the time required to analyse the entire dataset. Recently the term Amplicon Sequence Variants (ASVs) has been introduced for community studies (Callahan et al., 2015), with the recommendation that they replace OTUs (Callahan et al., 2017). Indeed, the latest version of the popular microbial ecology software QIIME (https://www.qiime2.org) no longer supports the use of OTUs and analyses individual sequences. We support the introduction of such high-resolution analyses and the phylogenetic insights they may elucidate. However, in our study, taxonomic identification of each individual

sequence for datasets comprising 500 samples each, averaging more than 40,000 sequence reads would currently be unfeasible outside of specialist computing facilities.

When evaluating the four different *rbcL* amplicons for use in elucidating diatom community structure from biofilm samples with the current Illumina short-read technology, we were mindful that the resulting amplicon and method were required to be high-throughput, contamination-free, and with a rapid bioinformatics pipeline. These considerations were required to ensure the method would be sufficiently robust and cost-effective to enable the UK Environment Agency to deliver ecological assessments on thousands of samples. Illumina short-read sequences have an increasing error-rate along the reads from 5' to 3' (Schirmer et al., 2015). As such, when amplicon sequences are overlapped and a consensus sequence produced, there is a small but significant risk of single nucleotide errors being introduced in the middle of the consensus sequence if the amplicon is too long, or quality control measures to remove low quality 3' bases are less than ideal (Schirmer et al., 2016). In applications such as ours, this carries the downstream implication of falsely increasing the diversity of the sample and subsequently increasing the potential for OTUs to be misidentified (Wen et al., 2017). As the TDI4 requires relative abundance estimates of species (Kelly et al., 2008), misidentified OTUs could have an impact on the overall TDI and subsequent water quality assessment. Each of the four amplicons evaluated here provided adequate length for species discrimination to varying degrees. It is likely that the limitations of Illumina sequencing for metabarcoding studies (short read length, error profile) may be negated with the introduction of nanopore sequencing technologies such as the MinION™ (Oxford Nanopore) which will allow the amplification of whole genes - and perhaps complete organelle genomes - for the assessment of communities.

Chapter 3. Efficacy of MinION sequencing to compare the diatom communities in three rivers

Introduction

There are many characteristics of diatoms and their assemblages which have made them useful in environmental studies and as a tool for monitoring water bodies. Diatoms contribute ~20% of global fixation of carbon dioxide and are the largest primary producers in water bodies, according to Hildebrand (Hildebrand, 2008) who also states that diatoms contribute 50% of the primary productivity of oceans. Diatoms have short lifespans and a quick response to a number of environmental disturbances. They are more sensitive to changes in nutrients and contamination with organic matter than fish, macrophytes and macroinvertebrates (Hering et al., 2006). They have been reported as being very sensitive to both organic toxicants (De Jonge et al., 2008; Hirst et al., 2002; Morin et al., 2016). The sensitive response of diatoms to organic matter, in particular to the nutrients nitrate and phosphate, is especially useful to the monitoring of rivers and water bodies where organic runoff from farming and effluent from water treatment plants and industrial processes may affect water quality (Morin et al., 2016; Stevenson, 2014; Stevenson et al., 2010).

The overarching aim of the Water Framework Directive is to protect water bodies and use ecological status as the assessment by which the protection is measured (Water Framework Directive, 2000). The ecological statuses used to classify water bodies within Europe are High, Good, Moderate, Poor and Bad. The Trophic Diatom Index (TDI) is a metric used in the United Kingdom to aid the calculation of an Ecological Quality Ratio (EQR) for a waterbody which determines classification, and therefore ecological status (United Kingdom Technical Advisory Group (WFD-UKTAG), 2014). Until 2017, the calculation of the TDI was based entirely on light microscopy counts of diatom species. We recently developed a cost-effective and robust Illumina metabarcoding method to replace light microscopy for determining the Trophic Diatom Index (TDI) of water bodies in the United Kingdom (Kelly et al., 2018). The metabarcoding method was carried out in parallel with the light microscopy for all monitoring by the

Environment Agency (EA) during 2016-2017 before full adoption of the metabarcoding method in 2017. While the TDI is used in the United Kingdom for calculation of the EQR and ecological status, different metrics exists in other EU member states (Almeida et al., 2014; Feio et al., 2009; Toudjani et al., 2017). Other groups within the EU have been actively working towards molecular methods for their water quality monitoring with diatoms (Kermarrec et al., 2014; Morin et al., 2016; Toudjani et al., 2017; Vasselon et al., 2017a; Visco et al., 2015; Zimmermann et al., 2015). However, none have been implemented at the same scale and level of operational readiness as has occurred in the United Kingdom.

Although the current Illumina metabarcoding method is being used by the Environment Agency, it does have its drawbacks as assessed in Chapter 1. Markers located in the chloroplast genome are subject to copy number variation within and among species, which may be associated with diatom size; therefore use of chloroplast genes such as *rbcL* is less quantitative than nuclear genes that occur at a fixed copy-number per cell. A further limitation is the restriction on amplicon length by the Illlumina short-read sequencing technology itself, with the amplicon used for diatom species identification being only 331bp. The region of *rbcL* was chosen such that it could be amplified in most, if not all, diatom species, yet provide enough sequence variation to discriminate species (Kelly et al, 2018).

The introduction of long-read nanopore sequencing by Oxford Nanopore Technologies (ONT) in June 2014 led to an explosion of applications. It has been used to scaffold bacterial genomes (Karlsson et al., 2015; Wick et al., 2017), for monitoring and surveillance of disease outbreaks (Quick et al., 2016, 2015; Walter et al., 2017) and, more recently for 16S bacterial community analysis (Benítez-Páez et al., 2016; Kerkhof et al., 2017; Shin et al., 2016) and real-time DNA barcoding in the field (Menegon et al., 2017; Pomerantz et al., 2018, 2017). While the portability of the ONT MinION sequencer is attractive for ecological studies in the field, the technology itself offers the ability to sequence much longer amplicons, albeit with an increased error rate.

This chapter aims to compare and evaluate three methods for the assessment of diatom species composition (light microscopy, Illumina metabarcoding and MinION metabarcoding) by the use of the methods to compare three rivers in

northern England. A discussion of MinION metabarcoding for the purposes of water quality assessment is undertaken and a comparison to the Illumina metabarcoding method developed in Chapter 1.

Materials and methods

Sample selection

Three rivers in the north east of England (Table 3.1) were chosen for analysis and each was sampled in three locations along the water course. The river Ehen (Cumbria) has high water quality and is a special area of conservation, the river Wear (County Durham) has good water quality and the river Derwent (County Durham) has moderate water quality. Cobbles in the rivers had been sampled for diatoms previously as part of an Environment Agency (EA) project to develop an Illumina metabarcoding method for water quality classification (Kelly et al, 2018). Each river was sampled in three locations, diatom species counted by light microscopy and DNA extracted as described in Kelly *et al*. (2018). The same DNA extracts were used for Illumina short *rbcL* barcode sequencing and MinION long *rbcL* barcode sequencing as described below.

Sequencing

Sampling and Illumina sequencing of nine diatom samples was carried out as described previously in Chapter 1. Amplification of *rbcL* from the same nine diatom samples with primers DPrbcL1 (5'-AAGGAGAAATHAATGTCT-3') and DPrbcL7 (5'-AARCAACCTTGTGTAAGTCTC-3') (Jones et al., 2005) and MinION sequencing of the ~1500bp products were carried out at FERA Science Ltd. Following sequencing 2D fastq reads were extracted from the MinION fast5 files using poretools v0.6.0 (Loman and Quinlan, 2014).

Illumina short *rbcL* barcode taxon assignment

Quality control, OTU clustering and taxon assignment for the Illumina *rbcL* sequences were performed with the previously developed diatom *rbcL* pipeline (Kelly *et al.,* 2018). Briefly, bad quality 3' ends of sequences were trimmed, forward and reverse sequences were merged and sequences with quality <Q30 also removed. The good quality Illumina reads were then clustered into

operational taxonomic units (OTUs) at 97% similarity with UCLUST (Edgar, 2010) and the most abundant sequence in the cluster selected to be the representative with QIIME 1 (Caporaso et al., 2010). The representative sequences were then assigned taxonomy by searching the diatom barcode reference database with blastn (Altschul et al., 1990) with an e-value threshold of 0.01. Relative abundances of each diatom species within each sample were also calculated within QIIME.

MinION long *rbcL* barcode taxon assignment

Taxon assignments for every MinION read in each sample were carried out with blastn (Altschul et al., 1990) with an e-value threshold of 0.01 against the diatom barcode reference database created during Environment Agency project SC140024 (Kelly *et al*., 2018). The taxon assignment and relative abundance calculations were automated using Enviropore v0.6 (Glover, 2018). The length of the sequence and the assigned taxonomy, along with the percentage identity with the reference sequence were appended to the sequence name in the FASTA file to aid downstream sequence analysis. Sequences where an identification could not be made with the diatom reference database were extracted from the sample dataset and searched against the NCBI NR protein database with blastx (Altschul et al., 1990) and subsequently visualised with MEGAN community edition (Huson et al., 2016).

Comparison of three rivers, light microscopy, Illumina sequencing and MinION sequencing

Statistical comparisons between the three rivers and three identification methods were carried out in R (R Core Team, 2017) with the packages ape (Paradis et al., 2004), vegan (Oksanen et al., 2007) and picante (Kembel et al., 2010). The Bray-Curtis dissimilarity metric (Bray and Curtis, 1957) was used to calculate community and sample dissimilarities. Non-metric multidimensional scaling (NMDS) was calculated with the metaMDS function within vegan and PERMANOVA analysis (Anderson, 2001) was carried out with the adonis function within vegan. *Gomphonema* and *Achnanthidium* reference *rbcL* sequences were aligned with MUSCLE (Edgar, 2004) and a neighbour-joining tree produced using the Kimura-2-parameter model (Kimura, 1980) within the software package Seqotron (Fourment and Holmes, 2016).

Results

MinION and Illumina sequencing data composition

From the nine samples from three rivers (Figure 3.1), there were a total of 53,507 MinION reads and 1,819,888 Illumina reads produced, which are broken down by sample and location in Table 3.1.

Table 3.1: Number of Illumina and MinION sequence reads for each sample from the rivers Ehen, Wear and Derwent along the grid references for each location.

Sample	River	Grid Reference	MinION reads	Illumina reads
114058	Ehen (Scout Camp)	NY 087 153	5,009	118,689
114061	Ehen (Mill, footbridge)	NY 081 152	4,388	124,492
114064	Ehen (Oxbow)	NY 072 157	2,714	209,118
114069	Wear (Stanhope)	NY 991 392	9,654	203,779
114072	Wear (Frosterley)	NZ 036 369	7,295	218,554
114075	Wear (Wolsingham)	NZ 075 369	7,963	271,808
114078	Derwent (Ebchester)	NZ 101 556	5,701	218,561
114081	Derwent (Low Westwood)	NZ 111 565	6,017	222,928
114084	Derwent (Blackhall Mill)	NZ 122 569	4,766	231,959

Figure 3.1: (A) Location of the three rivers in the United Kingdom; (B) River Ehen sampling locations; (C) River Derwent sampling locations; (D) River Wear sampling locations

The distribution of MinION sequence lengths by river are shown in Figure 3.2 and show a distribution of read lengths rather than a tight group around the expected amplicon size, as was observed in Illumina amplicon data (not shown). No difference in sequence length distribution is observed in any of the rivers.

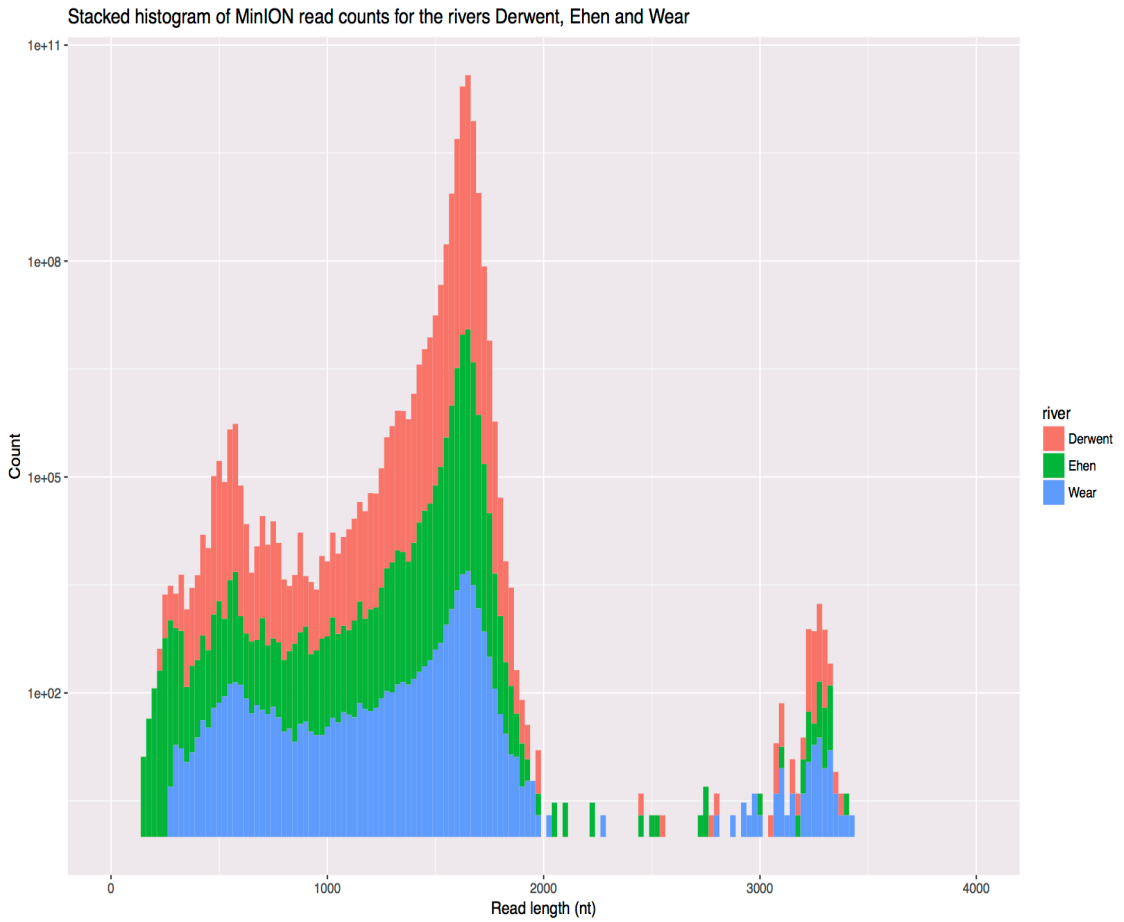


Figure 3.2: Stacked histogram of the MinION read lengths by river, with the counts shown as a log scale to demonstrate the abundance of read lengths. The largest peak is observed at a length of 1800 nucleotides, as expected, with smaller peaks at approximately 500nt and a small number of sequences with length >3000nt.

Comparison of the three methods used to determine relative abundance within diatom assemblages in three rivers

Following taxon assignment the relative abundance of each species within each sample was calculated. The relative abundances when determined by light microscopy were also included (M. Kelly, personal communication) and the results are shown aggregated for each river in **Error! Reference source not found.**. The largest differences between the next-generation sequencing metabarcoding methods and light microscopy (LM) were due to missing reference DNA barcodes for 10 species which were identified in the LM samples: *Achnanthidium caledonicum, Achnanthidium microcephalum, Adlafia suchlandtii, Aulacoseira* sp., *Cyclotella comensis, Encyonema lange-bertalotii, Gomphonema calcifugum, Gomphonema gracile, Gomphonema olivaceum* and *Synedra*

tenera) which resulted in these species being unidentified by the Illumina and MinION methods.

Table 3.2: Diatom species relative abundances as determined by light microscopy, Illumina short *rbcL* metabarcoding and MinION long rbcL metabarcoding for the rivers Ehen, Wear and Derwent. Only species present in >1% abundance in any sample are shown and a gradient of colour from red (0%) to yellow (6%) to green (>10%) has been applied to aid visualisation. Starred species (*) do not have a reference *rbcL* DNA barcode in the sequence database for identification.

	River Ehen			River Wear			River Derwent		
Species	**Light Microscopy**	**Illumina**	**MinION**	**Light Microscopy**	**Illumina**	**MinION**	**Light Microscopy**	**Illumina**	**MinION**
Achnanthidium caledonicum*	17.82%	-	-	0.00%	-	-	0.00%	-	-
Achnanthidium microcephalum*	17.82%	-	-	0.00%	-	-	0.00%	-	-
Achnanthidium minutissimum	31.68%	35.74%	29.73%	44.00%	24.71%	16.70%	10.77%	2.37%	2.93%
Achnanthidium pyrenaicum	0.00%	0.00%	0.00%	2.98%	0.00%	3.50%	0.00%	0.00%	1.26%
Adlafia suchlandtii*	0.00%	-	-	0.00%	-	-	2.25%	-	-
Amphora pediculus	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.20%	0.58%	0.83%
Aulacoseira sp.*	0.00%	-	-	0.00%	-	-	1.54%	-	-
Cocconeis euglypta	0.00%	0.00%	0.00%	1.44%	3.57%	12.41%	0.82%	0.81%	7.18%
Cyclotella comensis*	1.52%	-	-	0.00%	-	-	0.00%	-	-
Cymbella cf.	0.00%	0.00%	0.00%	0.00%	0.63%	6.43%	0.00%	4.37%	8.76%
Cymbella sp.	0.00%	0.00%	0.00%	0.00%	0.00%	3.05%	0.00%	0.00%	2.80%
Diatoma moniliformis	0.00%	0.00%	0.00%	1.24%	0.39%	0.37%	2.39%	0.58%	0.83%
Encyonema lange-bertalotii*	1.92%	-	-	0.00%	-	-	0.00%	-	-
Encyonema minutum	0.00%	0.54%	1.03%	2.61%	0.88%	3.20%	0.46%	0.31%	1.19%
Encyonema silesiacum	0.00%	0.00%	0.00%	0.32%	0.24%	1.59%	0.00%	0.00%	0.00%
Eunotia minor	0.00%	0.73%	1.53%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Fragilaria capucina	1.99%	0.51%	0.48%	0.00%	0.00%	0.00%	1.14%	0.00%	0.00%
Fragilaria gracilis	0.00%	1.20%	2.34%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Fragilaria pararumpens	0.00%	3.28%	0.69%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Fragilaria perminuta	0.08%	0.12%	1.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Fragilaria tenera	0.00%	2.37%	2.67%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Fragilaria vaucheriae	1.58%	0.45%	0.19%	0.00%	0.00%	0.00%	1.48%	0.10%	0.13%
Gomphonema acuminatum	0.07%	1.45%	0.34%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Gomphonema calcifugum*	0.00%	-	-	17.79%	-	-	15.26%	-	-
Gomphonema cf.	0.00%	3.20%	0.47%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Gomphonema clavatum	0.16%	0.43%	2.69%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Gomphonema exilissimum	5.56%	0.00%	11.46%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Gomphonema gracile*	5.41%	-	-	0.00%	-	-	0.00%	-	-
Gomphonema micropus	0.00%	0.00%	0.00%	0.00%	0.03%	3.75%	0.00%	0.00%	0.00%
Gomphonema minutum	0.00%	0.00%	0.00%	0.00%	0.26%	2.12%	0.07%	0.51%	2.20%
Gomphonema olivaceum*	0.00%	-	-	2.72%	-	-	16.46%	-	-
Gomphonema parvulum	2.55%	39.02%	25.67%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Gomphonema pseudobohemicum	0.00%	1.01%	0.18%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Gomphonema pumilum	4.88%	0.96%	3.40%	0.00%	0.00%	0.00%	0.00%	1.57%	1.05%
Gomphonema rosenstockianum	0.00%	0.00%	0.00%	0.00%	0.02%	3.64%	0.00%	0.00%	2.80%
Gomphonema sp.	0.00%	0.00%	0.00%	0.00%	34.48%	11.38%	0.16%	36.60%	10.52%
Hannaea arcus	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.36%	0.17%	0.29%
Mayamaea permitis	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	14.01%	0.00%	6.59%
Mayamaea atomus	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.47%	0.00%
Meridion circulare	0.00%	0.00%	0.00%	1.05%	0.02%	0.16%	0.00%	0.00%	0.00%
Navicula gregaria	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.31%	1.30%	2.53%
Navicula lanceolata	0.00%	0.00%	0.00%	0.92%	2.35%	1.58%	3.62%	25.26%	13.54%
Nitzschia dissipata	0.00%	0.00%	0.00%	1.49%	0.03%	0.10%	2.50%	0.05%	0.37%
Nitzschia inconspicua	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.65%	0.02%	0.14%
Nitzschia sociabilis	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.28%	0.00%	0.00%
Parlibellus protracta	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.05%	0.14%
Planothidium lanceolatum	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.31%	0.37%	1.02%
Reimeria sinuata	0.00%	0.00%	0.00%	15.43%	25.88%	19.15%	1.44%	1.81%	3.71%
Rhoicosphenia abbreviata	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.32%	7.40%	9.78%
Surirella brebissonii	0.00%	0.00%	0.00%	1.06%	1.33%	0.91%	2.35%	3.04%	2.68%
Synedra tenera*	1.75%	-	-	0.00%	-	-	0.00%	-	-
Tabellaria flocculosa	0.42%	0.87%	1.93%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Sub-total	**95.21%**	**91.88%**	**86.13%**	**93.05%**	**94.82%**	**90.04%**	**96.15%**	**93.74%**	**83.27%**

The river Ehen presented the most obvious example of missing DNA barcodes, where *Achnanthidium caledonicum* and *Achnanthidium microcephalum* were detected at 17.85% abundance each in the LM results but absent in the Illumina and MinION results. It was assumed that if a reference barcode from one species was missing from the database, that the sequences would be identified as closely related species, in this case another *Achnanthidium* species. However, the river Ehen also contained 31.68% of *Achnanthidium minutissimum* by LM, yet 35.75% *A. minutissimum* by Illumina metabarcoding and 29.73% by MinION metabarcoding. This suggested that sequences in the metabarcoding datasets originating from *A. caledonicum* and *A. microcephalum* were not being assigned to another *Achnanthidium* species as expected. The only other taxon in the river Ehen sequencing dataset to show significantly higher abundance in the metabarcoding than LM was *Gomphonema parvulum*. In contrast, *Gomphonema exilissimum* was not detected by Illumina metabarcoding in the river Ehen samples when it was detected in the LM (5.55% and MinION metabarcoding (11.46%).

Given the discrepancies in taxon assignment in the *Gomphonema* and *Achnanthidium* a neighbour joining tree was constructed of *Gomphonema* and *Achnanthidium* sequences from the *rbcL* reference DNA barcode database (Figure 3.3). This demonstrated that the *Gomphonema* and *Achnanthidium rbcL* sequences were very similar, that *Gomphonema* species are difficult to discriminate with *rbcL* and that potential cryptic species and sub-species may be present in both *Gomphonema* and *Achnanthidium*.

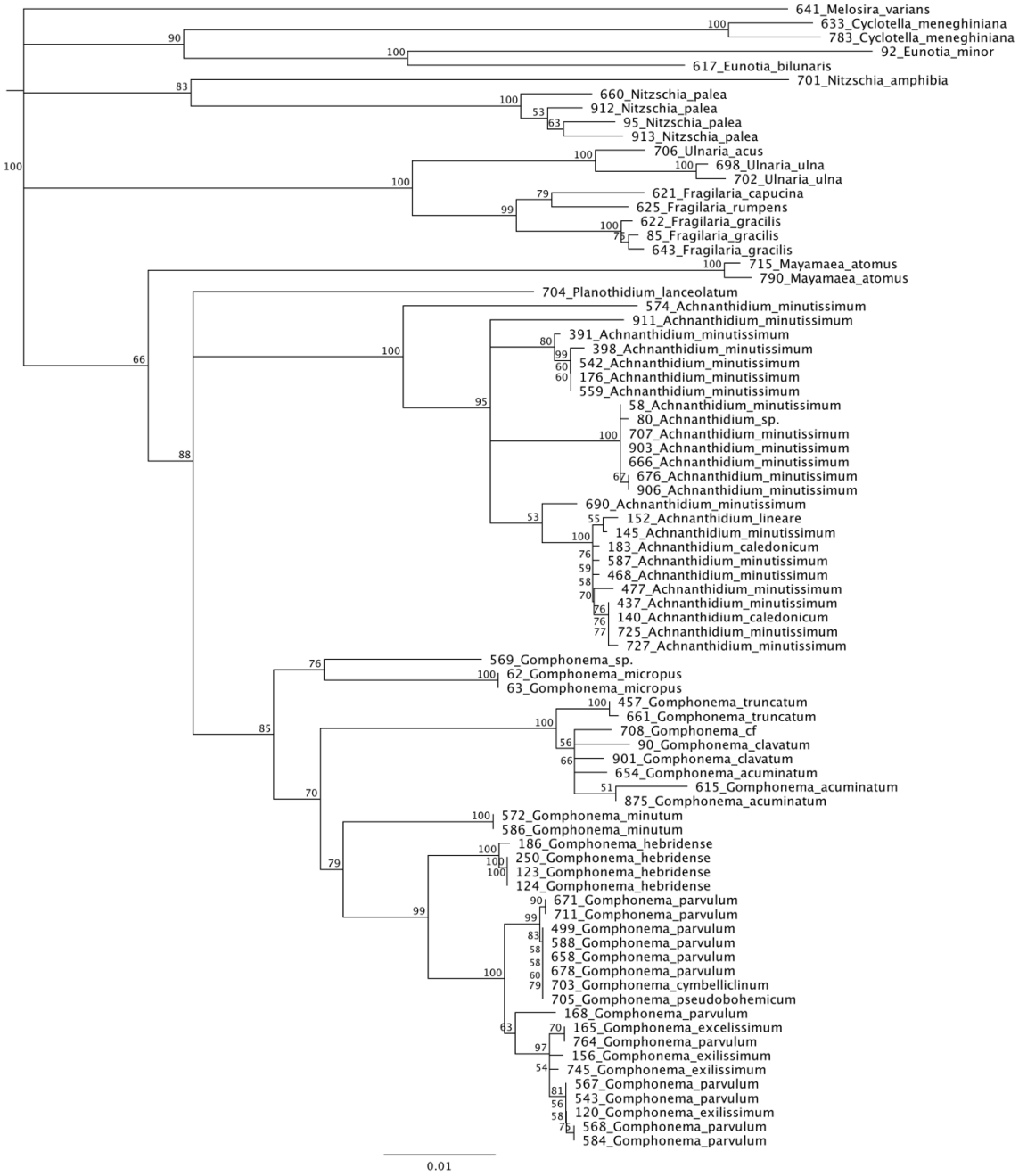


Figure 3.3: Neighbour-joining tree of *Gomphonema* and *Achnanthidium* species, along with a number of outgroup diatom species, demonstrating the difficulties in discriminating some *Gomphonema* species with full-length (Sanger sequenced) *rbcL* sequences and the potential cryptic species present within the genus *Achnanthidium*. The numbers prefixing each sequence are the isolate numbers within the *rbcL* reference sequence database. Branch labels are bootstrap support from 1000 replicates.

There was a difference in the number of species detected in all samples, with far fewer species detected by light microscopy compared to Illumina and MinION *rbcL* metabarcoding (Figure 3.4). Fewer species were detected by MinION long barcode metabarcoding than Illumina short barcode metabarcoding. However, the next-generation sequencing approaches demonstrated a greater diversity of species, even though a number of key species detected by light microscopy did

not have reference *rbcL* DNA barcodes and therefore could not be detected by sequencing. The total number of species detected, and shared, by each method are shown in Table 3.3 The number of species determined when the samples were grouped by river (Figure 3.5) is less informative but shows that overall the mean number of species detected in each river was similar and there were no large differences in the number of species detected in each river.

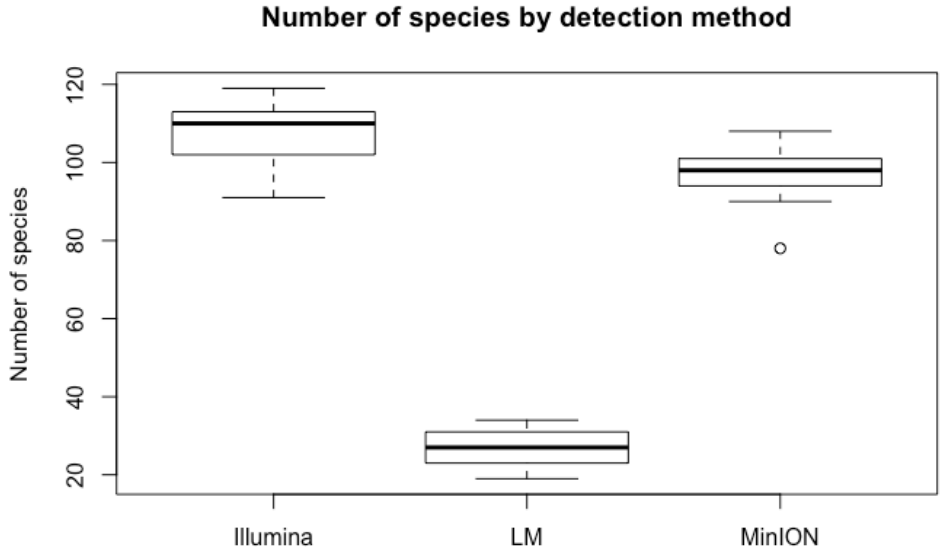


Figure 3.4: Boxplot showing the number of species detected across all samples by method (Illumina short *rbcL* metabarcoding, light microscopy and MinION long *rbcL* metabarcoding)

Table 3.3: Table showing the total number of species detected and shared by each method (Illumina, MinION and Light Microscopy). Very few species were shared between the two sequencing methods and light microscopy.

Total number of species	208
Number of species shared between Illumina, MinION and Light Microscopy	42
Number of species shared between Illumina and MinION	76
Number of species shared between Illumina and Light Microscopy	1
Number of species shared between MinION and Light Microscopy	1
Number of species only detected by Illumina	24
Number of species only detected by MinION	25
Number of species only detected by Light Microscopy	39

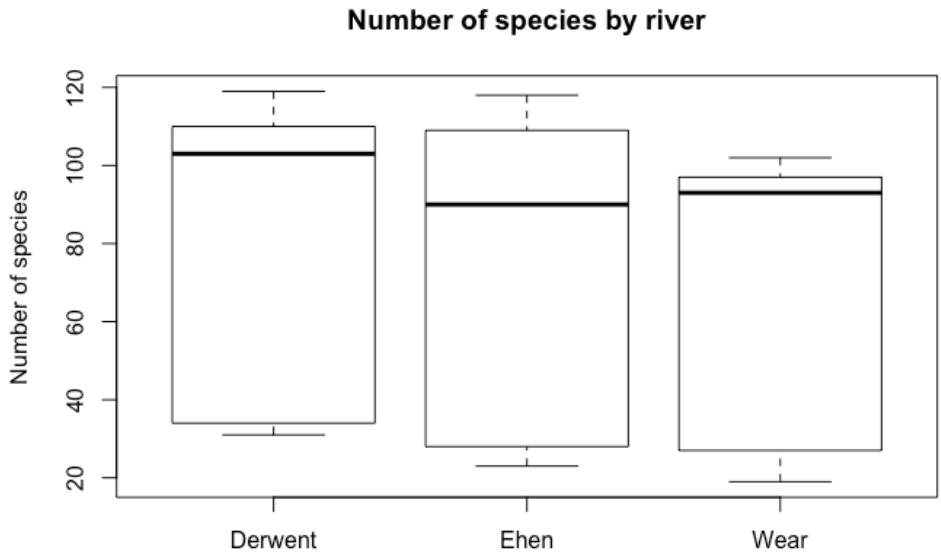


Figure 3.5: Boxplot showing the number of species detected across all samples by river (Derwent, Ehen and Wear). The number of species for each river is a combination of each of the three detection methods: Illumina, MinION and Light Microscopy.

Bray-Curtis dissimilarity metrics for all samples were calculated and average-linking hierarchical clustering applied to measure how similar the rivers were in terms of their species composition between samples, methods and rivers (Figure 3.6). This showed that the three rivers are in separate clusters, with distinct sub-clusters for each method (Illumina, MinION and LM). The rivers Wear and Derwent were found to be in a single larger cluster, separate from the river Ehen, which also corresponds with the geographically different locations of the rivers with Ehen being in Cumbria and the Wear and Derwent being in County Durham.

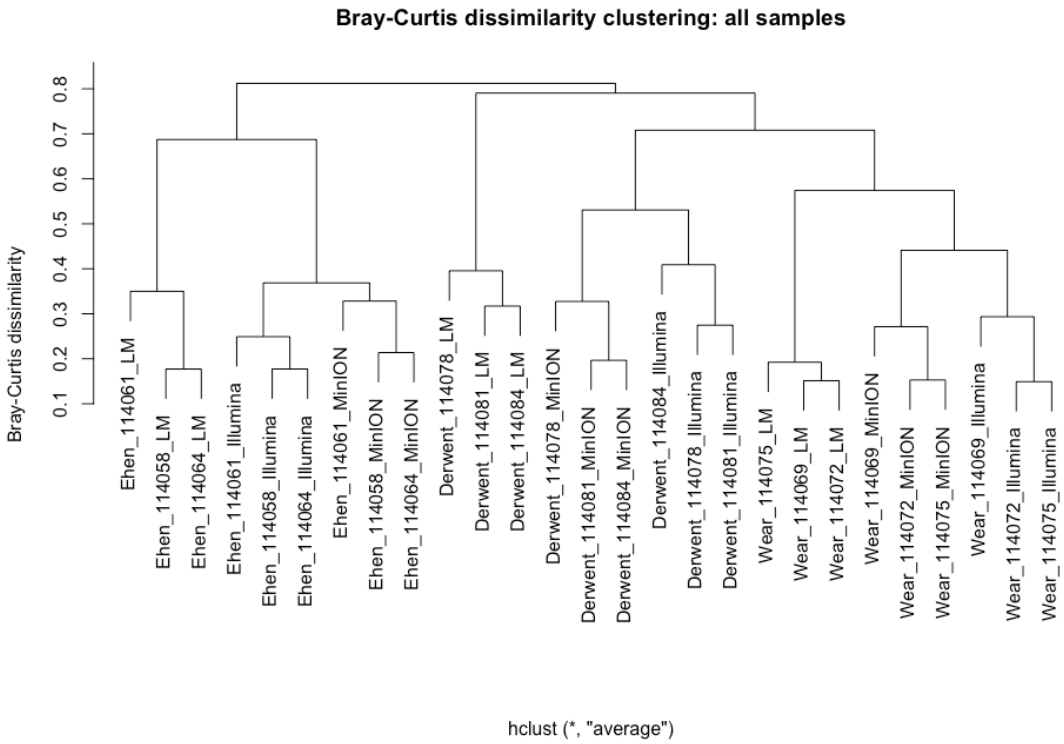


Figure 3.6: Hierarchical clustering of the Bray-Curtis dissimilarity metric between all 9 samples, three methods and three rivers.

Non-metric multidimensional scaling (NMDS) ordination of the Bray-Curtis dissimilarities was used to visualise the structure of the communities when split by river (Figure 3.7) and by method (Figure 3.8). When visualised by river the abundance results were clearly separated by the river of the original sample, despite different groupings observed when the abundance results were visualised by method (Figure 3.8). When split by method, two groups were apparent, one for light microscopy and a second for the two sequencing methods.

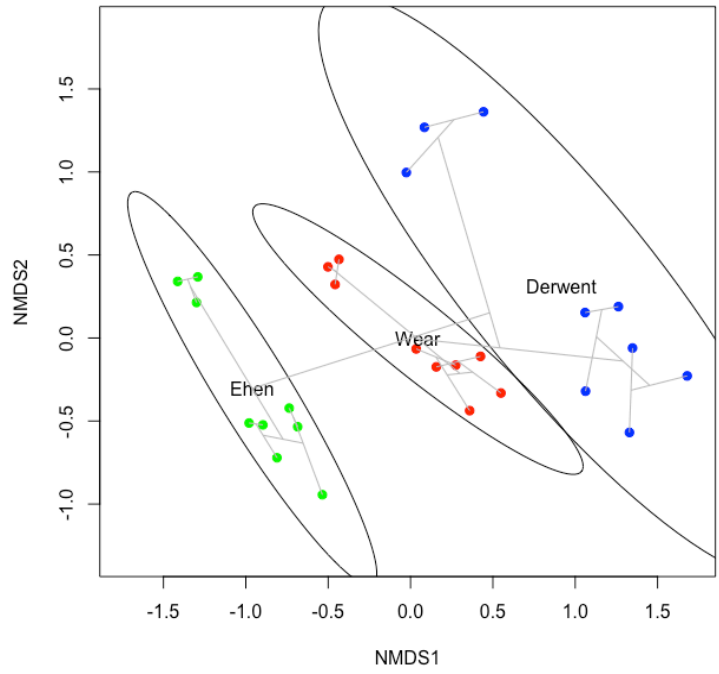


Figure 3.7: NMDS showing the relationships between the samples separated by the rivers Ehen (green), Wear (red) and Derwent (blue). The additional light grey lines show the hierarchical clustering results.

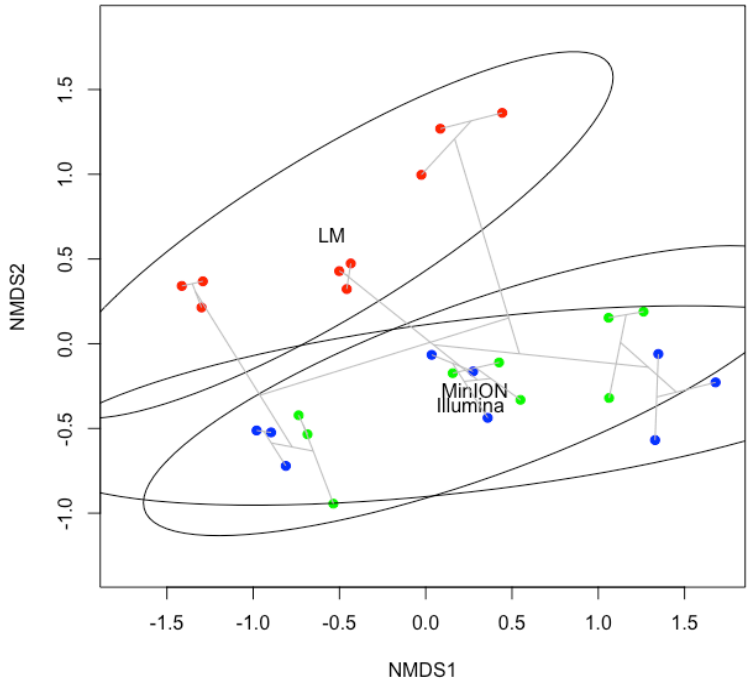


Figure 3.8: NMDS showing the relationships between samples separated by method: light microscopy (red), MinION (green) and Illumina (blue). The additional light grey lines show the hierarchical clustering results.

To test the variance in diversity between methods and rivers and to quantify the degree by which river and method explains the dissimilarity observed, the permutational multivariate analysis of variance (PERMANOVA) method adonis (Anderson, 2001) was used within the vegan package (Oksanen et al., 2007). Only 27% of the variance could be explained when the dataset is split by river (R^2 = 0.27, $p<0.001$) but 44% could be explained when the dataset is split by method (R^2 = 0.44, $p<0.001$) demonstrating that the diversity in the nine samples is affected more by abundance calculation method than geographical location.

MinION sequence lengths

The MinION read length distributions are shown in Figure 3.9 along with the number of MinION reads assigned to each genus. In most genera with more than 10 MinION reads assigned there was little length variation in the majority of sequences. The majority of shorter sequences were not found to have sufficient homology to the diatom *rbcL* sequences in the reference database and were assigned as “Unknown”. The genera where a low mean amplicon length is observed (e.g. *Biremis*, *Cylindrotheca*, *Halamphora*) had fewer sequences assigned to them and thus it could not be determined whether the shorter sequences had any true significance.

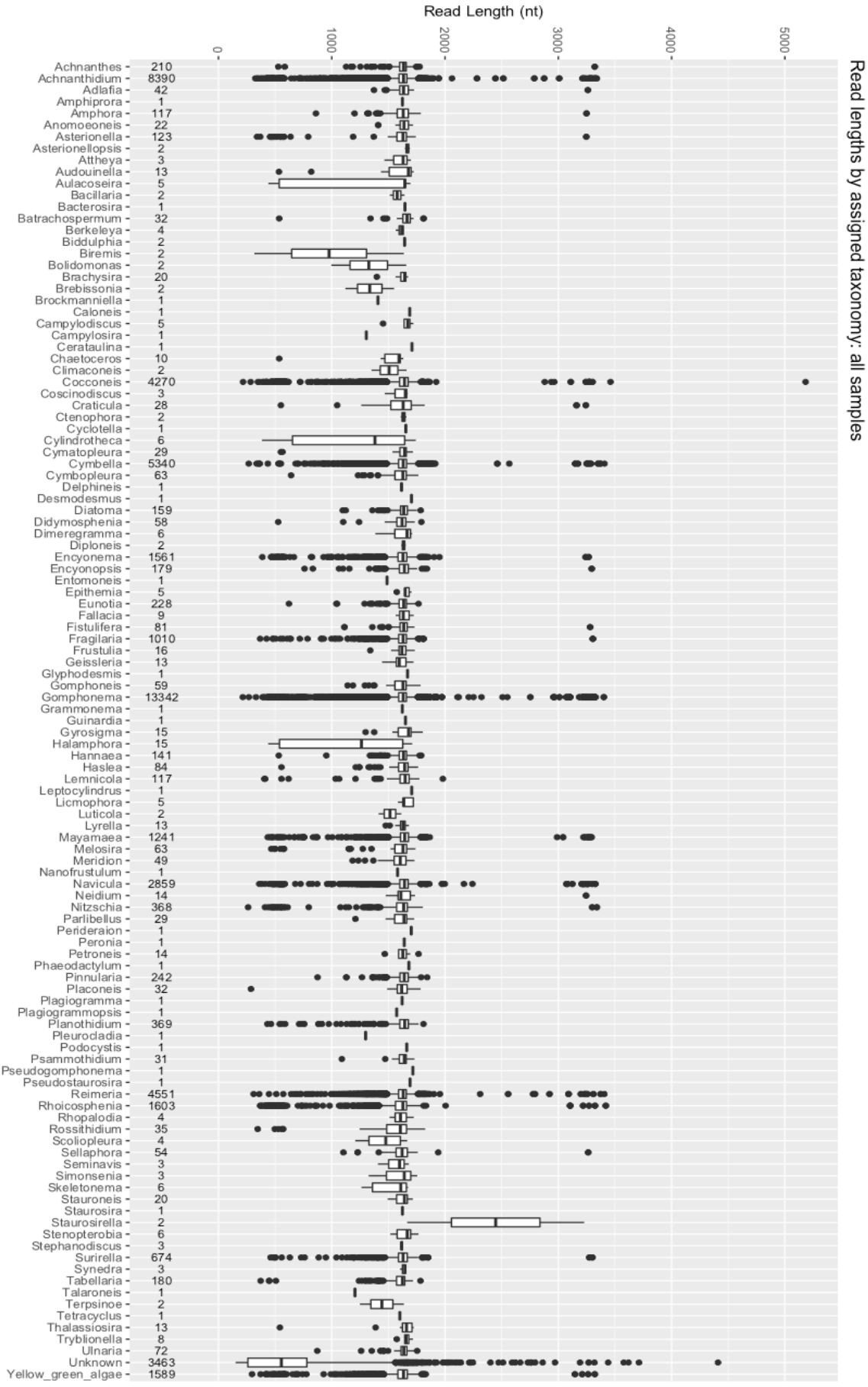


Figure 3.9: MinION read length by the genus the read was assigned to during analysis. The numbers to the right of the genus name are the number of MinION sequences assigned to that genus from all samples.

There were a number of genera where much longer sequences were observed than the expected amplicon size, predominantly *Achnanthidium*, *Cocconeis*, *Cymbella*, *Gomphonema*, *Mayamaea*, *Navicula*, *Reimeria* and *Rhoicosphenia*. These genera had large numbers of MinION sequence reads assigned to them. The longer reads were, in most cases, double the size of the expected amplicon. Investigation of the *Cocconeis* sequences more than 2kb in length (n=13, lengths 2.88kb-5.18kb) by blastn showed that the longer sequences did consist of two amplicons side-by-side with both amplicons being assigned to the same diatom reference sequence. However, the more 3' amplicon sequence in all cases was 8-11% more dissimilar to the reference sequence used for taxon assignment than the more 5' amplicon sequence. A subsequent blastn-short search for the MinION nanopore hairpin sequences was positive for all the longer sequences in the *Cocconeis* group suggesting that the longer sequences are not due to a duplication of *rbcL* or chimeric amplicon sequences but an artefact of the nanopore sequencing and/or 2D sequence assembly/production. The longer sequences were not associated with any particular sample or river.

A large number (n=3463) of MinION reads from across all samples could not be identified when searched with blastn against the diatom reference database of *rbcL* sequences. The primers used to amplify the ~1600bp *rbcL* region for MinION sequencing contain degenerate bases and were designed to amplify diatom species (Jones et al., 2005) only, but this does not preclude the possibility of the primers amplifying non-diatom sequences where no reference existed in our database. The taxonomic identifications elucidated from further searches with blastx against the non-redundant NCBI database are shown in Figure 3.10. Sequences assigned as bacterial in origin were generally short - around 650bp - and originated from many different genes with differing functions. The sequences assigned as diatoms (taxa below the class *Bacillariophyta*) contain species which were present in our validated diatom reference database of *rbcL* sequences and the majority of these sequences were identified as originating from *rbcL* suggesting successful amplification by PCR. However, these sequences only shared between 30% and 49% amino acid homology with their assigned taxonomy. This suggests that these sequences represent either undescribed or as-yet-unsequenced species of diatoms present in our samples.

Three taxa were assigned more of the unknown sequences than others (Figure 3.10): *Fistulifera solaris* (coloured red), *Phaeodactylum tricornutum* (coloured blue) and *Thalassiosira oceanica* (coloured green). Those taxa represented species where a genome sequence was available. The sequences assigned to these taxa contained a diverse range of genes rather than *rbcL* only and were found to originate from all nine of the MinION sequenced samples. The assignment of sequences to *Naviculaceae* (coloured purple) demonstrated that there are diatom sequences in our data which had not yet been fully sequenced or described. None of the sequences investigated by blastx showed potential for being the "missing" species present in the light microscopy counts but not represented in our reference sequence database.

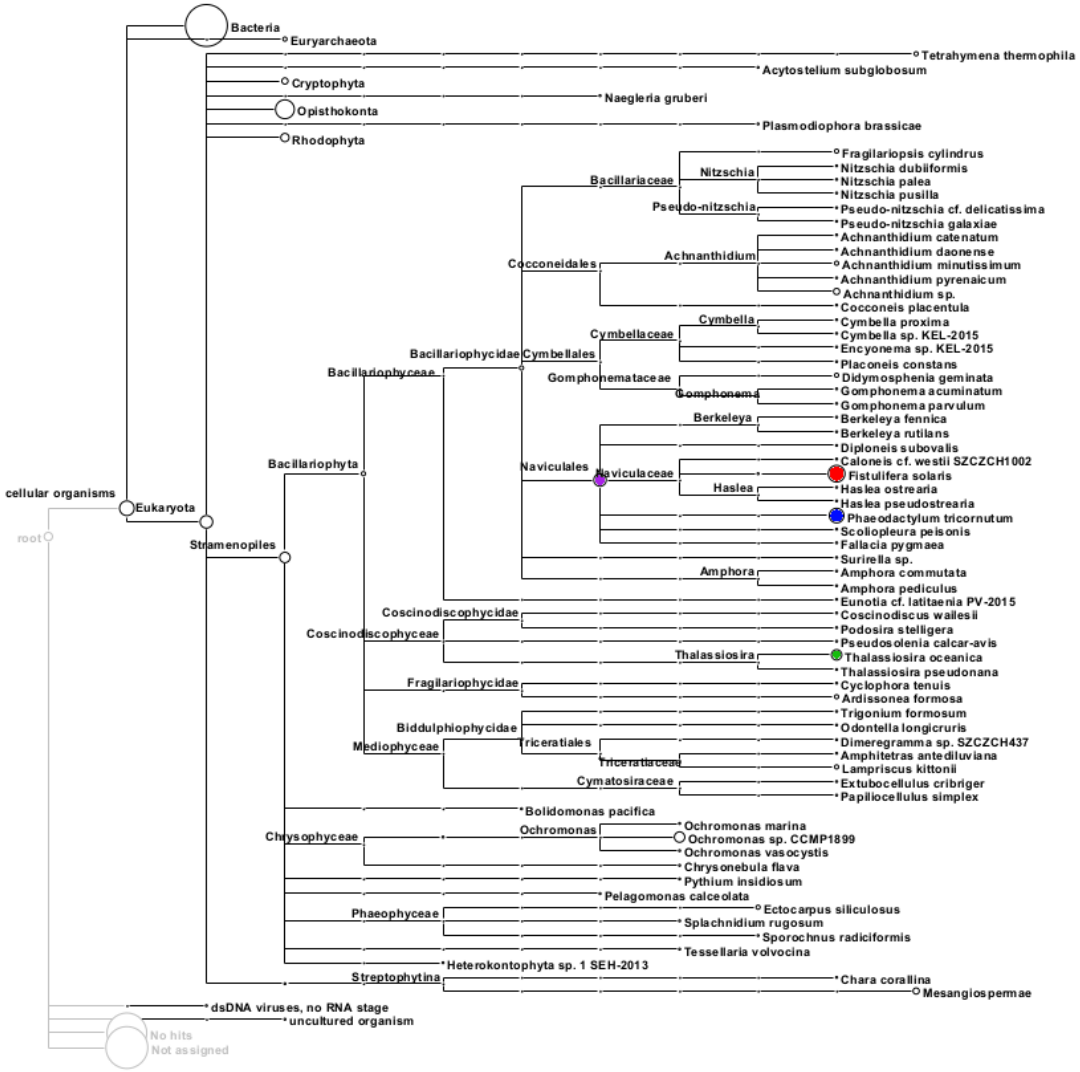


Figure 3.10: Taxonomic identifications of MinION sequences originally designed as "unknown" when searched against the validated reference database. Coloured dots represent species where a genome sequence is available.

Taxon assignment in *Achnanthidium* and *Gomphonema* with MinION

Further investigation into how the observed sequence diversity in *Achnanthidium* and *Gomphonema* might affect taxonomic assignment, particularly in the higher-error-rate MinION sequencing, is shown in Figure 3.11 and Figure 3.12. The top-hit blastn percent identity to the reference diatom sequence used to assign taxonomy was plotted for all MinION sequences identified as *Achnanthidium* species (Figure 3.11) and *Gomphonema* species (Figure 3.12).

The majority of MinION reads identified as belonging to species within *Achnanthidium* were assigned to *A. minutissimum* (8124 reads across the three rivers) and the mean percent identity used for species identification was high (>85%). However, the spread of sequence similarities was very broad, from 72% to 100%, which suggests again that there were cryptic species and/or undescribed species present in the river samples.

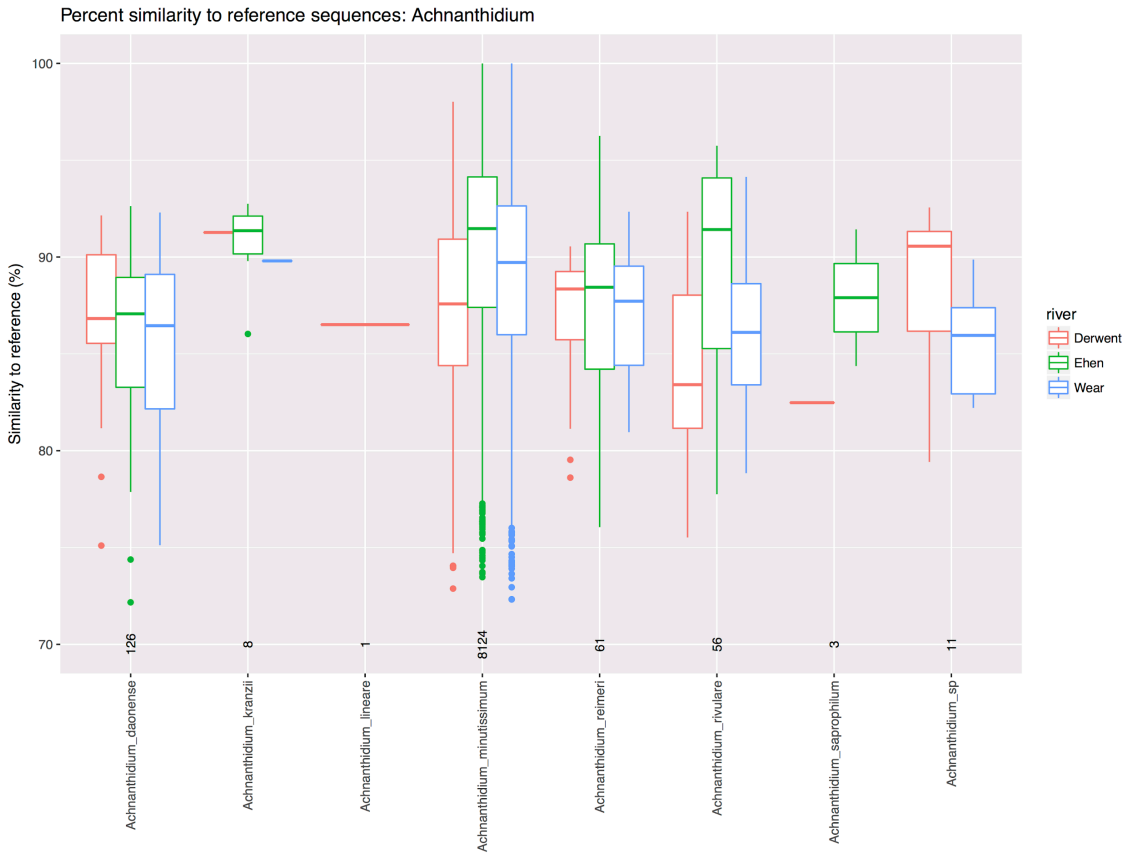


Figure 3.11: The best percentage sequence similarity used to assign each MinION read to species in the genus Achnanthidium. The numbers next to the species name are the number of MinION reads assigned to that species.

The predominant species identified in *Gomphonema* with MinION sequencing were *G. bourbonense, G. clavatum, G. exilissimum, G. micropus, G. parvulum, G. rosenstockianum* and *Gomphonema sp.* The latter assignments are all assigned to one reference *rbcL* DNA barcode only identified as "*Gomphonema* sp", rather than this describing a catch-all for *Gomphonema* sequences not assigned to other species. However, the breadth of sequence identities in this group points towards many yet-to-be described *Gomphonema* species.

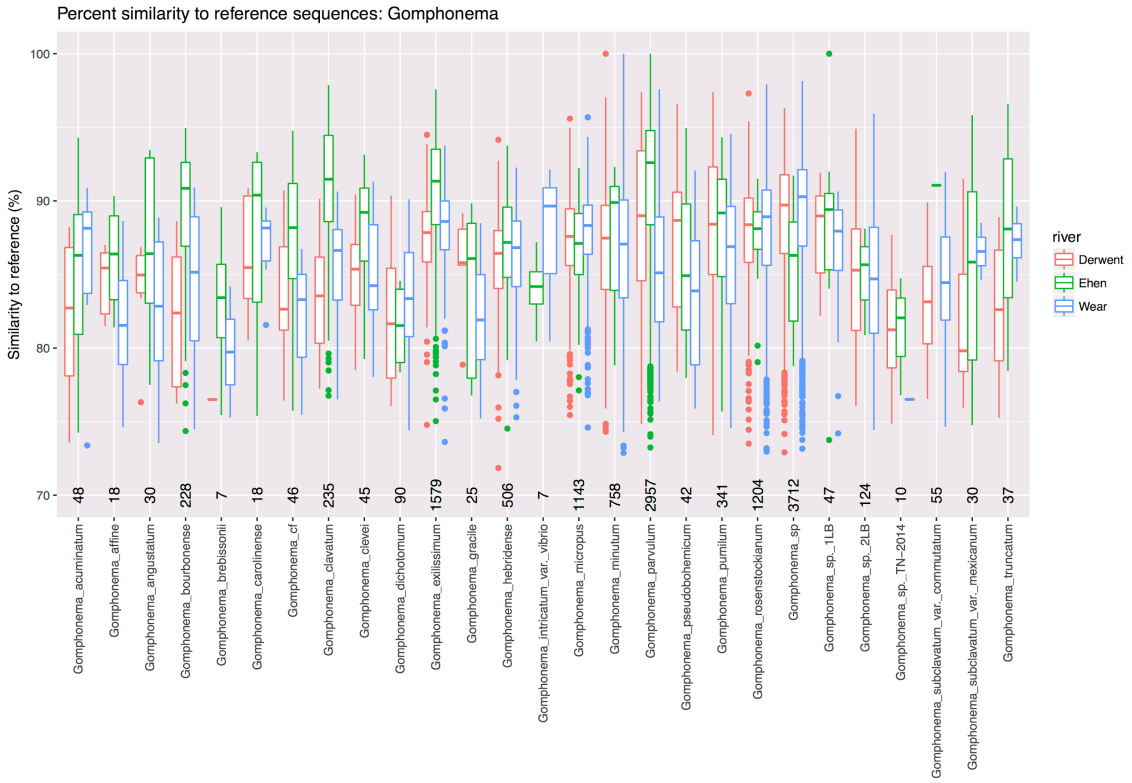


Figure 3.12: The best percentage sequence similarity used to assign each MinION read to species in the genus *Gomphonema*. The numbers next to the species name are the number of MinION reads assigned to that species.

Very few MinION sequences were assigned to taxa with >95% sequence identity. In most cases, for both *Achnanthidium* and *Gomphonema*, the best match taxon was determined with between 80% and 90% sequence identity.

Discussion

Diatoms are sensitive to changes in the availability of nutrients such as nitrate and phosphate and thus the species composition and relative abundance of diatom assemblages are used as a measure of ecological status for monitoring water bodies.

In this study we have compared three different methods for the determination of diatom species abundance in three different rivers. The largest effect on the comparison between light microscopy, Illumina metabarcoding and MinION metabarcoding was the absence of reference *rbcL* DNA barcodes for species identified by light microscopy. Previous work within the larger project for the Environment Agency (Kelly et al, 2018) had shown that while it was technically possible to infer potential barcode sequences from NGS data when compared with the light microscopy of the same sample, it was determined not to be an approach that could be used for the production of robust reference barcodes. Reference DNA barcodes produced in this manner would be subject to sequencing errors from either Illumina or nanopore technology and there would be significant risk that while we may use the barcodes with caution and with caveats, others may see them in repositories and assume they are correct, leading to spurious identifications. Further work is being undertaken by the Environment Agency in 2018-2019 to find and culture species of interest for the production of accurate DNA barcodes for these species.

We had made the assumption that when DNA barcodes of a particular species were absent from the database that sequences of the missing species would be assigned to a closely related species. Our results showed that sequences which may have originated from *Achnanthidium* species with no DNA barcode may have been assigned as *Gomphonema* species instead. This was unexpected but the taxonomy of diatoms is based upon frustule structure (Kaczmarska et al., 2007; Mann et al., 1996) and as names and classifications can change it could be that metabarcoding enlightens taxonomists to relationships previously unseen. Our metabarcoding methods used a blastn e-value threshold of 0.01 to filter out spurious identifications and the *rbcL* reference DNA barcode database had been thoroughly cleansed of misidentified sequences in the previous project (Kelly *et al.,* 2018) so we are confident that our taxon assignments were not different due to flaws in the reference database. An additional exploratory analysis of sequences left unidentified after comparison with the *rbcL* reference barcode database showed a number of sequences which had homology to diatom *rbcL* sequences. These are likely to represent currently undescribed diatom species or potentially some of the “missing” DNA barcodes from the reference database;

however, it is difficult to determine as the abundances were low in the unidentified sequences and inferring a species name from sequence similarity in a field sample would not be recommended. An interesting result in the exploratory analysis of unidentified sequences was the background of genomic DNA present in the MinION sequencing datasets. As MEGAN (Huson et al., 2016) uses a lowest common ancestor (LCA) algorithm for taxon assignment it is able to assign species to higher taxonomy which can be useful for identifying new species or diatom genomic DNA remaining in the PCR products. The number of non-*rbcL* sequences, in particular bacterial sequences, demonstrate that there appears to be a background contamination originating from DNA extraction kits, the PCR and library preparation reagents or aerosol contamination from the laboratory environment. This type of contaminating sequence have more recently been referred to as the "kitome" (Salter et al., 2014).

A serious consideration is that the light microscopy results were not as accurate as one might expect. The preparation of diatom slides for microscopy is chemically harsh, some species are not detected because they are small enough to be beyond the limit of detection of light microscopy, and the variability between samples when analysed with microscopy is high (Kelly, 1999; Prygiel et al., 2002). Species counts from light microscopy often come from a sample of a few hundred diatoms whereas next-generation sequencing enables the determination of counts from all diatoms in the sample. Our results show good correlation and statistical support for the similarity of the Illumina and MinION metabarcoding methods. The two metabarcoding methods use different primer sets, different amplicon lengths and the sequencing methods have different error profiles, so the similarity in relative abundance for samples when sequenced with the two methods gives support to the concept that the light microscopy is the least accurate of the three methods (Zimmermann et al., 2015). It would be typical to expect that microscopy gives the "correct" abundances and to assess a new method in its ability to match this. However, in this study we have compared the three methods without this assumption. With two metabarcoding methods showing comparable abundances when analysing the same sample, it is unlikely that primer bias is the cause of the differences between LM and metabarcoding.

The error rate for MinION at the time of sequencing and analysis (2016) was between 7.5% and 14.5% (Jain et al., 2017) but despite this disadvantage to taxon assignment, our MinION and Illumina results are comparable. Despite the error rate taxon assignment was on par with the Illumina method but interestingly, the breadth of percent identities used for taxon assignment was wider in the MinION results. Some taxon assignments in the genera *Achnanthidium* and *Gomphonema* were made at just 70-80% nucleotide identity. This would suggest that even with the MinION's relatively high error rate there is considerably more variability within the species of these genera. It is not unlikely that *Achnanthidium minutissimum* and *Gomphonema parvulum*, for example, represent species complexes or cryptic species.

The Illumina method has weaknesses in the length of the amplicon and its reduced ability to discriminate species due to the short length providing less variability but a good error rate of <1%. Our data showed that *Gomphonema exilissimum* could not be identified by Illumina metabarcoding but was present in the microscopy and MinION results. MinION currently has a high error rate but the distinct advantage of longer amplicons to exploit more species-specific variability in *rbcL*, which may account for its ability to detect species such as *G. exilssimum* where Illumina metabarcoding could not. It is worth noting that the Illumina sequencing methods are unlikely to improve in the future with regards to error profile or read length. However, the MinION platform is developing rapidly, with falling error rates and increasing read lengths along with the additional potential for field-based sequencing of diatoms at the riverside.

The calculation of the TDI relies upon the accurate determination of the relative abundance of species within the sample and so the method used to determine this is important. The three rivers were selected for further MinION testing because they had three different water quality statuses representing different abundances of species. The TDI4 was based upon light microscopy and a further iteration of the method, TDI5, was produced based upon the Illumina method developed in Chapter 1. The TDI5 (Kelly et al., 2018) adjusted the species-specific weightings based on abundances observed over the 500 samples analysed with LM and the Illumina metabarcoding method. The three rivers tested during this analysis only comprised single examples of "excellent", "good" and

“moderate” water quality. With the expected future improvement of nanopore sequencing accuracy, a much larger scale comparison would be required in order to determine if further adjustment of the TDI weightings would be needed to use nanopore sequencing for statutory water quality testing that would feed into the Water Framework Directive.

The next step in assessment of diatom communities may well be the sequencing of whole chloroplast genomes to fully embrace the current long-read sequencing technologies to enable better species discrimination. However, the sequencing of whole chloroplast genomes would have the same inherent issues as those proposed for sequencing whole mitochondrial genomes to replace COI DNA metabarcoding for invertebrate and fish studies (Deiner et al., 2017; Sato et al., 2018) insofar as the abundance would be skewed by the size of the organism and therefore the number of organelles isolated. To this end, accurate quantification of unicellular diatom species present in an environmental sample would require a single-copy nuclear region. Historically, mitochondrial and chloroplast genes were used for phylogenetic and population studies due to their abundance within cells and therefore greater yield of DNA for PCR amplification success. Their lack of introns was also an advantage allowing easier amplification with conserved primers and sequencing of the whole protein-coding region. These advantages have led to an abundance of sequence data for mitochondrial and chloroplast genes from morphologically verified specimens and compounds their present and future use in phylogenetic and species identification studies (Hebert et al., 2003a; Hebert et al., 2003b; Nilsson et al., 2008; Tedersoo et al., 2012). Yet the quantification problem remains with their continued use in metabarcoding studies. With the advent of cheaper and more accessible genome sequencing of morphologically verified and vouchered specimens, additional single-copy nuclear regions may be discovered which can discriminate species adequately and allow the accurate quantification of species when used as a metabarcoding locus. However, this would still be an interim measure between current metabarcoding methods and full power of metagenomic sequencing of environmental samples to determine species composition.

Our current monitoring methods use changes in diatom species abundance as a measure of response to changes in nutrients and inorganic contaminants in water

bodies. However, the future may see new eRNA monitoring methods developed where the up- or down-regulation of diatom gene expression in response to nutrients and contamination can be measured and assessed in real-time with RNA-seq of water samples by nanopore sequencing technologies in the field.

Chapter 4. Distribution of fungal plant pathogens over one month in eastern England

Introduction

Quarantine pests and pathogens are defined internationally as “pests of potential economic or environmental importance to an area, which are not present there or which, if present, are not widespread, and are being officially controlled” (IPPC, 2011). Additionally, there are regulated non-quarantine pests and pathogens, which are widely established and cannot be described as quarantine pests, but which require control and phytosanitary methods to a certain degree. The European Plant Protection Organisation (EPPO) is the regional organisation with responsibility for the promotion and harmonisation of approaches for the detection and control of plant pathogens. EPPO identifies the pathogens and pests which could pose a risk to its member countries and has arranged them into four lists for guidance, which are regularly updated: A1 (organisms absent from the EPPO region), A2 (organisms which are present in the EPPO region), Alert (non-quarantine organisms which present a phytosanitary risk) and Action (organisms on the A1 and A2 lists which are of particularly urgent phytosanitary risk). Each member state can also produce its own lists and the United Kingdom has the Plant Health Risk Register, which records and rates the risk to crops, trees, gardens and ecosystems from plant pests and pathogens (UK Plant Health Risk Register).

The causative agent of ash dieback, *Hymenoscyphus fraxinus*, is a fungal pathogen that causes shoot dieback, necrotic lesions and leaf wilting in ash trees. In young trees it is often lethal but in older trees ash dieback can result in a chronic infection, weakening the trees and predisposing them to other diseases (Cleary et al., 2016). It was first observed in Poland in 1992 (Kowalski, 2006) and has since been reported in more than 22 European countries (Timmermann et al., 2011), and is now considered widespread. Recent evidence has suggested the pathogen was recently introduced to Europe from East Asia (Gross et al., 2014) and in 2012 it was first reported in the United Kingdom (EPPO, 2012). The introduction of ash dieback to the United Kingdom gained widespread publicity

given the number of ash trees in the UK and was initially linked to the import of ash seedlings from continental Europe. A systematic and widespread survey of ash trees across the United Kingdom in late 2012 established that many of the affected sites were in eastern England and Scotland, leading to the conclusion that the disease had also been introduced by wind-borne spores from mainland Europe (Heuch, 2014). This conclusion led to the publication of a revised plant biosecurity strategy for Great Britain which aims to deliver an improved biosecurity system that is resilient and able “to respond effectively to new and emerging threats” (DEFRA, 2014).

In recent years, next-generation DNA sequencing technologies have matured and become a cost-effective way of both monitoring known species and undertaking surveillance for potential new threats (Bulman et al., 2018; Galan et al., 2016; Ji et al., 2013). Environmental DNA (eDNA) techniques - where the total DNA is extracted from a sample and either metabarcoding (targeted) or metagenomics (non-targeted) sequencing is carried out - have become popular and have been used widely (Douglas et al., 2012; Elbrecht and Leese, 2016; Pierre Taberlet et al., 2012). Standardised protocols for metabarcoding, coupled with cost-effective multiplexing of samples, deliver advantages in the simultaneous detection of both described and unknown organisms. With regards to fungal metabarcoding, the ribosomal internal transcribed spacer (ITS) region has become the most widely used locus (Schoch et al., 2012). The locus is widely utilised and reference sequences exist for many described and undescribed species within UNITE, the main repository for such sequences (Nilsson et al., 2013).

A number of studies have already shown that metabarcoding can be useful in assessing the presence of airborne fungal species. Air sampling of the indoor environment has been demonstrated and shown to provide high resolution fungal identification with ITS2 (Korpelainen and Pietilainen, 2015). In Canada, it has been shown that spore samplers can be used to collect fungal species for metabarcoding studies effectively, albeit with the conclusion that air samplers collect more Ascomycota and rain samplers collect more Basidiomycota (Chen et al., 2018). Air samples taken from urban environments in the UK and the

Netherlands have also shown that plant pathogens can be detected successfully with metabarcoding from rooftop air samples (Nicolaisen et al., 2017).

We present the results of a baseline study for the daily monitoring and surveillance of air-borne fungal species present in six locations in eastern England over one month in June 2015. The study had the dual aims of determining if we could use next-generation amplicon sequencing for routine monitoring and surveillance of fungal plant pathogens in the UK and whether the existing pollen network could be used for collection of representative samples.

Materials and methods

Spore sampling

Samples were obtained from two sources. Samples from Wansford, Alford, Haywold, and Stokesley came from the Crop Monitor network (CropMonitor, 2017) of spore samplers located in crop fields (Wansford: Latitude 53.55°N, Longitude -0.43°W; Alford: Latitude 53.26°N, Longitude 0.18°W; Haywold: Latitude 53.99°N, Longitude -0.60°W and Stokesley: Latitude 54.58°N, Longitude -1.16°W). The samplers were Burkard multi-vial Cyclone samplers (Burkard Manufacturing, UK). They continuously sample pollen and fungal spores, which were then deposited into a 1.5ml tube. The tubes were changed every day giving daily samples of spores during the period 29/5/2015 to 30/6/2015. Samples from York and Beverley came from the UK Pollen Monitoring Network. Both were Burkard volumetric spore traps (Burkard Manufacturing, UK) similar to the Crop Monitor traps but with the spores being collected on Melinex tape coated with petroleum jelly and paraffin wax and mounted on a rotating drum. Tape representing spore samples for each day was collected and slide mounted as described in Brittain et al (2013). Both traps were mounted on top of buildings in urban areas. The York trap was located on the roof of a building at the University of York (Latitude 53.95°N, Longitude 1.05°W) and the Beverley trap on the roof of an East Riding Council building in Beverley, East Riding, UK (Latitude 53.84°N, Longitude -0.42°W).

DNA extraction

Mounted slides were individually placed in 50ml falcon tubes (Fisher Scientific, UK) containing 2ml of CTAB buffer (10mM sodium phosphate buffer pH8, 50mM cetrimonium bromide, 1.5M sodium chloride) and incubated at 60°C on a rotating carousel for 30 minutes. The tubes were then centrifuged at 500g for 20 seconds in a Sigma 4K15 centrifuge (Sigma Laborzentrifugen GmbH, Germany) and the slide and coverslip removed.

Slide samples and Cyclone samples were then disrupted in 50ml falcon tubes by adding 1g of an equal weight mixture of 2.3mm and 0.5mm zirconia silica beads (Fisher Scientific, UK) and vortexed for 4 min at full speed on a Vortex Genie II (Fisher Scientific, UK) using a horizontal vortex adapter (Qiagen, UK). The tubes were centrifuged at 5000g in a Sigma 4K15 centrifuge and the clear lysate removed. The lysate was then extracted using a nucleospin Plant 2 kit (Macherey-Nagel, Germany) as per the manufacturer's instructions. Buffer-only samples were taken through the complete extraction process as extraction blanks.

DNA amplification and sequencing

Part of the ribosomal ITS1 region was amplified using primers Nex_ITS1_Ky02F (Toju et al., 2012) with the sequence:
5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTAGAGGAAGTAAAAGTCGTAA-3'
and Nex_ITS1R_Wobble with sequence:
5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCWGYGTTCTTCATCGATG-3'.
Thirty-microliter reactions consisted of 5µl HF buffer, 0.6U Phusion polymerase and 300uM dNTPs mix (all New England Biolabs, USA), 300nM forward and reverse primers (Eurofins, Germany) and 1µl extracted DNA. Extraction blanks, amplification blanks and a positive control consisting of a Gblock (IDT, UK) with artificial sequence separating Nex_ITS1_Ky02F and Nex_ITS1R_Wobble binding sites were also amplified. The resulting solution was amplified on a BioRad C1000 thermocycler (BioRad, USA) at 98°C for 2 min followed by 25 cycles of 98°C for 20 sec, 54°C for 30 se and 72°C for 90 sec. The reaction was completed with 10 mins at 72°C. Amplified DNA was assessed by agarose gel electrophoresis in a 1% gel. The reactions were then cleaned with Ampure XP beads (Beckman Coulter, UK), indexed and again cleaned with Ampure XP

beads as described in the Illumina 16S metagenomics library prep guide (Illumina, USA) with the exception that Phusion polymerase was substituted for Kappa polymerase. The final DNA library fragments were then quantified using Picogreen (Lifetech, UK), measured on a Fluoroskan Ascent fluorimeter (Thermo Scientific, UK) and mixed in equal quantities to create a 20nM pool. This pool was assessed using a D1000 tapestation tape (Agilent, UK) prior to running in the presence of 10% PhiX control (Illumina, UK) on an Illumina MiSeq with a V3 2x300 cycle flow cell (Illumina, UK).

Sequence quality control

A stringent quality control procedure was applied to the raw sequence data. Firstly, PCR amplification primers and any remaining Illumina adapter sequences were removed from both sequenced strands of DNA using Cutadapt v1.9.1 (Martin, 2011). Secondly, poor quality 3' ends of sequences from both strands were removed with sliding window trimming using Sickle v1.33 (Joshi and Fass, 2016) in paired-end (pe) mode. Thirdly, the trimmed read-pairs were merged to form single consensus sequences with PEAR v0.9.6 (Zhang et al., 2014). Finally, a further round of quality assessment was carried out to remove any sequence with an overall accuracy of less than 99.9%, length <100bp and sequences containing 'N' bases. This was achieved with Sickle v1.33 (Joshi and Fass, 2016) in single-read (se) mode. Remaining PhiX positive sequencing control sequences were identified by mapping the good quality reads to the PhiX reference sequence (NC_001422) with bowtie2 (Langmead et al., 2009) and were removed from the dataset with bedtools (Quinlan and Hall, 2010). Chimeric amplicon sequences were identified using the usearch61 method (Edgar, 2010) implemented within QIIME v1.9.1 (Caporaso et al., 2010) and removed from the dataset.

Sequence analysis

The number of sequence reads per sample were rarefied to 15000 reads prior to the analysis to ensure the clusters were not biased by sequencing depth per sample. Sequences were clustered at 100% similarity within QIIME 1.9.1 (Caporaso et al., 2010) in order to collapse identical amplicon sequences in pseudo-OTUs to reduce computational time. Taxonomy was assigned to each

unique amplicon sequence by searching the UNITE ITS database (version 7.2, dynamic) (Kõljalg et al., 2005) with megablast (Camacho et al., 2009). Additional ITS sequences (n=44) were obtained from NCBI and added to the UNITE ITS blast database to cover fungal species present in the UK Plant Health Risk Register, however 22 risk register pathogens had no ITS1 sequence available. The data for each sample was rarefied to 15,000 reads and counts for each taxon identified within each sample were outputted with the summarize_taxa script within QIIME1 and imported into R (R Core Team, 2017) for downstream investigation and analysis. The version of the United Kingdom risk register used was downloaded on 13/07/2017 from https://secure.fera.defra.gov.uk/phiw/riskRegister/.

Statistical comparisons between samples, dates and locations were carried out in R (R Core Team, 2017) with the packages ape (Paradis et al., 2004), vegan (Oksanen et al., 2007) and picante (Kembel et al., 2010). The Bray-Curtis dissimilarity metric (Bray and Curtis, 1957) was used to calculate community and sample dissimilarities. Non-metric multidimensional scaling (NMDS) ordination was calculated with the metaMDS function within vegan. Identification of species contributing to location, date and cluster differences was carried out with SIMPER (Clarke, 1993) which assesses dissimilarity between groups.

Results and Discussion

Spatial and temporal characterisation of fungal communities

Figure 4.1: Location of the sampling sites (1) Haywold; (2) Stokesley; (3) Wansford; (4) Alford; (5) Beverley and (6) York. York and Beverley (Light blue) samples were collected as part of the national pollen monitoring network sample collection.

The location of each air sampler is shown in Figure 4.1. In order to identify any location specific grouping within the dataset, hierarchical clustering of Bray-Curtis dissimilarity distances was performed on the rarefied dataset. The clustered dendrogram (Figure 4.2) shows a number of groups; however, none of the groups correlated with sampling location or date upon initial inspection. The original number of reads which passed the stringent QC process are shown in Table 4.1.

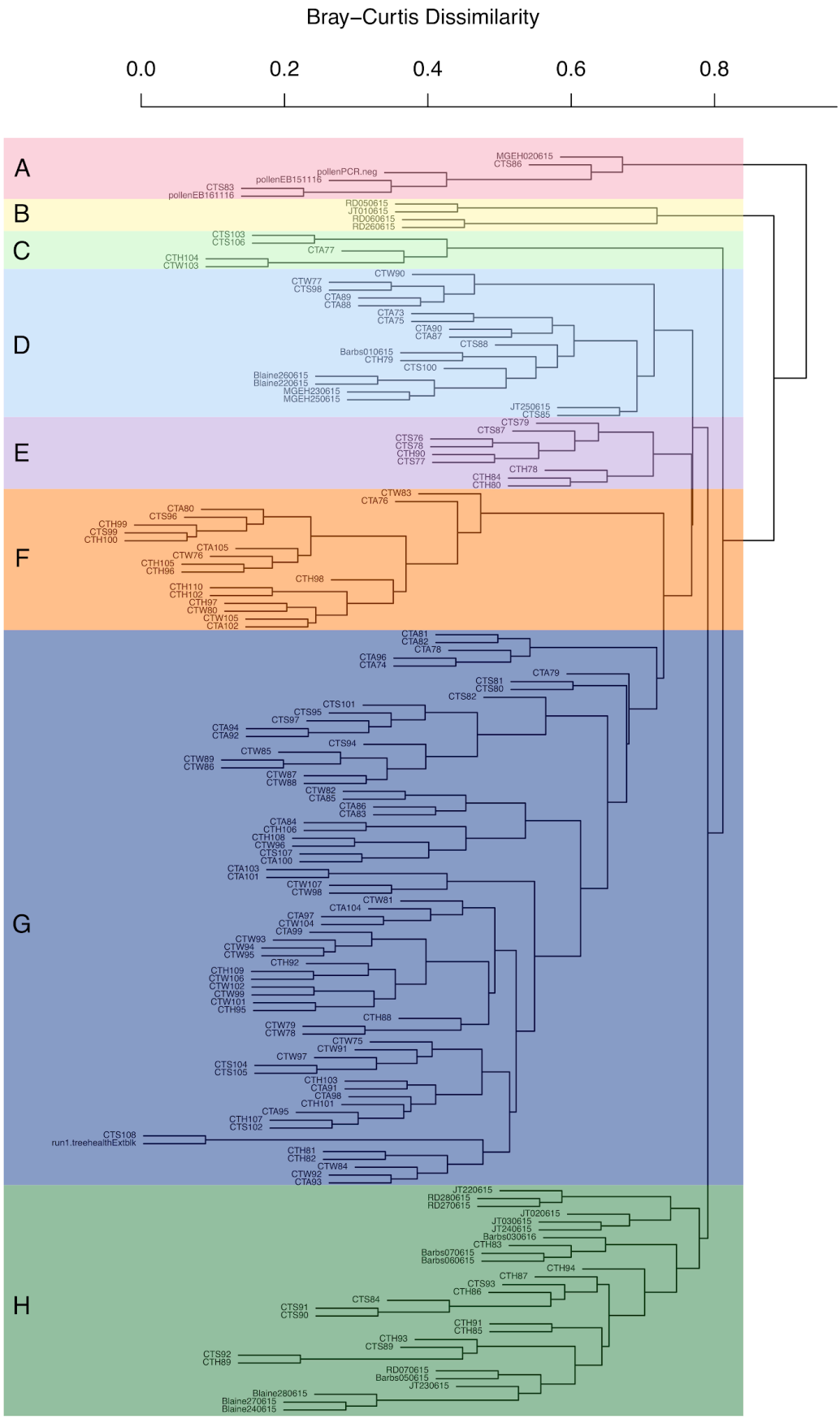


Figure 4.2: Hierarchical clustering of Bray-Curtis dissimilarities across all samples with rarefaction to 15000 reads per sample. Whilst 8 groups are observed, A-H, these are not consistent with either sampling location or date.

The hierarchical clustering by Bray-Curtis dissimilarities demonstrated 8 separate groups, A-H (Figure 4.2). Four control samples contained enough sequence reads to be included in the rarefied dataset and three of the controls (pollenPCRneg, pollenEB151116, pollenEB161116) were found to be clustered with sample CTS83. This would suggest that the three controls were contaminated with DNA from this sample, although at an unknown time in the protocol. Unfortunately the staff who carried out the DNA extraction, PCR amplification and library preparation did not keep records of any potential contamination events they may have noticed. The fourth control sample containing significant sequence reads was found to cluster closely with sample CTS108. All other controls had fewer than 15,000 sequence reads in total and so were excluded from the rarefied dataset (Table 4.1).

Table 4.1: Number of Illumina reads which passed quality control for each sample (prior to rarefaction), organised by date and sampling location

Date	Alford	Alford seqs	Wandsford	Wandsford seqs	Haywold	Haywold seqs	Stokesley	Stokesley seqs	York	York seqs	Beverley	Beverley seqs
29/05/2015	CTA73	157027	CTW75	153045	CTH78	356849	CTS76	247450	-	-	-	-
30/05/2015	CTA74	169135	CTW76	30391	CTH79	282043	CTS77	242332	-	-	-	-
31/05/2015	CTA75	182287	CTW77	268116	CTH80	302439	CTS78	299577	-	-	-	-
01/06/2015	CTA76	130168	CTW78	382553	CTH81	446821	CTS79	287996	Barbs010615	288477	JT010615	219706
02/06/2015	CTA77	225421	CTW79	329305	CTH82	601821	CTS80	227587	MGEH020615	216420	JT020615	293218
03/06/2015	CTA78	124967	CTW80	79009	CTH83	308456	CTS81	197943	Barbs030616	429080	JT030615	670384
04/06/2015	CTA79	250379	CTW81	47090	CTH84	468103	CTS82	274677	MGEH040615	15023		
05/06/2015	CTA80	226516	CTW82	218467	CTH85	407387	CTS83	69387	Barbs050615	332815	RD050615	256300
06/06/2015	CTA81	291303	CTW83	198601	CTH86	166660	CTS84	66352	Barbs060615	766707	RD060615	313962
07/06/2015	CTA82	69831	CTW84	37611	CTH87	183346	CTS85	185581	Barbs070615	317811	RD070615	310324
08/06/2015	CTA83	397147	CTW85	160384	CTH88	205071	CTS86	194760	-	-	-	-
09/06/2015	CTA84	183337	CTW86	153019	CTH89	232884	CTS87	156601	-	-	-	-
10/06/2015	CTA85	90477	CTW87	147420	CTH90	178862	CTS88	139351	-	-	-	-
11/06/2015	CTA86	138116	CTW88	140409	CTH91	114410	CTS89	167227	-	-	-	-
12/06/2015	CTA87	125890	CTW89	166792	CTH92	244333	CTS90	82715	-	-	-	-
13/06/2015	CTA88	189287	CTW90	179586	CTH93	74423	CTS91	167697	-	-	-	-
14/06/2015	CTA89	248099	CTW91	157029	CTH94	173105	CTS92	59399	-	-	-	-
15/06/2015	CTA90	240290	CTW92	113446	CTH95	151997	CTS93	188403	-	-	-	-
16/06/2015	CTA91	203632	CTW93	185098	CTH96	55031	CTS94	236433	-	-	-	-
17/06/2015	CTA92	115514	CTW94	171901	CTH97	92475	CTS95	347445	-	-	-	-
18/06/2015	CTA93	282953	CTW95	182507	CTH98	133052	CTS96	231948	-	-	-	-
19/06/2015	CTA94	201403	CTW96	164224	CTH99	186364	CTS97	381527	-	-	-	-
20/06/2015	CTA95	136462	CTW97	103895	CTH100	186981	CTS98	380563	-	-	-	-
21/06/2015	CTA96	119858	CTW98	207509	CTH101	492936	CTS99	189327	-	-	-	-
22/06/2015	CTA97	441346	CTW99	641375	CTH102	76102	CTS100	35536	Blaine220615	312007	JT220615	258546
23/06/2015	CTA98	183226	CTW100	0	CTH103	184664	CTS101	540500	MGEH230615	191197	JT230615	308935
24/06/2015	CTA99	178620	CTW101	400402	CTH104	546767	CTS102	677130	Blaine240615	334246	JT240615	294607
25/06/2015	CTA100	275301	CTW102	436225	CTH105	385273	CTS103	502582	MGEH250615	327792	JT250615	333630
26/06/2015	CTA101	230052	CTW103	345091	CTH106	949404	CTS104	576966	Blaine260615	1145941	RD260615	460166
27/06/2015	CTA102	502355	CTW104	253333	CTH107	400369	CTS105	576635	Blaine270615	703666	RD270615	106999
28/06/2015	CTA103	406151	CTW105	263791	CTH108	437232	CTS106	491272	Blaine280615	1611508	RD280615	269096
29/06/2015	CTA104	272378	CTW106	61028	CTH109	146687	CTS107	234978	-	-	-	-
30/06/2015	CTA105	185380	CTW107	202275	CTH110	202534	CTS108	123233	-	-	-	-

Testing of associations between clusters, sampling location and sampling date with adonis/PERMANOVA (Anderson, 2001) determined that only 6.2% of the differences between samples could be attributed to sampling location (p=0.001) and only 21% could be attributed to date (p=0.23). The same hierarchical testing was applied to the individual clusters A-H (Table 4.2) and statistically supported associations were discovered for sampling location in clusters G and H and for sampling date in cluster E. However, despite a link between the samples in cluster G and their original location, only 12% of sample clustering could be explained this way (n=71), and similarly only 18% for cluster H. In contrast, the

association between samples in cluster E and their collection date is present for 83% of the samples (n=9).

Table 4.2: Adonis (PERMANOVA) testing of association between samples in a cluster (as determined by Bray-Curtis dissimilarities) and location or date. Statically supported associations are highlighted in green (Clusters G and H samples by location; cluster E samples by date).

Cluster (samples)	**A (n=6)**	**B (n=4)**	**C (n=5)**	**D (n=19)**	**E (n=9)**	**F (n=18)**	**G (n=71)**	**H (n=29)**
R^2 (location)	0.56	-	0.79	0.32	0.11	0.13	0.12	0.18
p-value (location)	0.1	-	0.4	0.15	0.7	0.8	0.001	0.003
R^2 (date)	0.86	1	1	0.79	0.83	0.89	0.43	0.54
p-value (date)	0.1	1	1	0.32	0.02	0.41	0.67	0.44

The detailed hierarchical clustering for clusters E, G and H are shown in Figure 4.3a, Figure 4.3b and Figure 4.3c respectively. While there were clear visual associations between some samples, for example the similarity between the blue Beverley samples and orange York samples in Figure 4.3c, this served to highlight the lack of statistical association between most samples based on location in clusters H and G. Overall, the hierarchical clustering and statistical analysis showed very little support for correlation between samples and their location and/or date.

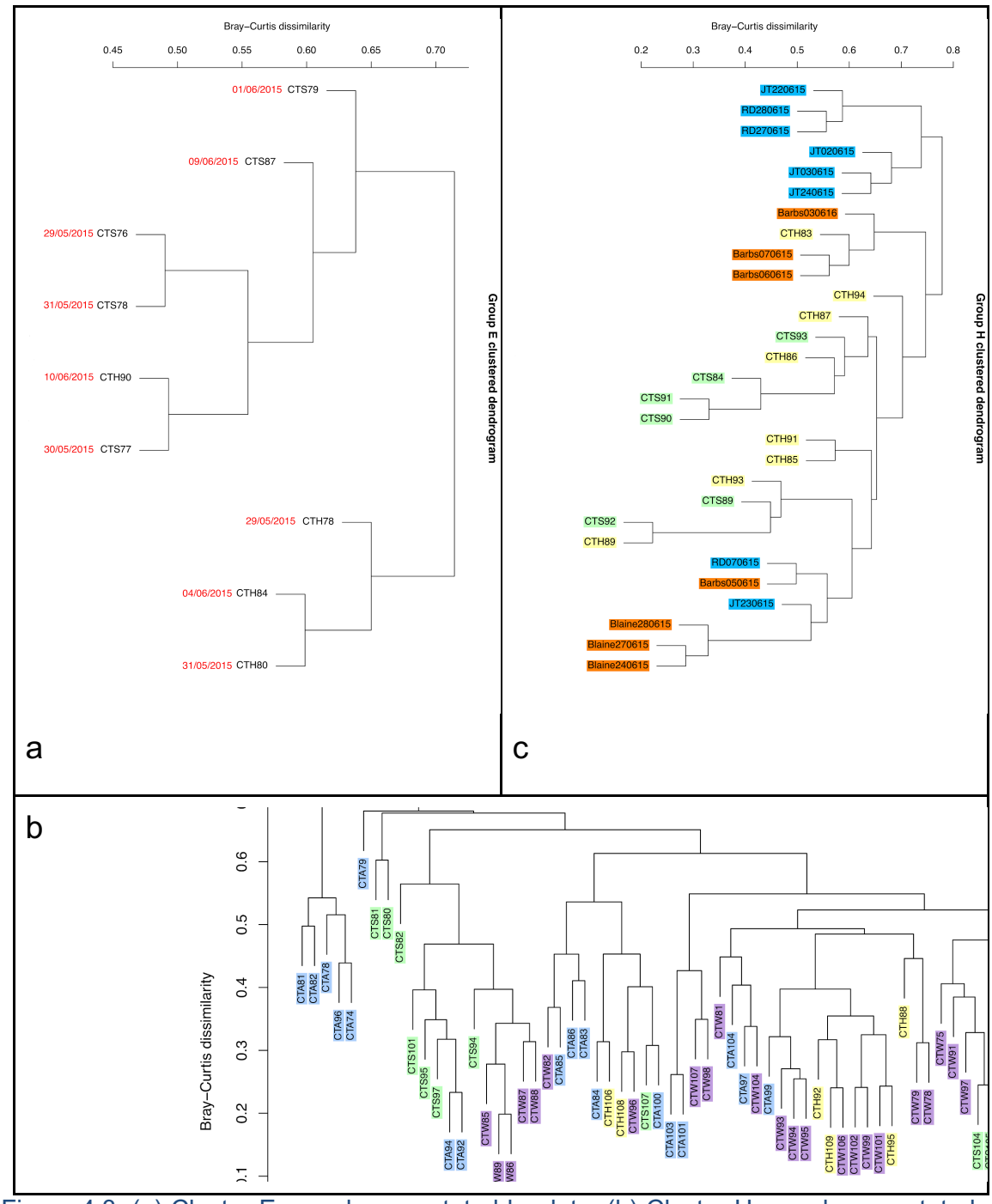


Figure 4.3: (a) Cluster E samples annotated by date. (b) Cluster H samples annotated by location (Blue=Beverley, Orange=York, Yellow=Haywold, Green=Stokesley). (c) Cluster G samples annotated by location (Pale Blue=Alford, Purple=Wansford, Red=Contaminated control

The clustering of the data into eight clusters, A-H, demonstrated that there was enough variability between samples for comparison. However, the initial analysis showed no support for correlation between the six sampling locations, nor the dates over the month sampled. This was interesting given the sampling locations were in some cases hundreds of miles apart, with two locations being in urban areas, as the York and Beverley air samplers used in pollen studies were on top

of buildings. The samples in cluster H appeared to show some supported within-cluster differentiation of York and Beverley from the field-based samples (Wansford, Alford, Stokesley and Haywold). However, this represented a small amount of the dataset and was not further supported by differences between sampling dates. It did suggest that there are location-specific taxonomy differences even if they are being lost in the noise of contamination subsequently found in the dataset.

Identification of contamination in samples and controls

With almost no supported correlation between the data and either location or date, and the presence of contamination, further interrogation of the species composition was required. We hypothesised that the samples had been contaminated during storage, DNA extraction, PCR amplification or library preparation. The three controls remaining in the dataset after rarefaction which were placed in cluster A by hierarchical clustering of the Bray-Curtis dissimilarities were two DNA extraction negative controls and a PCR negative control. These three controls were all from the pollen slide samples (York and Beverley sampling sites) and had similarity in species composition to samples CTS83, CTS86 (both Burkard samples) and MGEH020615 (pollen slide sample). This suggested contamination of both pollen and Burkard samples during both DNA extraction *and* PCR amplification given the pollen slide and Burkhard extractions were carried out on different days by different people. The DNA extraction control from the Burkard samples (Alford, Stokesley, Haywold and Wansford) was also positive with significant similarity in species composition to sample CTS108.

The evidence of both insidious and gross contamination in the results was worrying and changed the course of the analysis dramatically. We were not in a position to carry out the detailed spatial and temporal analysis as the samples were clustering based upon contamination events rather than true diversity, a result supported by PERMANOVA testing. The removal of suspected contaminating sequences would have reduce the sequences in most datasets to very low levels which would have further weakened a downstream analysis. The United Kingdom Plant Health Risk Register describes 125 fungal species but reference DNA barcodes were not available in the public databases for 22 of these species, which represents a gap in our results. These species may be

present in varying quantities, yet described as “unidentified”. UNITE provides many sequences for unidentified species and it may be that some of the Risk Register species are within these groups, for example, “unidentified *Pleosporales”*, which was very abundant over the whole dataset. The random way in which some species were present in samples from very different dates and very different locations is stark (e.g. *Endocronartium harkenessii*), and, given the large abundance of contaminating species suggests gross contamination of these samples through unknown mechanisms. Other forms of contamination were more subtle and perhaps the result of batch effects resulting from DNA extraction orPCR amplification.

Identification of species driving the observed clustering of samples and potential batch effects

SIMPER analysis determined a number of species across all clusters that contribute at least 70% to the difference between clusters A-H: *Endocronartium harknessii*, *Mycosphaerella tassiana*, *Cryptococcus pseudolongus*, an unidentified *Mycosphaerellaceae* species, *Microdium phyllanthi*, *Blumeria graminis*, an unidentified *Pleosporales* species, *Elsino australis*, *Aspergillus cibarius*, *Sporobolomyces roseus* and *Puccinia striiformis*. Many other species with a lesser effect on the clustering were also elucidated with SIMPER; however, interpretation proved difficult. SIMPER determines the percentage by which each individual species is responsible for within-group and between-group differences, so species which are more abundant contribute more to the outputted difference. In this case the number of samples and the number of species present within them made elucidating any additional species presence or abundance subtleties that contributed to the clusters very difficult.

To further investigate the effect of overly-abundant species on the clustering, the rarefied counts of each species in each sample were plotted with the hierarchical clustering of the Bray-Curtis dissimilarities (Figure 4.4). This identified the species responsible for certain clusters. Cluster F is characterised by very large relative abundance of *Endocronartium harknessii* (>80% in some samples). *Cryptococcus pseudolongus* influences the groups within clusters A and H. The abundance of *Elsinoe australis* in five samples explains cluster C; however, these samples were from four different locations on four different dates. There is an

interesting relationship between *Blumeria graminis, Endocronartium harknessii* and *Cryptococcus pseudolongus* across the samples: where *B. graminis* and *E. harknessii* are present, *C. pseudolongus* is absent. *Sporobolomyces roseus* was present in almost all samples in varying abundances with the largest abundances observed within five samples in cluster D. This may represent a batch effect where *B. graminis* and *E. harknessii* were introduced during one set of DNA extractions and *C. pseudolongus* was introduced during another set of DNA extractions. Unfortunately, the researcher who carried out the DNA extractions did not keep a record of which samples were extracted together. The unidentified *Pleosporales* was also present in almost all samples in varying abundances but with larger abundances within cluster G.

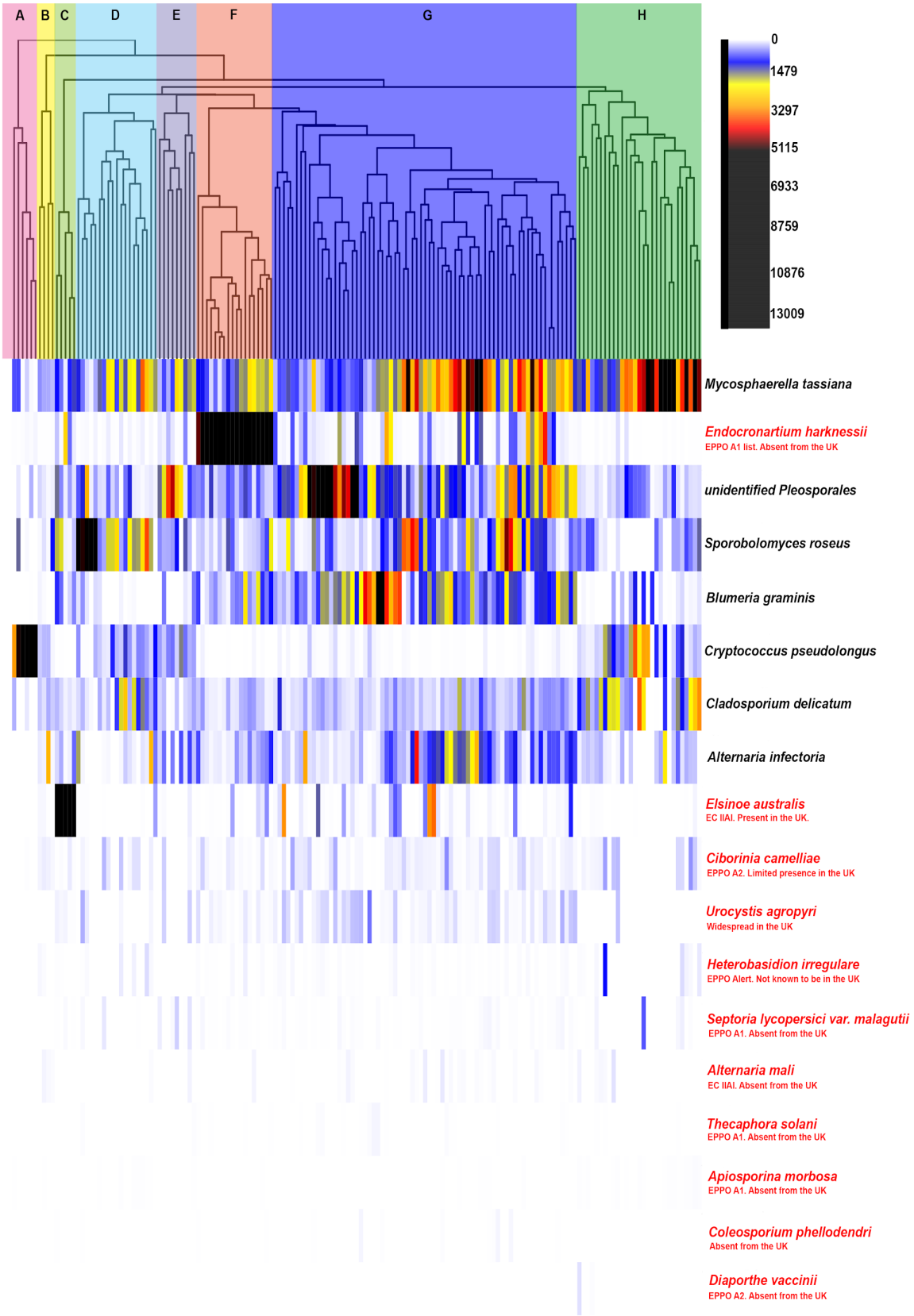


Figure 4.4: Heatmap of the rarefied read numbers (max=15000) for the 8 most abundant species detected in the whole dataset (labelled in black) and for species found on the UK Risk Register (labelled in red). The hierarchical clustering and groups A-H are also shown to demonstrate the effect of single species on the clustering across the dataset.

Batch effects are well known in genomic datasets (Parker and Leek, 2012) and attempts have been made to develop algorithms to counter their effect on different types of genomic datasets (Akulenko et al., 2016; Gibbons et al., 2017; Manimaran et al., 2016; Nyamundanda et al., 2017). Despite the progress made with tools for identifying and reducing batch effects in expression studies with limma (Ritchie et al., 2015) and Combat (Johnson et al., 2007), there are few such tools for metabarcoding studies. Hypothetically, a positive mock community control, a negative DNA extraction control, a negative PCR amplification control and a negative PCR library preparation control would all be taken through the metabarcoding pipeline to determine contaminants from the laboratory environment, reagents and kits. Any contaminants discovered in these controls would then be subtracted from the main dataset prior to analysis in order to minimise batch effects across the dataset. This strategy was first proposed by (Salter et al., 2014) who determined that contaminating DNA present in reagents (the “kitome”) can have a significant effect on experiments with low biomass samples. They published a list of contaminating genera they had found in negative controls from various kits and found that while contaminants were predominantly environmental bacteria (rather than the human samples they were studying) the contamination between batches/lots of consumables was not predictable. Recent studies have also confirmed that common routes of contamination in metabarcoding experiments are DNA extraction and PCR amplification, and that kit box number or even reagent lot number can affect the taxonomic composition of a sample (de Goffau et al., 2018; Glassing et al., 2016; Kim et al., 2017). Mock communities as positive controls have been shown to help understand the effects of PCR amplification and accurate community representation in the final data (Bakker, 2018), however no such control was used in this experiment due to the diversity in the range and abundance of species expected to be found. However, in the fungal spore data analysed here the abundance of species associated with batch effects was considerable and removal of these species would have significantly weakened the dataset further.

United Kingdom Plant Health Risk Register plant pathogens present in the samples

The United Kingdom Plant Health Risk Register is a register of pests and pathogens of significance to plant health and includes species which are already

known to be present in the UK as well as those which have yet to arrive and which are hoped will never arrive given their pathogenicity. A number of fungal pathogens present on the United Kingdom Plant Health Risk Register (DEFRA) were expected to be found in the dataset due to their widespread distribution in the United Kingdom. However, a number of highly quarantine species believed to be absent from the UK were present in the rarefied sequence data (Figure 4.4 and Table 4.3).

Further interrogation of the rarefied sequence counts, sampling date and sample location for *Endocronartium harknessii* showed a random distribution of large abundances of this species strongly suggesting laboratory contamination rather than presence in the field (Figure 4.5). Given its EPPO A1 list status as a quarantine pathogen, it is highly unlikely that this pathogen comprised 8.4% of the spores collected in the field. Even if the sequence reads assigned as *E. harknessii* originated from a closely related species, rather than the quarantine species, the introduction of the organism to the dataset looks like a random spotting of huge abundances of this organism, rather than a gradually increasing or decreasing level, or a steady background level as had been observed for other species in the dataset. For example, the sequences classified as *E. harknessii* comprised more than 14,000/15,000 on some days and 0/15,000 the next.

Table 4.3: Fungal species present on the UK Risk Register (accessed 13-07-2017) and also present within the rarefied dataset. The EPPO/EU classification and their presence/absence in the UK is also listed.

Name	EPPO listing	Presence in the UK	Relative abundance (in all samples)
Alternaria mali	*A1*	*Absent*	*0.039%*
Alternaria panax	*-*	*Absent*	*0.003% (only in samples JT240615 and CTW107)*
Apiosporina mobosa	*A1*	*Absent*	*0.011% (found primarily in the pollen slide samples)*
Ciborinia camelliae	*A2*	*Present (limited distribution)*	*0.376% (see figure 4)*
Coleosporium phellodendri	*-*	*Absent*	*0.007% (found in four unrelated samples)*
Cronartium quercuum	*A1*	*Absent*	*0.0001%*
Diaporthe vaccinii	*A2*	*Absent*	*0.006% (only in sample JT220615)*
Elsinoe australis	*A1*	*Present*	*2.596% (see figure 4)*
Endocronartium harknessii	*A1*	*Absent*	*8.455% (see figure 4)*
Gymnosporangium asiaticum	*A2*	*Absent*	*0.00004%*
Heterobasidion abeitium	*-*	*Absent*	*0.00008%*
Heterobasidion irregulare	*Alert*	*Distribution unknown*	*0.099% (found primarily in the pollen slide samples)*
Heterobasidion parviporum	*-*	*Absent*	*0.002% (only in samples CTA91 and Blaine280615)*
Hymenoscyphus fraxinus	*A2/Alert*	*Present (limited distribution)*	*0.0005% (only found in sample CTH100)*
Kabatiella zeae	*-*	*Present (limited distribution)*	*0.0001%*
Neonectria neomacrospora	*-*	*Present (unknown distribution)*	*0.00004%*
Ophiognomonia claviginenti-juglandacearum	*A1*	*Absent*	*0.00008%*
Podosphaera euphorbiae-helioscopiae	*-*	*Present (unknown distribution)*	*0.00004%*
Puccinia graminis	*-*	*Absent*	*0.00025%*
Puccinia komarovii	*-*	*Present (unknown distribution)*	*0.00008%*
Septoria lycopersici var malagutii	*A1*	*Absent*	*0.084% (see figure 4)*
Stemphylium vesicarium	*-*	*Present (widespread distribution)*	*0.0003%*
Thecaphora solani	*A1*	*Absent*	*0.011% (see figure 4)*
Tilletia indica	*A1*	*Absent*	*0.0005%*
Urocystis agrophyri	*-*	*Present (widespread)*	*0.304% (see figure 4)*
Verticillium albo-atrum	*A2*	*Present (widespread)*	*0.001% (only in sample JT030615)*
Verticillium dahliae	*A2*	*Present (widespread)*	*0.0003%*

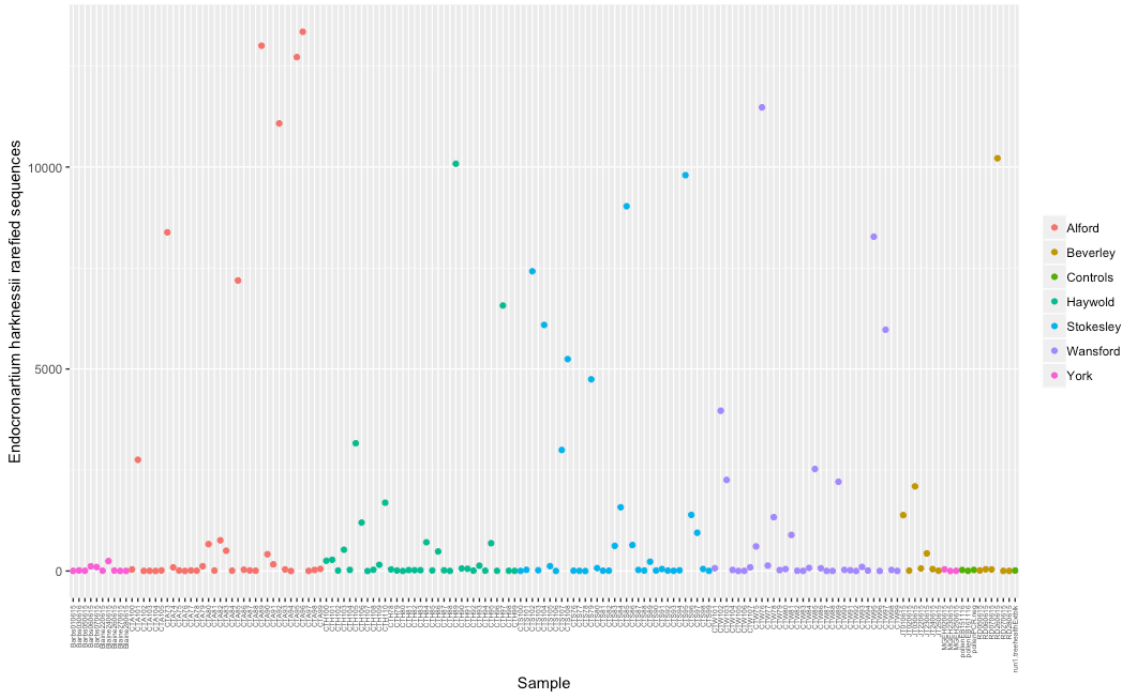


Figure 4.5: Rarefied counts of *Endocronartium harknessii* by sample and location. Despite being an EPPO A1 listed quarantine pathogen which is absent from the UK, this species constituted 8.4% of all sequences in the dataset. This is an overt example of the laboratory-acquired contamination during DNA extraction. Note that many samples are negative for this pathogen.

We plotted the percentage similarities between each read and its corresponding database reference sequence for all Risk Register species identified in >0.01% abundance across the dataset (Figure 4.6). This analysis demonstrated that the identifications of sequences as *Elsinoe australis*, *Endocronartium harknessii* and *Septoria lycopersici* may not be as robust as might be assumed without this check. However, the plot also demonstrates that the identifications for the majority of sequences assigned as *Alternaria mali*, *Apiosporinia morbosa*, *Ciborinia camellae, Diaporthe vaccinii, Heterobasidion irregulare, Thecaphora solani* and *Urocystis agroyri* are robust. The “comet trails” below the box plots in Figure 6 represent sequences with lower sequence identity that were assigned that that species. Given the comprehensive reference database used, UNITE, which also included references for unidentified species, it is likely that the majority of lower percentage identifications represented sequence errors.

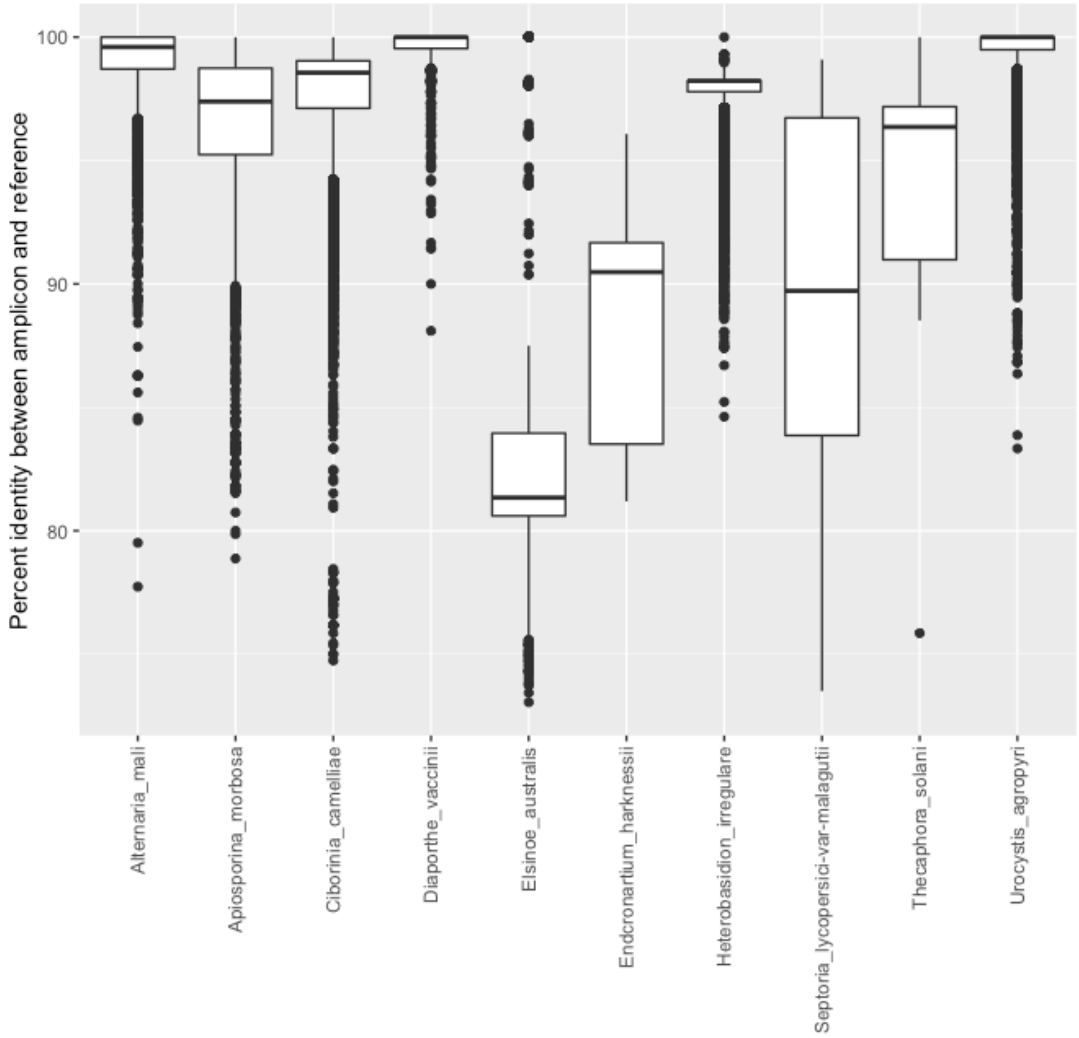


Figure 4.6: Boxplots of the percent identity between the reference sequence used for identification and the sequence read being identified. This figure demonstrates that the identifications for *Elsinoe australis, Endocronartium harknessii* and *Septoria lycopersici* may not be as robust as those for other species.

A species of particular interest at the time of sampling was the causative pathogen of ash dieback, *Hymenoscyphus fraxinus*. However, it was only detected in very low abundance in a single sample using short-read Illumina sequencing and due to sequencing limitations that sample was not included for long-read nanopore sequencing. The low abundance of *H. fraxinus* was unexpected, but it is possible that the weather was not conducive to the transport of *H. fraxinus* spores during the sampling period.

Conclusions and recommendations

In contrast to the published studies on metabarcoding contamination (Salter et al., 2014; Weiss et al., 2014), the major contaminating species in our data were unlikely to have been introduced by the kits or reagents used during DNA extraction as they are quarantine species under licencing restrictions (*Heterobasidion irregulare, Alternaria mali, Thecaphora solani, Apiosporina morbosa, Coleosporium phellodendri* and *Diaporthe vaccinii*) and thus would be highly unlikely to be present in the kit manufacturing facilities. However, they could have been introduced during DNA extraction, PCR amplification and library preparation from the laboratory environment itself as these steps were carried out in quarantine licensed facilities where fungal pathogens were actively being worked with, and which had been grown and worked with previously. None of these species are known to be present in the United Kingdom and if found to be present would require significant response from the UK government's Animal and Plant Health Agency for management and control. This presents an interesting conundrum: does our data show a low level of these quarantine pathogens in the field, in which case they would be found in multiple locations on different dates at low level, or, does the data show that there is a low background level of contamination introduced from the laboratory in which DNA extraction was carried out. We believe that the latter is considerably more likely; however, without further field samples being collected and tested with more robust diagnostic methods such as validated species-specific real-time PCR assays, the former cannot be fully ruled out. Further investigation with the researchers who carried out the sample collection, DNA extraction and library preparation revealed that DNA extraction had been carried out in a quarantine laboratory where fungal cultures were present and being actively worked with. To this end, we propose that aerosol contamination from the laboratory itself is an additional significant route for contamination of samples being prepared for metabarcoding and may explain why the "kitome" (Salter et al., 2014) varies so unpredictably between kits and batch numbers: the "labome" may be an additional factor contributing to the variability.

A number of studies have discussed the considerations for reducing contamination in metabarcoding studies. In environmental DNA (eDNA) studies, Goldberg et al (2016) recommend that negative field controls be used, field

equipment and field staff should be separated from the testing laboratory prior to sampling and analysis. Their recommendation of negative field controls (clean water collected using the same protocol and equipment as field samples) would have been difficult to achieve for air sampling in the field. However, given the hypothesis of aerosol contamination within the laboratory environment it would have been an improvement on our methods to include a 24h air sample from within the lab to assess the background of fungal spores in the laboratory. Goldberg et al (2016) and Salter et al (2014) both make similar recommendations for reducing the level of background contamination in eDNA metabarcoding and clinical 16S sequencing, respectively. Both papers stress the importance of "clean laboratory" protocols at all stages of the process, to the extent that staff travelling from laboratories where PCR products or DNA extractions are handled to the clean laboratory should shower and change into fresh clothing before entering the clean laboratory. They also recommend that deep cleaning using 50% bleach solutions and UV treatment of surfaces should also be regularly carried out.

In addition to the general laboratory considerations stated above, should the contamination not be introduced by the random aerosol introduction of spores into tubes but by processes we recommend an additional number of basic, and rather obvious, laboratory processes to reduce contamination and enable batch effects to be more easily identified in the downstream data. The following recommendations would have enabled the contamination found in our fungal dataset to be identified and characterised more readily and accurately.

1. *Use of an incremental numbering system for samples*

   Given 100 samples, they should be numbered 1-100, independent of their actual sample names or numbers in a LIMS system. All manipulations, for example pipetting of DNA extraction reagents, should be carried out in numerical order. This would allow the downstream identification of any dilution effect on a single tube being contaminated as manipulations are carried out in the same order.

2. *Accurate record keeping of which samples were tested together.*

This should include which samples were extracted and by what method, the lot numbers and kits used, along with any differences from protocol noted at the time.

3. *DNA extractions should not be carried out in a laboratory where growth of fungal cultures or bench-top analysis of fungal spores is being carried out.*

   In scenarios such as ours, where we were actively looking in the downstream dataset for the presence or absence of particular pathogens, sample storage and DNA extraction should not be carried out in a laboratory where these pathogens might be present. Goldberg et al. (2006) state implicitly that separate rooms should be available for pre- and post-PCR steps, potentially with positive air pressure and filtration.

4. *Staff training should be kept up-to-date and carried out by the most experienced senior scientist in the lab*

   Refresher sessions on good pipetting practice, the effect and impact of contamination on large projects, and a culture where reporting of contamination events is welcomed and encouraged. The same person should be responsible for training all staff to reduce the dilution effect of second-hand training. Senior scientists should not assume that a one-off training session is enough to reduce contamination events in the long term; continuous training and improvement should be encouraged.

5. *Routine swabbing and metabarcoding monitoring of laboratories and equipment where DNA extraction, PCR reactions and library preparation are carried out.*

   Ideally if bacterial 16S, fungal ITS or any other loci is routinely used within a sequencing facility, those metabarcoding experiments should be carried out on the monitoring swabs. This would monitor background equipment contamination levels and illuminate issues prior to them appearing in datasets.

In conclusion, metabarcoding is an incredibly promising tool for surveillance and monitoring of fungal plant pathogens in the field. However, the propensity for contamination-based batch effects needs to be fully understood by researchers wanting to utilise it - this also applies to metagenomic studies which are subject

to the same effects. We have made a number of recommendations to aid the identification of contamination-based batch effects which could be easily introduced into standard operating procedures in most laboratories.

Chapter 5. The potential for full ribosomal tandem repeat metabarcoding

Introduction

It has been estimated that there are between 2.2 to 3.8 million fungal species, with only an estimated 3-8% having been named (Hawksworth and Lücking, 2017). The ability to discriminate species by their DNA sequence can be limited by the choice of genetic/genomic locus. If the locus does not contain enough variable positions within the length of sequence amplified then species- or isolate-level identification may not be possible. The locus also needs to have enough conservation to allow the design and use of universal primers. However, universal primers, amplicon length and the variability within the locus are all assessed based upon knowledge of the species which have already been described. Recent studies have shown that there are likely to be many fungal species which have not yet been identified and are completely unknown (Tedersoo et al., 2017). With so many species yet to be described, the longer the sequence used to first identify it, the better it can be placed within the current taxonomy.

Studies utilising targeted Illumina sequencing of the ribosomal internal transcribed spacers (ITS) have demonstrated the ability of this technique to resolve fungal species, including the indoor environment (Korpelainen and Pietilainen, 2015; Korpelainen et al., 2015), soil (Schmidt et al., 2013), urban environments (McGuire et al., 2013) and even hot water systems (Ma et al., 2015). The ability of Illumina metabarcoding to be used for environmental fungal spore monitoring for plant pathogens has been demonstrated in Chapter 4, despite the contamination observed. However, targeted Illumina sequencing is subject to a number of disadvantages when applied to the surveillance for unknown or emerging threats. The restriction in amplicon length for Illumina sequencing can reduce the ability to resolve some species and the requirement for conserved primers flanking variable regions can drastically reduce candidate loci for sequencing studies. The clustering of quality-filtered sequence reads into operational taxonomic units (OTUs) - often at 97% similarity - can cluster separate sub-species and isolates into the same OTU. This loss of resolution can

be negated by clustering at 100%, thereby only clustering identical sequences for computational efficiency, but should species, sub-species or isolates have the same ITS1/ITS2 sequence they are still impossible to discriminate. The introduction of long-read sequencing technologies such as PacBio and nanopore negates the short-read disadvantages and allow the sequencing of longer regions, however with their own set of disadvantages, including error rate (Jain et al., 2017).

The nuclear ribosomal tandem repeat contains genes for the ribosomal 18S small subunit (SSU), 28S large subunit (LSU) and 5.8S subunit. These genes are interrupted by the two internal transcribed spacers, ITS1 and ITS2, and the non-transcribed intergenic spacer (IGS). A fourth ribosomal gene, 5S, is present in some fungal groups, splitting the IGS into two spacers, IGS1 and IGS2, with the 5S rDNA being transcribed by a different RNA polymerase to the 18S, 5.8S and 28S rDNA (Bergeron and Drouin, 2008). The ITS regions can vary significantly in length between fungal species (Hausner and Wang, 2005; Taylor and McCormick, 2008; Tedersoo et al., 2015). The ITS regions are hypervariable, yet also contain more conserved regions, and form secondary structures (Mullineux and Hausner, 2009; Nazar, 2004; Rampersad, 2014). The ribosomal genes - LSU, SSU, and 5.8S - are highly conserved and are routinely used for the forward and reverse primers for ITS sequencing studies. The ribosomal genes and spacers are present in multiple copies in the nuclear genome as a tandem repeat. Their abundance has made them an attractive region for species identification studies due to their ease of amplification and variability. The ribosomal tandem repeat regions have been used extensively in the past for species identification, with different regions being used for different taxa prior to the adoption of a more standardised approach to fungal identification using ITS (Schoch et al., 2012).

In the late 20th century, the IGS region was commonly amplified and digested with restriction enzymes (RFLP) for species and isolate discrimination (Erland et al., 1994; Gardes and Bruns, 1996; Henrion et al., 1992). The IGS region varies in length from around 2kb in yeast to 21kb in mammals and the length is determined by the number of repeats present within the region (Moss and Stefanovsky, 1995). Due to its hypervariability, the IGS region has been used to discriminate not only species but isolates too (Pantou et al., 2003). Despite these

obvious advantages for species discrimination, the IGS region is not a panacea. Its length can make amplification problematic and the resulting product is too long to be sequenced with Sanger sequencing. However, with the newer long-read sequencers this region shows renewed promise as an additional marker for biosurveillance and monitoring.

Short-read next-generation sequencing technologies have enabled the characterisation of fungal communities in many different environments using conserved primers to amplify the ITS regions (Ghannoum et al., 2010; Korpelainen and Pietilainen, 2015; Li et al., 2016; Mosier et al., 2016; Sugiyama et al., 2010). However, there are disadvantages to using such short loci for community assessments, including the inability to discriminate the species of certain genera and primer bias during amplification (Bokulich and Mills, 2013). Such disadvantages are not just applicable to next-generation sequencing technologies but also Sanger sequencing.

The introduction of long-read sequencing methods introduced the potential to sequence the entire ribosomal tandem repeat. A recent study used PacBio long-read sequencing to amplify the 18S-ITS1-5.8S-ITS2-28S section of the repeat in fungi (Tedersoo et al., 2018) but did not attempt to sequence the entire repeat to include IGS. A further recent study sequenced the whole tandem repeat for using nanopore sequencing, but used three overlapping long amplicons and cultured material in order to demonstrate sequencing of the whole repeat for reference sequence database construction (Wurzbacher et al., 2018). Here we demonstrate the utility of long-read nanopore sequencing to characterise the whole ribosomal operon for fungal biomonitoring and surveillance by the metabarcoding of aerosol samples from four locations over one week.

Materials and Methods

Spore sampling

Samples from Wansford, Alford, Haywold, and Stokesley came from the Crop Monitor network (CropMonitor, 2017) of spore samplers located in crop fields (Wansford: Latitude 53.55°N, Longitude -0.43°W; Alford: Latitude 53.26°N, Longitude 0.18°W; Haywold: Latitude 53.99°N, Longitude -0.60°W and

Stokesley: Latitude 54.58°N, Longitude -1.16°W). The samplers were Burkard volumetric Cyclone samplers (Burkard Manufacturing, UK). They continuously sample pollen and fungal spores, which were then deposited into a 1.5ml tube. The tubes were changed every day giving daily samples of spores during the period 29/5/2015 to 30/6/2015. A subset of these samples comprising 24/06/2015 to 30/06/2015 was selected for MinION sequencing.

DNA extraction

Cyclone samples were disrupted in 50ml falcon tubes by adding 1g of an equal weight mixture of 2.3mm and 0.5mm zirconia silica beads (Fisher Scientific, UK) and vortexed for 4 min at full speed on a Vortex Genie II (Fisher Scientific, UK) using a horizontal vortex adapter (Qiagen, UK). The tubes were centrifuged at 5000*g* in a Sigma 4K15 centrifuge and the clear lysate removed. The lysate was then extracted using a nucleospin Plant 2 kit (Macherey-Nagel, Germany) as per the manufacturer's instructions. Buffer-only samples were taken through the complete extraction process as extraction blanks.

Ribosomal tandem repeat region primer design

The GenBank nucleotide database was queried on 25th October 2016 for fungal rDNA tandem repeat sequences up to 5kb in length with the following query which retrieved 236,064 DNA sequences:
(28S[All Fields] AND 18S[All Fields]) AND (fungi[filter] AND biomol_genomic[PROP] AND ddbj_embl_genbank[filter] AND ("1"[SLEN]:"5000"[SLEN]))
As there were too many sequences to produce an alignment to investigate for primer design, a list of existing large rDNA subunit (LSU) primers were retrieved from https://sites.duke.edu/vilgalyslab/rdna_primers_for_fungi/ (Table 5.1). A blast database was created with the GenBank sequences and each of the primers searched against the database with blastn and the number of (unrestricted) hits were recorded. The final selected amplicon sequences were 5.8S (5'-CGCTGCGTTCTTCATCG-3') and 5.8SR-MinION (5'-TGCSRGARCCAAGAGATCCG-3').

Table 5.1: Primers investigated for their efficacy in amplifying large numbers of fungal species

Primer name	Sequence (5'-->3')	Position within *S. cereviseae* rRNA
5.8S	CGCTGCGTTCTTCATCG	51-35 (5.8S RNA)
5.8SR	TCGATGAAGAACGCAGCG	34-51 (5.8S RNA)
LR0R	ACCCGCTGAACTTAAGC	26-42
LR1	GGTTGGTTTCTTTTCCT	73-57
LR2	TTTTCAAAGTTCTTTTC	385-370
LR2R	AAGAACTTTGAAAAGAG	374-389
LR3	CCGTGTTTCAAGACGGG	651-635
LR3R	GTCTTGAAACACGGACC	638-654
LR5	TCCTGAGGGAACTTCG	964-948
LR6	CGCCAGTTCTGCTTACC	1141-1125
LR7	TACTACCACCAAGATCT	1448-1432
LR7R	GCAGATCTTGGTGGTAG	1430-1446
LR8	CACCTTGGAGACCTGCT	1861-1845
LR8R	AGCAGGTCTCCAAGGTG	1845-1861
LR9	AGAGCACTGGGCAGAAA	2204-2188
LR10	AGTCAAGCTCAACAGGG	2420-2404
LR10R	GACCCTGTTGAGCTTGA	2402-2418
LR11	GCCAGTTATCCCTGTGGTAA	2821-2802
LR12	GACTTAGAGGCGTTCAG	3124-3106
LR12R	CTGAACGCCTCTAAGTCAGAA	3106-3126
LR14	AGCCAAACTCCCCACCTG	2616-2599
LR15	TAAATTACAACTCGGAC	154-138
LR16	TTCCACCCAAACACTCG	1081-1065
LR17R	TAACCTATTCTCAAACTT	1033-1050
LR20R	GTGAGACAGGTTAGTTTTACCCT	2959-2982
LR21	ACTTCAAGCGTTTCCCTTT	424-393
LR22	CCTCACGGTACTTGTTCGCT	364-344

MinION nanopore sequencing and analysis

Amplification and MinION nanopore sequencing was carried out at Fera Science Ltd (York, UK) with nanopore R9.4 chemistry. Base calling was carried out with MinKNOW v1.4.3 and the Recurrent Neural Network (RNN) algorithm. FastQ 2D sequences were extracted from the MinION fast5 files using poretools v0.6.0

(Loman and Quinlan, 2014). Taxonomy was assigned to each 2D nanopore sequence by searching the UNITE ITS database (version 7.2, dynamic) (Kõljalg et al., 2005) with blastn (Camacho et al., 2009) with a maximum number of alignments and descriptions set to 50. In addition to isolating the top blast hit from this data, it was imported into MEGAN (Huson et al., 2016) to calculate an identification using the lowest common ancestor (LCA) algorithm. Additional ITS sequences (n=44) were obtained from NCBI and added to the UNITE ITS blast database to cover fungal species present in the UK Plant Health Risk Register, however 22 risk register pathogens had no ITS1 sequence available. GC content and fastq quality scores were calculated with BioPython (Cock et al., 2009) and all results were imported into R for exploratory analysis (R Core Team, 2017). Comparisons with fungal relative abundances obtained with Illumina sequencing from in Chapter 4 were carried out within the vegan package in R (Oksanen et al., 2007) and searches between the Illumina and MinION sequence reads were performed with Blast+ v2.7.1 (Altschul et al., 1990)

Results and discussion

Primer design for the rDNA tandem repeat

The primers listed in Table 5.1 had variable performance when searched against the 236,064 fungal sequences retrieved from GenBank. The 5.8S and 5.8SR primers performed considerably better than primers located within the LSU, those named starting with ‘LR’ in Figure 5.1.

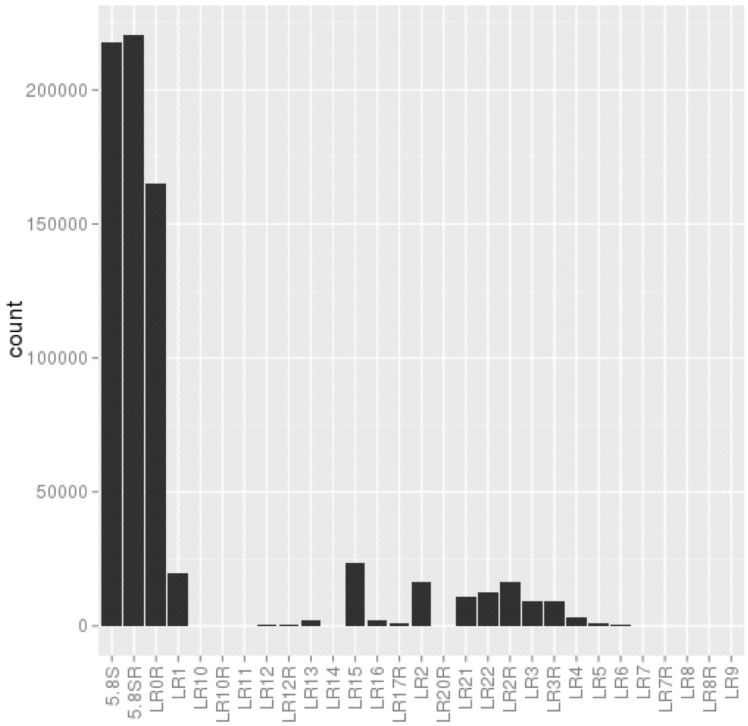


Figure 5.1: Proportion of the 236,064 fungal sequences where each primer sequence was detected

A decision was made to not only amplify the full tandem repeat, but also to design the long PCR amplicon in such a way that it exploited as much taxonomically-informative sequence as possible. The results from the primer search experiment (Figure 5.1) showed superior performance of primers 5.8S and 5.8SR. These two primers were reversed so that the amplified DNA exploited back-to-back primers placed within 5.8S (Figure 5.2). This approach sacrificed the taxonomically conserved 5.8S region from the amplicon, yet allowed the full amplification of SSU, LSU, ITS1, ITS2, and the IGS region(s). Primer 5.8SR was amended on the basis of the primer search experiment to include three degenerate bases for the downstream amplification and nanopore sequencing. There was a risk using this approach would amplify PCR products spanning multiple copies of the repeat region, however the PCR extension time was optimised for the amplification of products ~6kb to reduce this.

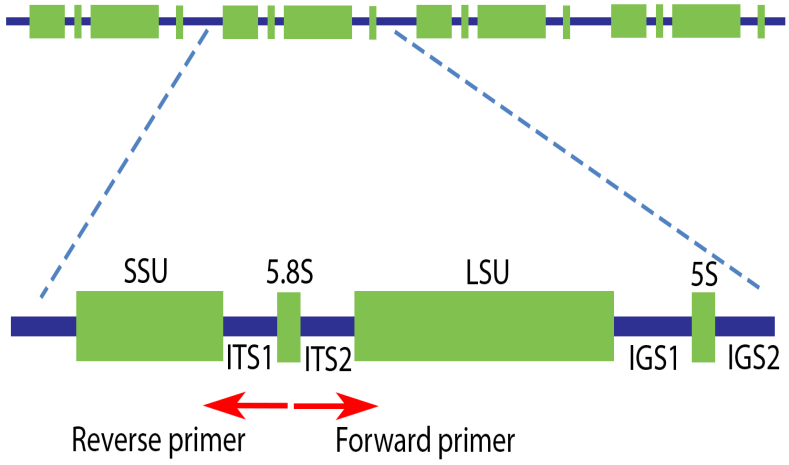


Figure 5.2: Long PCR primer design to amplify the full tandem repeat.

Basic sequence composition and identification

The MinION nanopore sequencing produced 19,235 2D sequences. There were large differences in the number of sequences produced from each dated sample for each site, with the lowest being 129 sequences and the highest being 2137 sequences (Table 5.2). The total number of sequences per site, excluding controls, was: Alford (5290), Haywold (5784), Stokesley (3810), Wansford (3944). While some of the samples had low numbers of sequence reads it was decided to proceed with the analysis.

Table 5.2: Total number of nanopore sequences produced per sample

Sampling date	**Alford**		**Haywold**		**Stokesley**		**Wansford**	
24/6/2015	CTA99	579	CTH104	2137	CTS102	270	CTW101	279
25/6/2015	CTA100	489	CTH105	847	CTS103	360	CTW102	359
26/6/2015	CTA101	1894	CTH106	288	CTS104	360	CTW103	930
27/6/2015	CTA102	537	CTH107	430	CTS105	237	CTW104	1137
28/6/2015	CTA103	146	CTH108	167	CTS106	215	CTW105	739
29/6/2015	CTA104	680	CTH109	1058	CTS107	1334	CTW106	129
30/6/2015	CTA105	965	CTH110	857	CTS108	1034	CTW107	371
Controls	CTAEB	194	CTHEB	199	CTSEB	2	CTWEB	2

Sequence length frequency histograms were produced to assess the length of sequences in each sample (Figure 5.3). The histograms showed that in many samples there were large numbers of sequences less than 1kb. There were 19 2D sequences >10kb which when investigated further with blast were shown to be extra long amplicons with two copies of the tandem repeat. However, the majority of the samples produced a cluster of high quality 2D sequences around the expected amplicon size of 5-6kb. The 1D sequences are considerably longer than the 2D reads and likely reflect the true amplicon size, however the accuracy of 1D sequences at the time of analysis was poor (Rang et al., 2018). All downstream analysis was carried out with the 2D sequences.

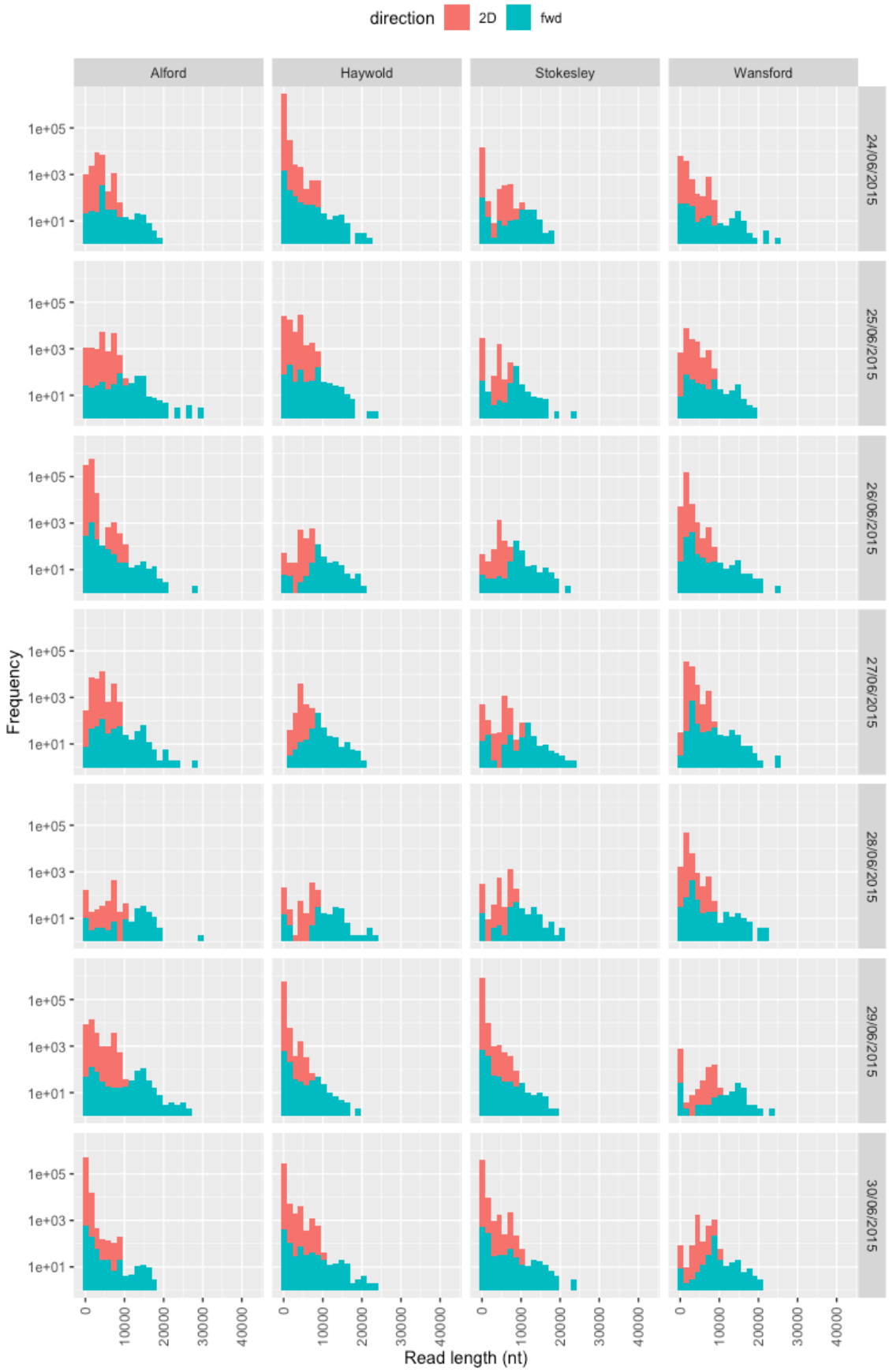


Figure 5.3: Sequence length histograms for each sampling location and date. The number of sequences is shown on a log scale to aid visualisation between samples. Individual read length differences between the 2D sequences (red) and 1D sequences (blue) are also shown.

The mean quality scores for all sequences in each sample was plotted to give a broad overview of the quality of the nanopore sequencing (Figure 5.4). The mean phred quality score across all 19,235 sequences was 19.8, indicating a 99% base call accuracy. However, this is very likely to be a gross overestimation of the true accuracy of the nanopore sequencing of these samples.

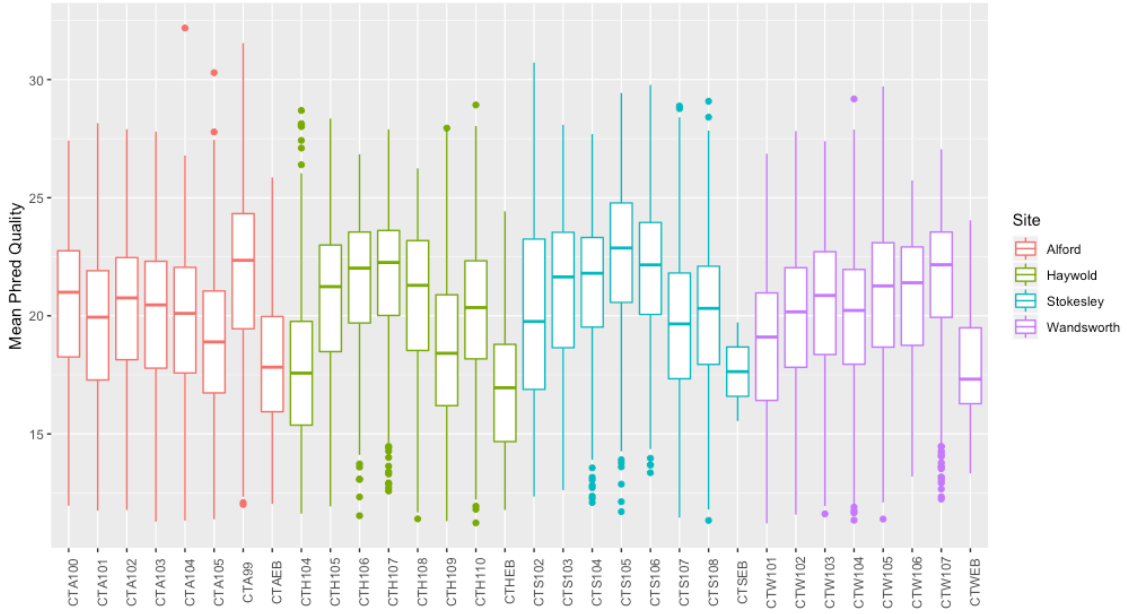


Figure 5.4: Boxplot of the mean quality scores for the nanopore sequences in each sample.

One strategy for determining the accuracy of nanopore sequencing which was not employed during these experiments, but which would improve the quality estimations, would be to use a spike-in of known DNA. The nanopore reads from this DNA would then be mapped back to a Sanger or high-coverage Illumina reference sequence and accuracy estimated based upon the number of SNPs in the reference assembly. The inclusion of PhiX DNA in Illumina sequencing is an example of this as it enables an assessment of the overall performance of that run. Similar approaches have been applied to microbiome and metabarcoding studies where a mock community containing a mixture of species with known DNA concentration is also sequenced (Nguyen et al., 2015). This provides a two-fold assessment of the sequencing as both sequencing accuracy and the sensitivity of the sequencing run to species detection can also be assessed.

Taxonomic identification of the nanopore sequence reads

The nanopore sequences were searched against the same UNITE/UK risk register sequence database used previously to analyse the Illumina sequencing of the same samples. Only 5421 (28%) nanopore reads could be assigned to a taxon with 13,814 sequences left unidentified. The sequences with a putative identification had also been subject to a lowest common ancestor (LCA) analysis of the blast results using MEGAN. To assess whether the sequence quality may have been a factor in the identifications, the mean quality scores for each sequence was plotted with the putative identification made by the LCA algorithm (Figure 5.5). This demonstrated that there was no obvious difference in the mean quality score of the sequences that could not be identified. In fact, many of the highest quality sequences in the dataset could not be identified with blastn (Figure 5.5, pink boxplot).

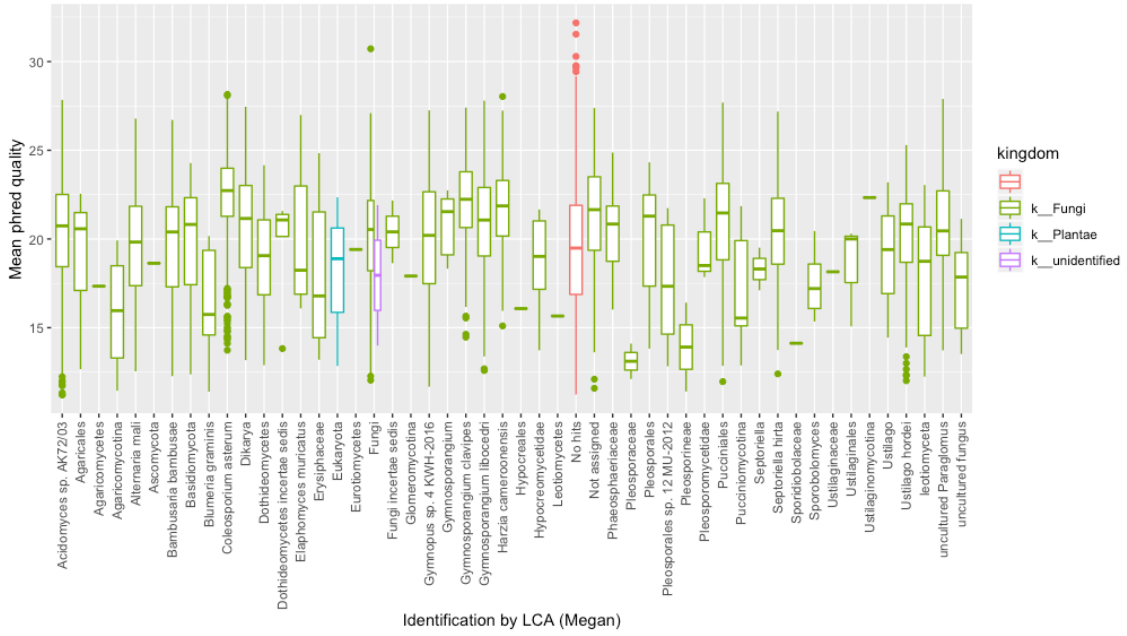


Figure 5.5: Mean sequence quality scores for sequences with identifications made by MEGAN using a lowest common ancestor.

The mean sequence quality, sequence length and the percent identify found between the sequence and the hit in the UNITE/risk register sequence database were plotted to elucidate any correlations (Figure 5.6). This demonstrated that many of the high quality identifications had adequate sequence quality, but were predominantly less than 1kb in length. Most of the longer sequence identifications were made between 80-90% identity to the reference sequence, which is also the

reported accuracy for nanopore sequencing (Rang et al., 2018) and suggests the potential of genus- or family-level identifications but not species-level identifications.

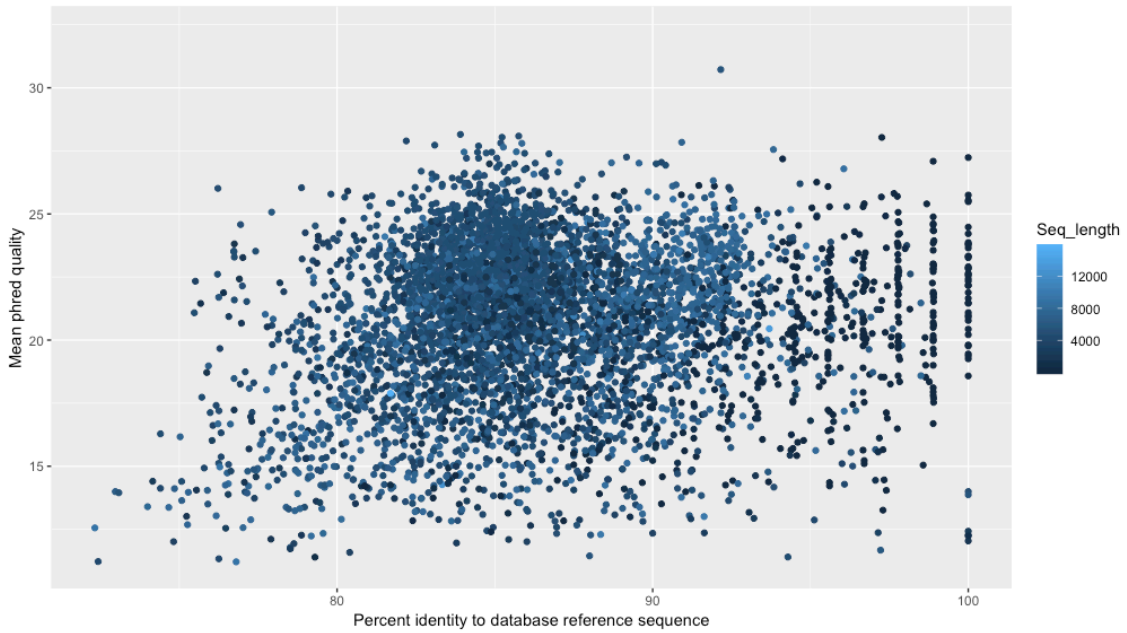


Figure 5.6: Scatter plot of sequence quality and the percent identity to the database for sequences with blastn hits to the UNITE/risk register sequence database. The points are coloured by sequence length.

Investigating the link between sequence length and putative identification

It has been reported that the intergenic spacer region can vary in size between different fungal species and groups, and it is well known that ITS1 and ITS2 can differ in size between species (Vilgalys et al., 1992). In the earlier years of sequence analysis, where RFLP was a popular way to investigate the differences between species and strains, the fungal IGS region was a target of first choice given its hypervariability (Jackson et al., 1999; Konstantinova and Yli-Mattila, 2004; Ranjard et al., 2001). The IGS region can vary from 2kb upwards and in filamentous ascomycetes it can reach up to 5kb in length. In yeasts and basidiomycetes the IGS region also contains a single coding region for the 5S RNA. In light of this, the lengths of the nanopore sequences with putative identification were plotted to investigate any correlation between putative identification (by LCA) and sequence length (Figure 5.7)

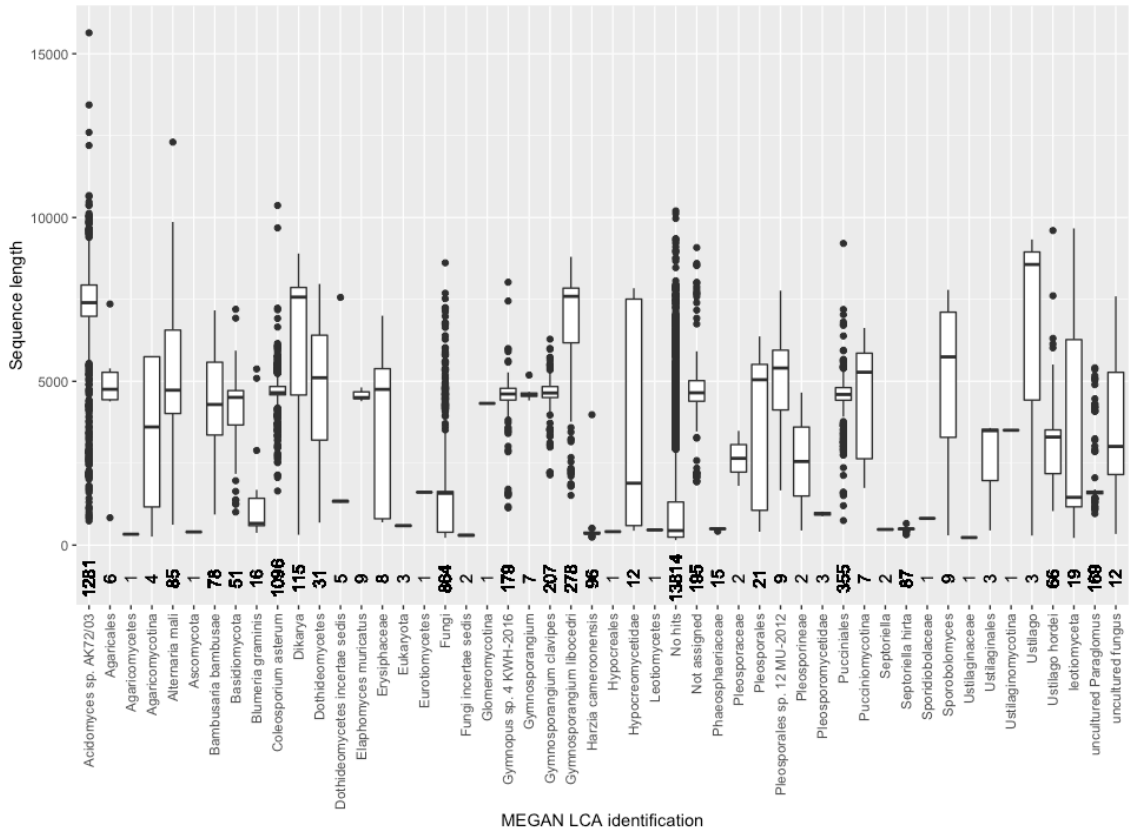


Figure 5.7: Sequence length boxplots for each of the taxa identified with a lowest common ancestor algorithm in MEGAN. The numbers next to the taxa describe the number of nanopore sequences assigned to that taxon.

The sequences putatively identified as belonging to three taxa were considerably longer than other sequences: *Acidomyces* sp. AK72/03, *Dikarya* spp., and *Gymnosporangium libocedri*. The reference sequence for *Acidomyces* sp. AK72/03 was published in 2010 (FJ430711) and is 2.8kb in length, which will have enabled more sequences to find similarity to this reference. The sequences identified as *Dikarya* by MEGAN's LCA algorithm were predominantly identified as *Acidomyces acidothermus* using blast alone and account for all the sequences longer than >7kb in the *Dikarya* group. The sequences putatively identified as *Gymnosporangium libocedri* had only 75-84% identity to the reference sequence and so are only related to this plant pathogenic rust fungus. Figure 5.7 also shows taxa which have mean sequence lengths ~5kb with a wide spread of sequence lengths longer and shorter, although these taxa have lower numbers of sequences.

The results demonstrate that nanopore sequencing is able to sequence the entire tandem repeat. While the IGS region was popular in the 1990s and early 2000's

as a species identification marker in combination with RFLP, there are very few sequenced IGS regions with which to compare nanopore sequences. The error rate of nanopore sequencing adds a little uncertainty to the identifications made, however it may well be the hypervariability of the IGS region itself. A recent paper demonstrated the sequencing of the entire tandem repeat using three long PCR products from cultured material for the production of IGS and full-length ribosomal DNA barcodes (Wurzbacher et al., 2018). Should efforts be made by the fungal barcoding community to extend the length of barcodes in databases such as UNITE, then the combination of full-length tandem repeat barcoding could yield higher resolution identifications and therefore higher resolution community studies.

Investigation of unidentified nanopore sequences: fungal species or amplification from other taxa?

Given the large number of unidentified sequences (n=13,814; 72%) an exploratory analysis of the GC content of the sequences was undertaken to identify any groups of sequences which may originate from related organisms, whether fungal or another taxonomic group. Differences in GC content between genomic DNA and ribosomal DNA have been described for both prokaryotes and eukaryotes, with higher %GC being observed in ribosomal DNA (Schattner, 2002). In order to aid the identification of the sequences to higher taxonomy the GC content of each sequence was plotted against its length for all sequences in the dataset (Figure 5.11A). Further plots coloured by putative identification (Figure 5.11B), mean sequence quality (Figure 5.11C) and site (Figure 5.11D) demonstrated that the identified sequences were centrally grouped, sequence quality did not influence the groups of sequences but that site-specific groups existed.

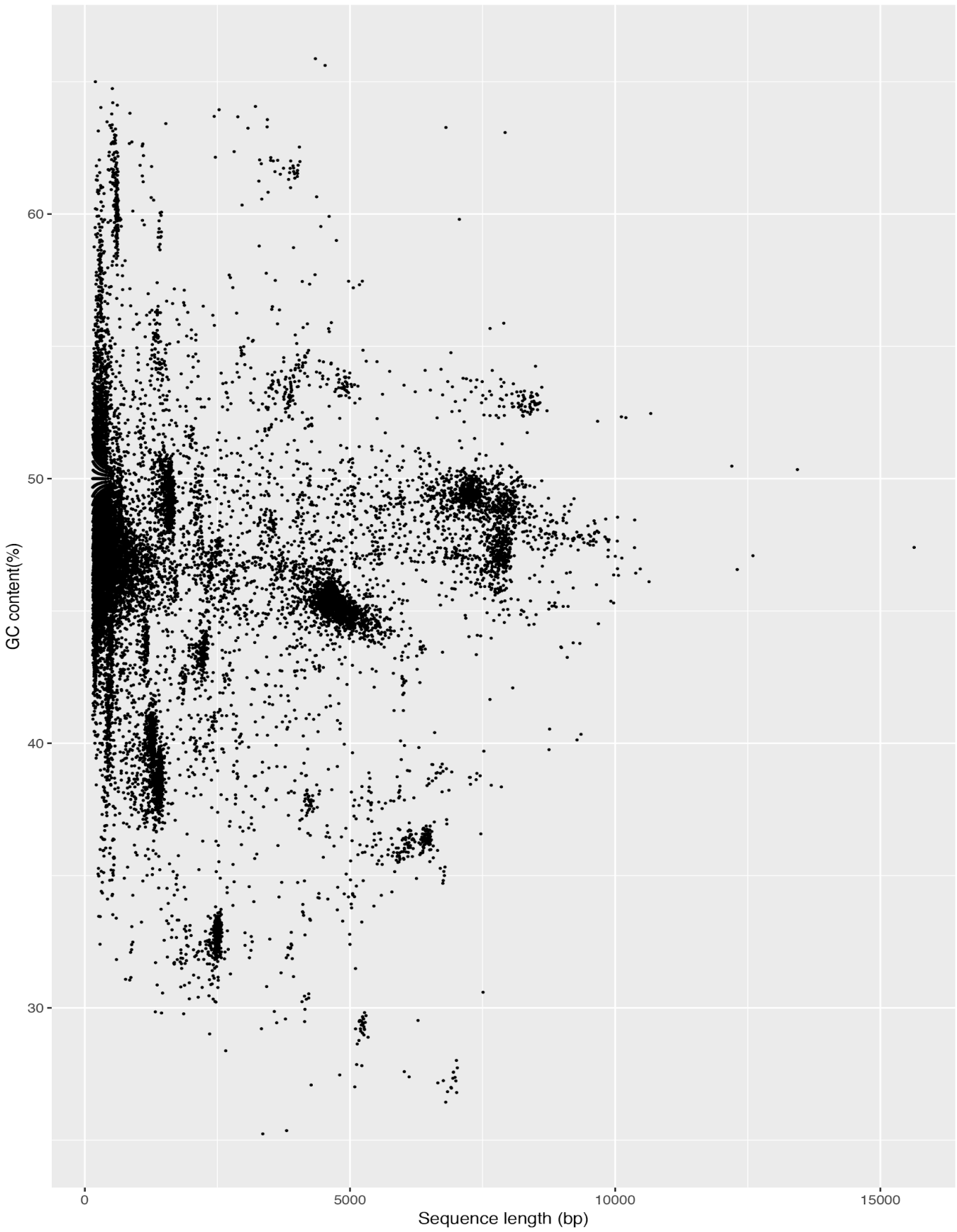


Figure 5.8: Scatter plots of the GC content against sequence length of each 2D nanopore read. (A) Points uncoloured;

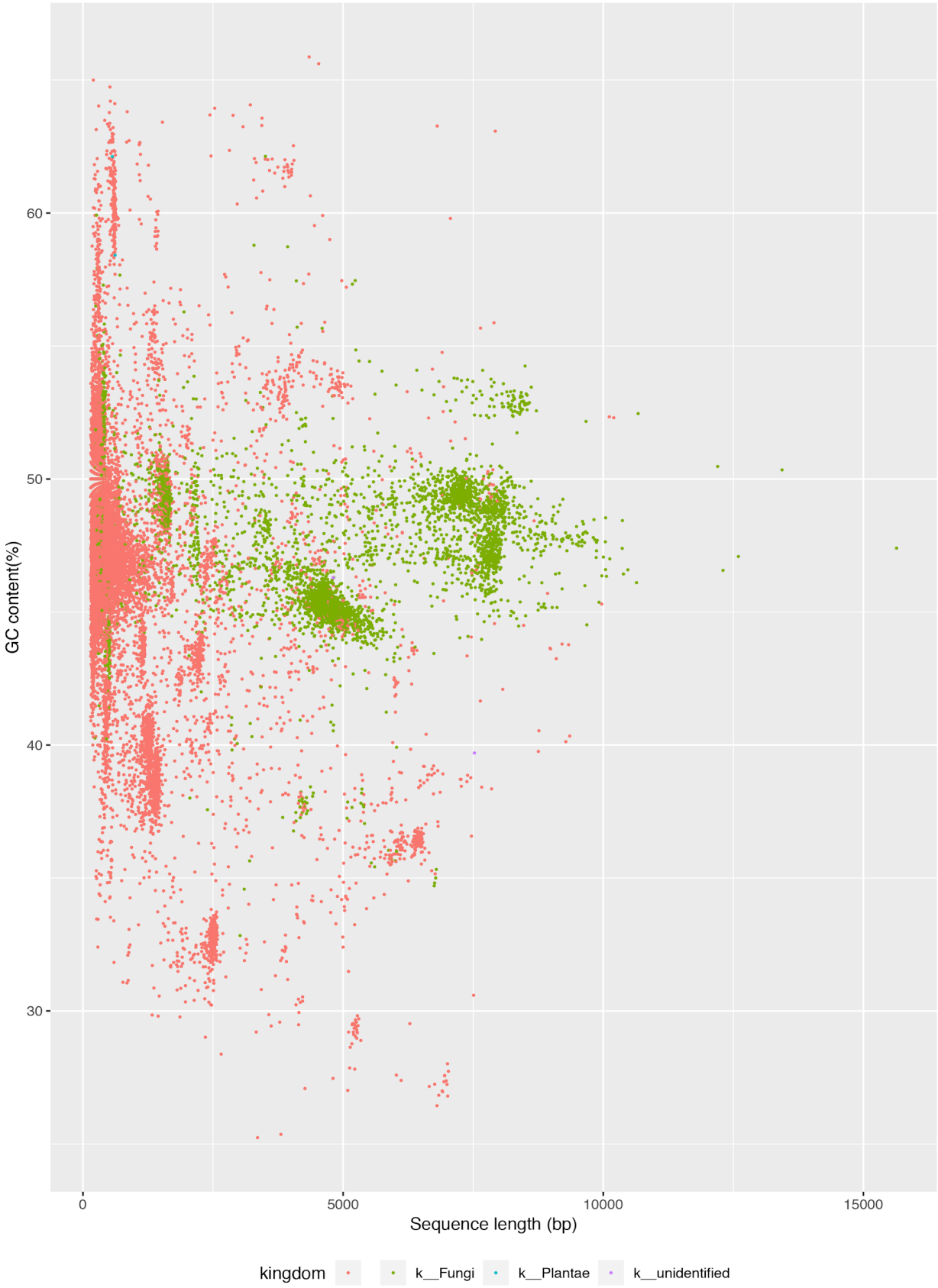


Figure 5.9: Scatter plots of the GC content against sequence length of each 2D nanopore read. B) Points coloured by taxonomic kingdom: fungi (green), unidentified (pink), plant (blue);

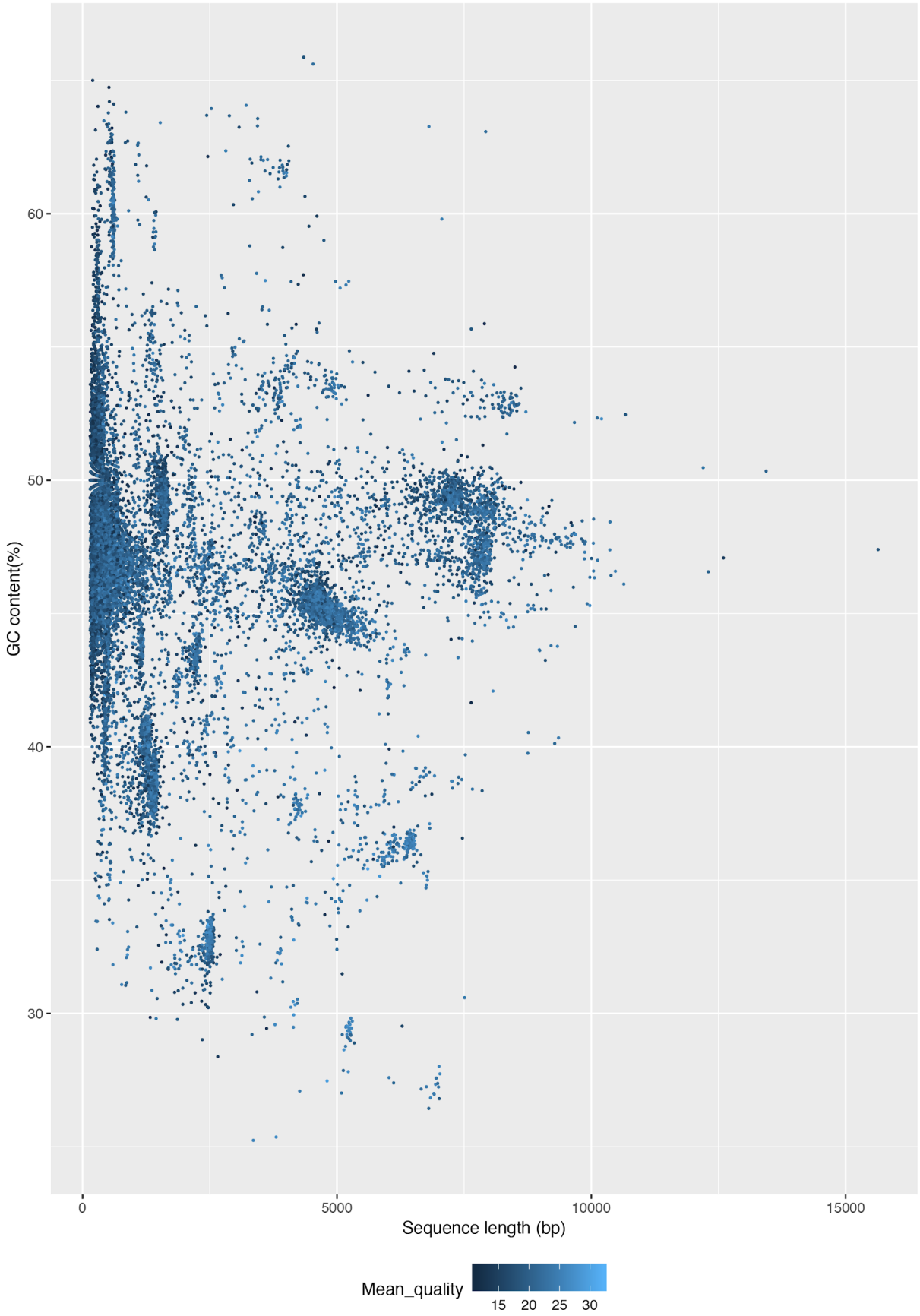


Figure 5.10: Scatter plots of the GC content against sequence length of each 2D nanopore read. (C) Points coloured by mean sequence quality score: low (black), high (blue

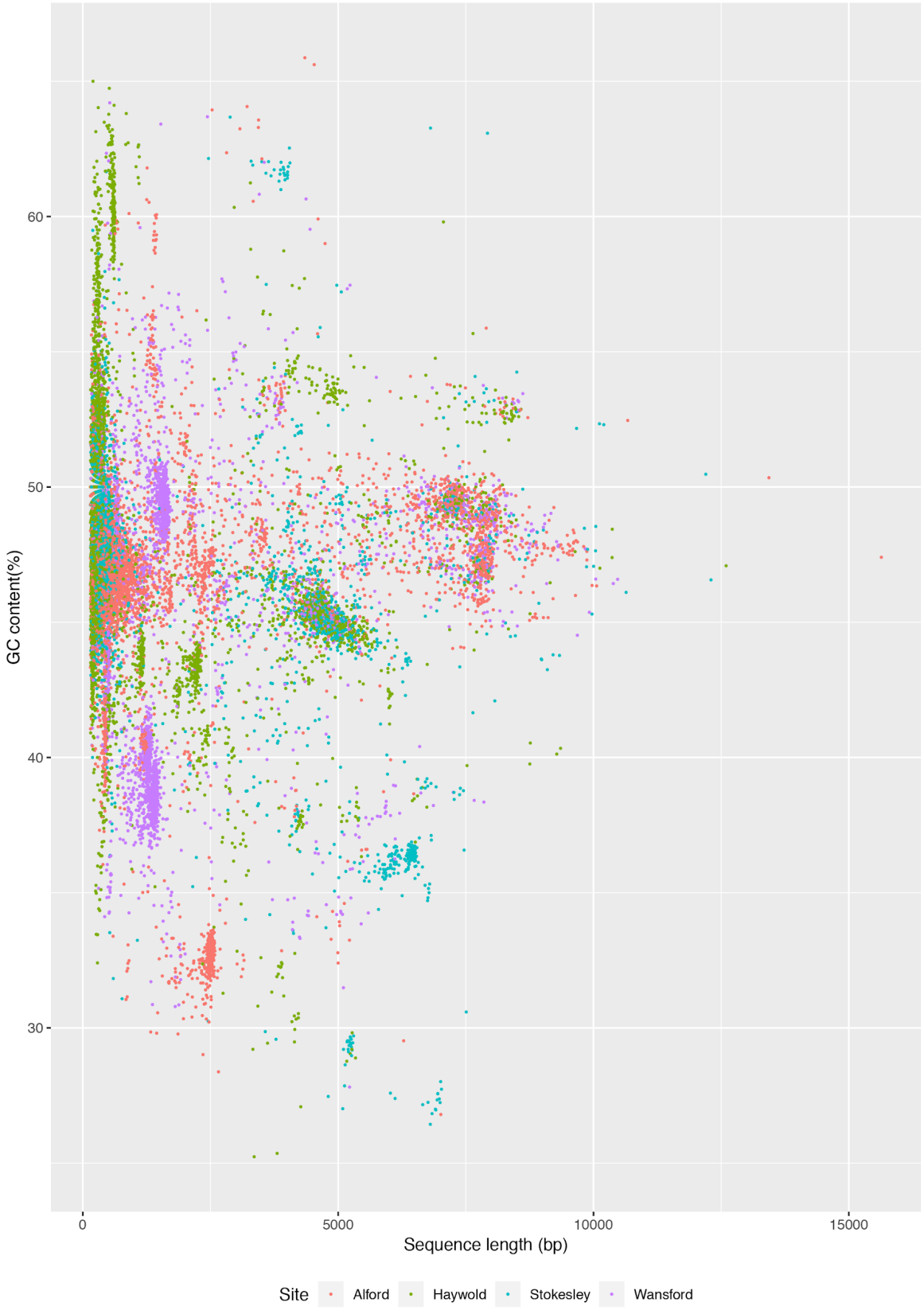


Figure 5.11: Scatter plots of the GC content against sequence length of each 2D nanopore read. (D) Points coloured by location of the originating sample: Alford (pink), Haywold (green), Stokesley (blue), Wansford (purple).

To further investigate the relationships between the GC content and sequence length, the plots were split by both date and sampling location (Figure 5.12). This was carried out with the hypothesis that closely related sequences may group together due to similar ribosomal tandem repeat length and %GC content. Four interesting groups of sequences were selected for further analysis due to their extreme %GC content (either high or low), or discrete correlation with date and location.

Group 1 comprised sequences from the Alford sample from 24th June 2015 with low GC content and potentially also present in the Wansford sample on 25th June 2015. Group 2 comprised sequences in the Haywold sample from 25th June 2015 with a high GC content and potentially only present in this location on this day. Group 3 comprised sequences present in the Stokesley sample from 27th June 2015 with very high GC content. Group 4 was a group of sequences in the Stokesley sample from 28th June 2015 with very long read length and very low GC content.

Group 1

The length of these sequences ranged from 2050-2816bp, the mean %GC content was 32.6% (min=30.2%, max=34.9%) and mean sequence quality score was 22.6 (min=13.2, max=31.5). No identification could be made for any of the sequences against the fungal ITS UNITE/risk register database with blastn which would be expected to find sequence identity greater than approximately 75%. Alignment of sequences within this group with clustalw did not identify any sub-groups of sequences and blastn against the NCBI nucleotide database did not return results useful for identification. A further search with blastx against the NCBI nr protein database did not return any results and as such this group of sequences remains unidentified. A low GC content may be suggestive of bacterial origins but the indels prevalent in nanopore sequencing made open reading frame (ORF) finding extremely difficult and may have ultimately affected the ability of blastx to find significant homologies within the nr database.

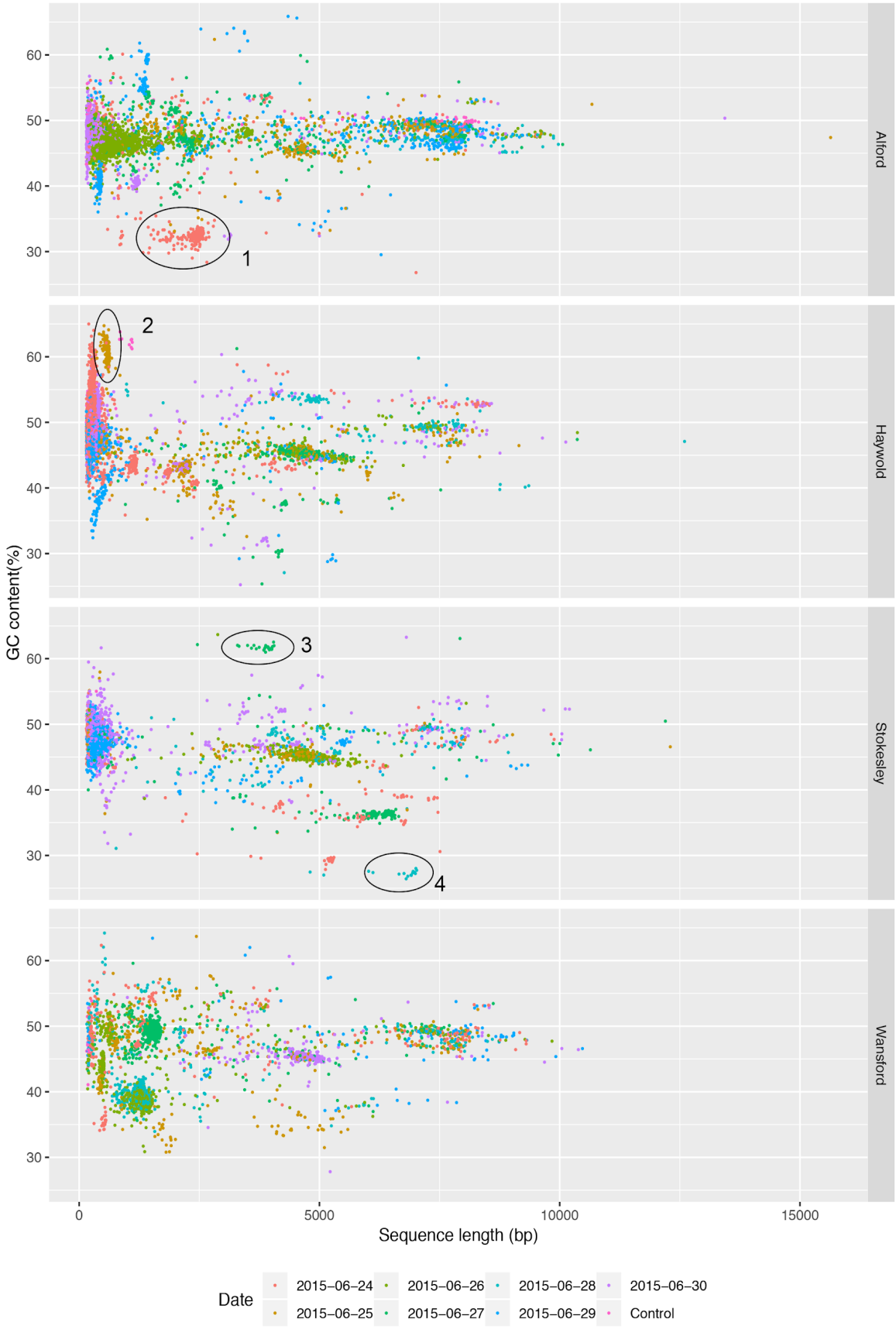


Figure 5.12: GC content plotted against sequence length for all nanopore sequences, separated by site and coloured by date. Four interesting groups are annotated: 1. Alford, 24th June; 2. Haywold, 25th June; 3. Predominantly Stokesley, 27th June; 4. Stokesley, 28th June.

Group 2

Sequences ranged from 177-1106bp with the vast majority originating from Haywold samples CTH104 and CTH105. Searches against the fungal UNITE/risk register sequence database returned a small number of very short high identity hits to fungi in the family *Ceratostomataceae*. The alignment to the reference sequences in these cases was at the extreme 5' end of the sequences, suggesting the hits were in the highly conserved primer region within 5.8S. Alignment with ClustalW showed that there were four different sub-groups of sequence within this group. A representative from each sub-group was searched against the NCBI nucleotide database with blastn to further characterise these sequences. The sequences from two of the sub-groups showed significant similarity to *Triticum aestivum* (bread wheat) and the other two sub-groups to bacterial sequences, suggesting the back-to-back conserved 5.8S primers are capable of amplification in other eukaryotes and in non-ribosomal regions in prokaryotes.

Group 3

The sequences in this group predominantly originated from Stokesley on 27th June but a small number were also observed in the sample from Alford on the 29th June. The GC content ranged from 60.3-65.8 and sequence lengths ranged from 2532-4531bp. Only one sequence out of 41 could be putatively identified with the UNITE/risk register sequence databases. The single sequence had 75% identity to 343bp of the Basidiomycete *Malassezia restricta*. When aligned with clustalw, three sub-groups became apparent and a representative sequence from each was searched with blastn against the NCBI nt database. Two of the sub-groups could be identified as *Leifsonia* spp., with 81% identity to *Leifsonia xyli*, a bacterium that causes sugarcane stunting disease. *Leifsonia* is a bacterial genus, adding further evidence that the 5.8S primers are capable of amplification outside of fungal taxa. The bacterial amplicon covered a number of genes, including the bacterial LSU, but the priming sites appeared to be in the *DnaA* and *SpoIIIJ-associated protein* genes. The third sub-group was also bacterial in origin, with the closest hit being *Variovorax* spp., with the primers having amplified 3.5kb of the GGDEF-containing protein. The bacterial identifications were significantly

better in both identity and alignment length than the original putative identification determined from the UNITE/risk register database (*Malassezia* spp.). The sequences within this group appear to be bacterial in origin and are predominantly found in the Stokesley sample.

Group 4

The sequences in this group were only found in the Stokesley sample from 28th June. The GC content ranged from 26.4 to 28.0 and the length from 6024bp to 7024bp. No identification could be made when the sequences were searched against the UNITE/risk register database with blastn. Alignment of the 15 sequences in this group with clustalw showed two distinct groups which were very different to each other.

The four small groups characterised above demonstrate that in some cases more accurate identifications can be made by expanding the searches to larger databases. A better search strategy for identifying amplicons encompassing the entire tandem repeat would likely be blastx against a eukaryotic rDNA database followed by subsequent searches in whole genome sequence databases for amplicons where an adequate identification could not be made. However, this would be very computationally intensive and outside the current capabilities of most applied organisations where these techniques would have the most impact. The current nanopore pipelines incorporating species identification for use in the field are mainly k-mer based and therefore unlikely to find distant homology to place a sequence to the lowest taxonomic rank possible (Oxford Nanopore, *What's In My Pot [WIMP])*. Current field-based studies rely on offline searching of sequence data using local databases installed on a laptop (Edwards et al., 2016), which restricts the identifications that could be made to the computational power and disk capacity of the machine.

Discovery of dark taxa with long-read metabarcoding

The bacterial sequences identified in the Stokesley samples (group 3) by a combination of GC content, sequence length and blastn against the full GenBank nucleotide database were not introduced as contaminants during the extraction or library preparation process through reagents or consumables as has been described previously (de Goffau et al., 2018; Salter et al., 2014). Rather, the

sequences originated from mis-priming of the conserved 5.8S primers to genomic DNA from bacterial species present in the Burkard trap samples, as the bacterial species are only present in specific samples. If these were reagent contamination they would likely be present in all samples or in particular batches of samples processed at the same time. The combination of degenerate primers and long PCR thermocycling resulted in spurious amplification of a non-target taxa. However, for the purposes of surveillance for new fungal threats, the amplification of non-fungal sequences is preferable to the opposite scenario where undiscovered taxa could be missed by primers which are too conserved (Hugerth et al., 2014).

The plotting of GC content against sequence length may be a useful tool to elucidate groups of novel sequences where local alignment searches against known fungal species have failed to produce an identification. The majority of sequences which could be identified as fungal in origin in the nanopore dataset had a GC content of ~50% (+/- 5%). Yet there were many sequences in the same region which could not be identified. Sequences suspected to have a fungal origin but which have no further identification can be described as “dark taxa”. That is, the sequences have no lower taxonomic placement and have not been described or cultured previously (Ryberg and Nilsson, 2018). In many community ecology studies such sequences are, at worst, ignored as they have no taxonomic name and at best are grouped together as “unidentified”, despite the efforts of UNITE to include such sequences under the concept of “species hypotheses” (Kõljalg et al., 2005; Nilsson et al., 2013). These taxa represent important parts of the fungal communities being studied and their abundances may be important in elucidating differences between samples. For the purposes of plant health surveillance, there is a small possibility that these dark taxa could contain emerging threats which have yet to be described.

Comparison between Illumina and nanopore amplicon sequencing of the same sample

Chapter 4 demonstrated that one or more contamination events had occurred during either DNA extraction, PCR amplification or Illumina library preparation of the Burkard samples subsequently used for the MinION sequencing presented in this chapter. Surprisingly, given they were compared to the same reference

database, the Illumina and MinION reads had very different proportions of the samples which could be identified when searched against the UNITE/risk register sequence database (Table 5.3). The percentage of sequences identified in the Illumina samples rarely fell below 99%, whereas the percentage of sequences identified when the same sample was sequenced with long-read nanopore sequencing varied greatly from a low of 1.7% (sample CTH104) to a high of 94.4% (sample CTS104). The range in percent identified to taxa in the MinION sequences can be partly attributed to the far lower number of sequences in the MinION sequenced samples. The samples had been sequenced on a MinION flowcell which had previously been used and washed and which still had active pores but which may have been underperforming in the number of pores available for sequencing, which explains the lower number of sequences obtained than expected. The very high proportions of Illumina sequences able to be placed to taxa seems incredibly optimistic given these are field samples where we would expect a proportion of the sample to be “dark taxa”, those which have not yet been sequenced or described. These results suggest that the truly unknown Illumina sequences are being assigned to a taxon, even if incorrect, whereas the longer MinION sequences are not being assigned taxonomy. This is an interesting observation given the blastn databases and search parameters were identical. We would expect that the species-specific ITS loci would be sequenced in the MinION reads, perhaps with a higher rate of sequence errors and single nucleotide indels, but still with the ability to be assigned to a taxon with blastn (i.e. with sequence identity >70% to a reference in the database). One possible explanation for the lower number of taxa assignments for the nanopore sequences is that the UNITE database is almost entirely populated with shorter ITS1-5.8S-ITS2 sequences. As the nanopore amplicon begins and ends in the 5.8S rDNA - effectively a circular amplicon - the ITS2 region would be at the 5’ end of the nanopore sequence and the ITS1 region would be at the 3’ end of the nanopore sequence. While there is no reported loss of quality towards the end of reads in nanopore sequences, the splitting of the ITS1-5.8S-ITS2 sequence into two distant regions of the sequence may have affected the ability of the blastn algorithm to find the best high-scoring pair for that sequence. As the reference sequence database contained complete sequences, our MinION sequences for identification were effectively split (5’ and 3’ end of the MinION amplicon) and this would have negatively affected the ability of the blastn algorithm to find

identifications for the MinION sequences. An improvement to the search strategy would have been to take the 5' and 3' ends of the MinION amplicons and re-join them to improve the taxonomic identifications made, but time constraints prevented this from being tested further. The primer locations for the amplification of the MinION long amplicon were chosen because 5.8S is highly conserved in Fungi (Nilsson et al., 2008) and so potentially allowed the amplification of many more species than primers placed in less conserved areas. However, the placement of the primers in 5.8S clearly had an impact on the bioinformatics algorithms used to identify the resulting amplicons.

The effect of the difference in sequences identified was obvious when the identified sequences from both the Illumina and nanopore methods were compared with standard molecular ecology analyses. The number of species present over all samples showed a striking difference between the two technologies (Figure 5.13) which was further backed up by splitting the technologies by location (Figure 5.14). The difference can be partly explained by the large difference in the number of sequences used to calculate the relative abundances used for the Bray-Curtis dissimilarities, indicating a difference in resolution. NMDS of the Bray-Curtis dissimilarities also confirmed that the significant difference between the species present (Figure 5.15).

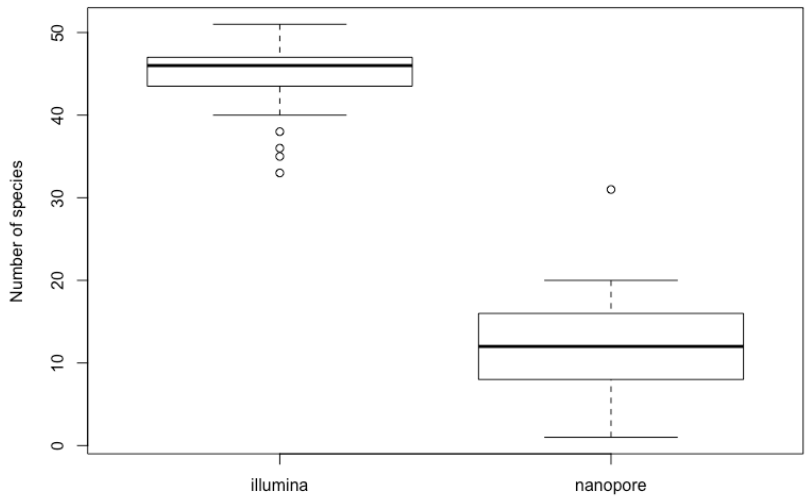


Figure 5.13: Number of species across all samples sequenced with both Illumina and nanopore technologies

Table 5.3: Percentage of sequences which could be assigned to a taxon from samples sequenced with both nanopore MinION and Illumina metabarcoding sequencing.

Sample	MinION (% identified)	Illumina (% identified)
CTA99	10.7%	99.5%
CTA100	77.5%	99.5%
CTA101	31.0%	99.8%
CTA102	52.5%	99.4%
CTA103	85.0%	99.9%
CTA104	53.4%	99.6%
CTA105	4.0%	98.3%
CTH104	1.7%	98.8%
CTH105	33.9%	99.1%
CTH106	34.5%	99.6%
CTH107	89.7%	99.5%
CTH108	34.1%	99.5%
CTH109	5.7%	99.5%
CTH110	55.3%	99.2%
CTS102	18.8%	99.2%
CTS103	78.8%	99.1%
CTS104	94.4%	99.3%
CTS105	10.9%	99.2%
CTS106	15.8%	98.8%
CTS107	2.7%	99.4%
CTS108	8.4%	99.2%
CTW101	34.4%	99.8%
CTW102	41.5%	99.9%
CTW103	20.1%	98.7%
CTW104	59.1%	99.4%
CTW105	9.6%	99.3%
CTW106	58.9%	99.5%
CTW107	89.7%	99.9%

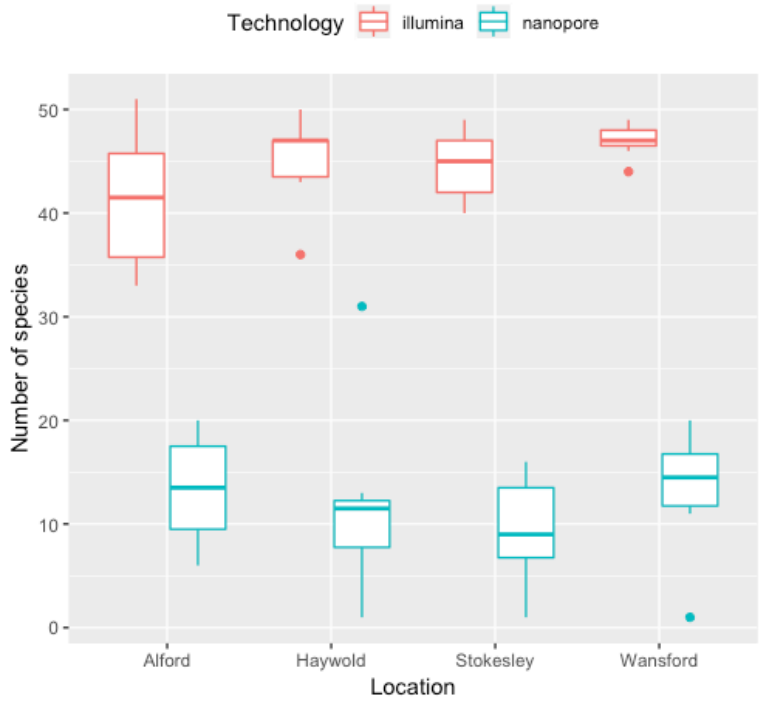


Figure 5.14: Number of species across all samples sequenced with both Illumina and nanopore technologies when split by location

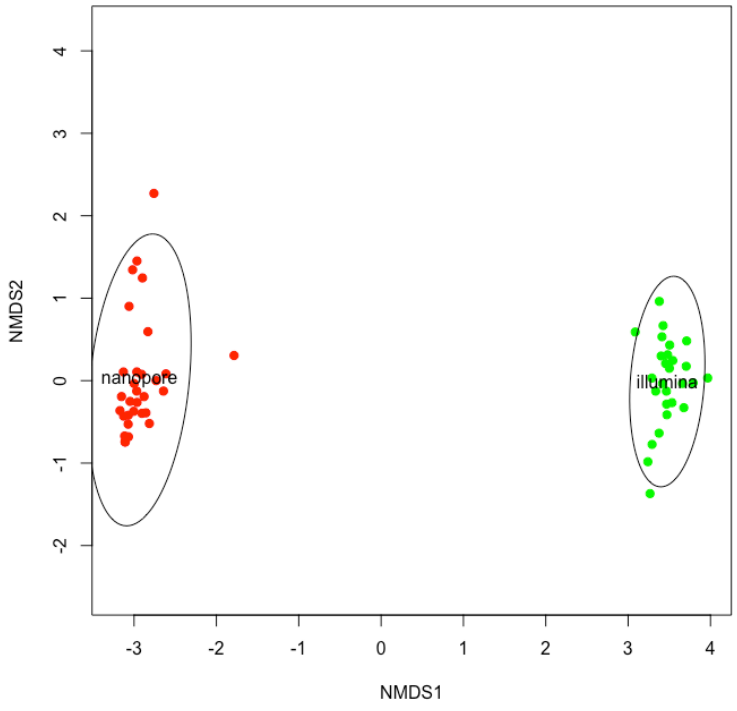


Figure 5.15: NMDS showing the relationships between the samples separated by sequencing technology: nanopore (red) and Illumina (green)

Hierarchical clustering of the Bray-Curtis dissimilarities further demonstrated a distinct divide between the samples when sequenced with the two technologies (Figure 5.16). Manual inspection of the relative abundances (abundance >1%) showed that only 7 species could be detected by both methods: *Holtermanniella takashimae*, *Vishnacozyma victoriae*, *Sporobolomyces roseus*, *Ciborinia camelliae*, *Ustilago hordei*, *Blumeria graminis* (mostly Wansford samples when Illumina sequenced and throughout the nanopore sequenced samples), *Alternaria mali* (predominantly in the Illumina sequencing and a small number nanopore sequenced samples).

One prior hypothesis was that carrying out nanopore and Illumina sequencing of amplicons from the same DNA extract would produce similar communities, albeit with differences in resolution due to amplicon length. However, this does not seem to be upheld by the data. Only 7 species of 115 present in any sample in more than 1% abundance are shared by the two amplicons and sequencing methods. The differences could well be due to primer bias but this seems unlikely in the long amplicon nanopore sequencing due to the evidence of non-fungal eukaryotic amplification of the ribosomal tandem repeat. The accuracy of the nanopore sequencing may have made identifications more difficult for the blastn algorithm, yet nucleotide identities as low as 70% between nanopore sequence and reference database sequence were observed, suggesting that if an adequate reference sequence was present a distant identification could have been made.

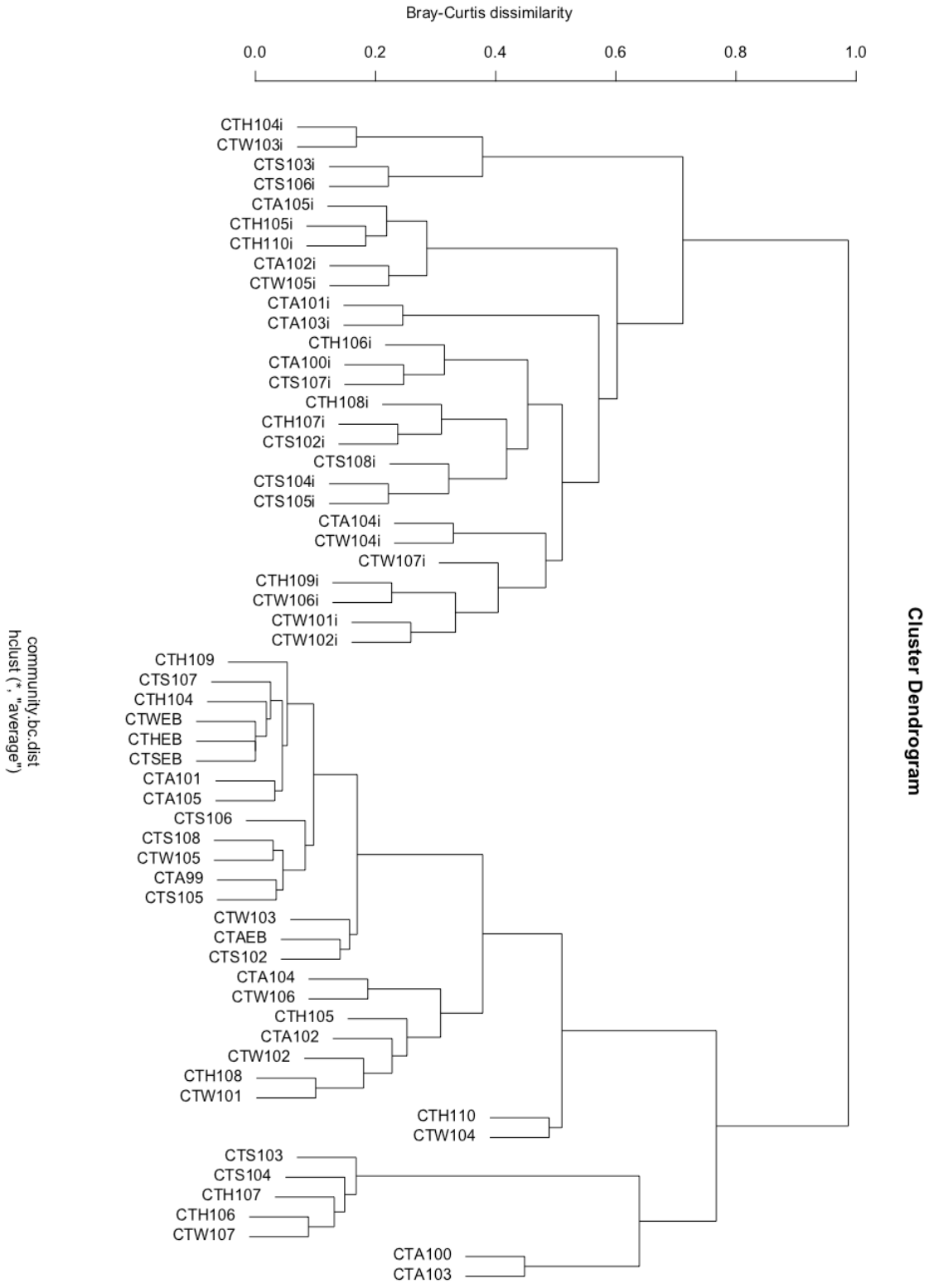


Figure 5.16: Hierarchical clustering of Bray-Curtis dissimilarities for samples sequenced with Illumina (suffixed with 'i', left hand group) and nanopore (no suffix, right hand group) sequencing technologies

Chapter 6. Discussion

This project was the first to demonstrate that DNA metabarcoding can produce comparable results to those obtained from traditional light microscopy for diatom assemblage analysis, with the advantages of increased resolution and the ability to elucidate cryptic taxa. Prior to this work there had been limited demonstrations of the use of metabarcoding for the characterisation of diatom assemblages (Kermarrec et al., 2014, Visco et al., 2015, Zimmerman et al., 2014). Of these Visco et al., (2015) was the only work to achieve quantification, using 18S rDNA on a small dataset and with lower agreement with light microscopy than was achieved in this project. As a direct result of this work, following a transition period where light microscopy and Illumina metabarcoding were used in parallel, the Illumina metabarcoding method is the sole method used by the UK Environment Agency to calculate the Trophic Diatom Index and thus determine the Ecological Quality Ratio as required by the EU Water Framework Directive.

The results from the water quality testing part of this project have shown that there are certain species of diatom, for example *Melosira varians*, which consistently had higher relative abundances in the metabarcoding data compared to light microscopy. These differences were hypothesised to be due to *rbcL* copy number, which can vary between species of diatoms due to the number of chloroplasts present (Mann et al., 1996). Multi-copy loci such as *rbcL* (and other chloroplast, mitochondrial and tandem repeat loci) are advantageous as species identification markers due to their ease of PCR amplification (Nilsson et al., 2008; Hebert et al., 2003b). This has led to markers such as *rbcL*, *Cytochrome oxidase I*, and *ITS2* being utilised extensively for single specimen identification using Sanger sequencing (Hebert et al., 2003a; Hebert et al., 2004; Hollingsworth et al., 2009; Koljalg et al., 2005). When a more quantitative result is required, such as the metabarcoding methods described here, a single-copy nuclear gene is typically more appropriate, enabling more accurate calculation of relative abundances without knowing the copy number in each species. However, single-copy nuclear regions are not a panacea as the degenerate PCR primers required to amplify multiple diverse species present in a sample can also co-amplify pseudogenes, which can potentially be misidentified as a species not actually

present in the sample (Song et al., 2008) and the exon/intron structure of nuclear genes can make degenerate primer design surrounding informative regions challenging.

The use of *rbcL* for the diatom metabarcoding experiments presented here had been decided by an earlier project funded by the EA and completed by Cardiff University to deliver a metabarcoding method using 454 pyrosequencing (Kelly et al., 2018). However, 454 pyrosequencing was discontinued in 2013 and the original method required updating to a newer sequencing technology with a more appropriate operational design that would work for the processing of thousands of samples per year. This was an iterative process over a number of sequential projects over multiple years and so the ability to change from *rbcL*, a multi-copy gene, to a more quantitative single-copy nuclear gene was not possible within the budgetary restraints of the EA. With hindsight, it may have been more appropriate for the EA to invest in identifying a new marker region in the earliest projects given the limitations of *rbcL* due to copy number as determined in this work.

A comparison of barcoding genes for diatoms published after the development of our *rbcL* method (Guo et al., 2015) demonstrated that *18S* rRNA, *COI* and *ITS* may perform better than *rbcL* in the identification of certain diatom taxonomic groups. However, these loci are all multi-copy and would have required considerable investment in the re-creation of the DNA barcode reference database for their use. Other European countries developing similar biotic indices for water quality with diatoms have used different markers but none are yet in operational use by their respective statutory bodies at the time of writing. Despite other barcoding genes being available, the choice of *rbcL* as a marker did not prove to be too detrimental as there was a predictable relationship between the number of individual diatoms and the number of reads. An approach was developed for the selection of a short-read sequencing region of *rbcL* from longer DNA barcode sequences, such that the maximum variability was retained for species discrimination, yet degenerate primers could be designed to highly conserved regions. The approach developed during this project could be applied to any similar scenario where longer barcodes exist but where sequencing technology restricts the size of the amplicon.

The relationship between the diatom LM and NGS results is complicated for a number of reasons. Firstly, the relative abundance of a species appears to be influenced by the number of chloroplasts but there have been no studies to date fully investigating the number of chloroplasts per species. It is likely that the number of chloroplasts varies between species and between environmental conditions in the same species (Rauwolf et al., 2010). Secondly, LM does not record the number of cells, rather the number of frustules (the number of valves, or half cell walls). In very small diatoms it can be difficult to determine whether a single valve or complete frustules are present. Finally, the relationship between the LM and NGS for a particular diatom species will be determined from within a mixture of typically more than 20 species. Therefore, the proportion of one species will be influenced by changes in the proportion of other species in the sample when comparing species between samples. Species with multiple chloroplasts, for example *Cyclotella meneghiniana* and *Diatoma vulgare*, had the potential for over-estimation of their abundance in the metabarcoding while being present at much lower abundance in LM. Other diatom species, for example *Fistulifera saprophila* were thought to be at lower abundance in the LM due to being weakly silicified and therefore susceptible to damage during the chemically aggressive process during slide preparation (Zgrundo et al., 2013).

Following the work in this project the version of the Trophic Diatom Index originally used with light microscopy data (TDI4) was updated to a version which could be used with metabarcoding data (TDI5) which included species-specific weightings which accounted for the differences seen between light microscopy and metabarcoding (Kelly et al., 2018).

The reproducibility and repeatability between individuals and with different MiSeq machines demonstrated a very consistent method that was robust to change in personnel and physical sequencing location. The ability of a routine method to withstand these changes is important as batch effects can introduce significant biases into analyses, as demonstrated in the fungal surveillance work in Chapter 4 and by Balint et al. (2018).

The acceptance of the short-read diatom metabarcoding method for routine high-throughput use by the UK Environment Agency demonstrates its utility and

robustness. However, the genomics technology landscape is constantly changing and evolving and, just as the initial 454 technology used prior to this project became obsolete, so will Illumina short-read sequencing (Slatko et al., 2018). Long-read sequencing technologies such as PacBio Sequel and Oxford Nanopore Technologies nanopore sequencing demonstrate that there is a demand for sequence length as well as quality. The use of a short barcode was led by the available technology and did not provide the taxonomic resolution of the full length *rbcL* barcode, yet it offered good enough resolution and a cost-effective method to provide a solution to the Environment Agency.

With this in mind, this project also included a comparison between light microscopy, the accepted short-read Illumina method and a nanopore sequencing method using a longer amplicon covering the entire *rbcL* gene. The results from this comparison showed that there are still major challenges to be solved within diatom taxonomy and the correct assignment of sequence reads to diatom taxa which may be very similar in their *rbcL* sequence but are currently "distantly" related by naming conventions based upon frustule structure. This comparison also demonstrated that the light microscopy counts were significantly different in species abundance from the short- and long-read sequencing methods. Light microscopy is universally assumed to be the "ground truth" of diatom identification as the specimen can be seen and the frustule characterised yet diatom frustules can be very difficult to tell apart with light microscopy alone and the accuracy can be influenced by the experience of the microscopists. Higher resolution scanning electron microscopy has shown previously that identifications made by light microscopy can often be inconsistent and that the low taxonomic resolution can lead to an overestimation of geographical distributions and ranges of tolerance to environmental and grown conditions (Morales et al., 2001). The harsh chemical preparation methods used to prepare slides for LM can also affect the distribution of species observed, with weakly silicified species often being under-represented after slide preparation (Zgrundo et al., 2013). The results from this project showing that two different sequencing technologies using different primer sets and different amplicon lengths gave very similar relative abundance counts could not be discounted and it can be suggested that light microscopy is actually more inaccurate than traditional taxonomists in this area would concede. It has been suggested by some studies

that traditional taxonomic approaches should be abandoned in favour of sequencing based approaches (Baird & Hajibabaei, 2012; Woodward et al., 2013). There is potential to explore aspects of ecosystem function and biodiversity with NGS methods where traditional approaches struggle, for example, elucidating highly informative, but cryptic, species. However, a baseline of NGS versus LM was important to provide in order to provide comparable information given that LM was the method accepted by regulatory bodies at the start of the project.

The results of the three rivers experiments also provided the initial groundwork for future proofing the use of metabarcoding diatom assemblages in biomonitoring for water quality analysis. Should Illumina sequencing be entirely replaced by long-read sequencing in the future, the switch to such a technology for the Environment Agency should be simpler and less costly than an entirely new project to redevelop the method again from scratch. As current ecological classifications are based in part on assessment of diatom assemblages it is important that new methods are compared with the methods currently accepted by the regulatory bodies. The results showed that even with the relatively low quality of the nanopore sequencing undertaken on the diatom samples, they provided very similar results to the short-read Illumina sequencing of the same sample. At the time of writing we could not find any other studies which had directly compared Illumina short-read and Oxford Nanopore MinION metabarcoding from the same biological samples. Given that the Illumina sequences were at their technological limit of length and sequence quality and as the nanopore sequencing quality will only improve over time to match the quality of Illumina, the future is promising for long-read metabarcoding for ecological studies.

It is worth noting that the short-read Illumina sequences for the diatom assemblage experiments and the airborne fungal community experiments were treated differently with regards to initial post-QC processing. The diatom sequences were clustered into Operational Taxonomic Units (OTUs) with a similarity threshold of 97% (Nilsson et al., 2008). The fungal sequences were clustered into OTUs, but with a similarity threshold of 100% to reduce the computational requirements of searching the same sequence millions of times.

As such, the fungal short-read sequences can be thought of as Amplicon Sequence Variants (ASVs) rather than OTUs (Callahan et al., 2017). This difference in methodology was decided upon to enable higher resolution identifications in the fungal surveillance experiments as the intra specific difference between the ITS of some closely related fungal species can be much smaller than the recommended clustering percentage of 3% (Nilsson et al., 2008). There is an increasing number of studies to date which have evaluated nanopore sequencing for community ecology (Tedersoo et al., 2018, Krehenwinkel et al., 2019, Wurzbacher et al., 2019) and a minority decided to cluster their nanopore amplicon sequences into OTUs (Wurzbacher et al., 2019). The long-read nanopore amplicons in this project were not clustered into OTUs as it was felt to be a futile exercise given the indel-prone data. MinION nanopore sequences produced in early 2017 with the R9.4 pores and RNN basecaller were of lower quality than those produced at the time of writing (late 2018). The errors present in early 2017 nanopore sequences include small (1-5nt) insertions and deletions throughout the sequence (Laver et al., 2015) and it was assumed that the nanopore sequences would ultimately cluster into OTUs with only one sequence per OTU. Even if the clustering of nanopore sequences were successful and the number of sequences in each cluster were significant, it would be very difficult using current methods to decide upon a representative sequence for that cluster, as the most prevalent is typically recommended (Edgar et al., 2010).

In contrast to the experiments demonstrating methodology for monitoring of known diatom species, the aerosol fungal samples collected with Burkard samplers demonstrated the potential of metabarcoding for surveillance, where the species of interest may not yet be described. The aim of the surveillance experiments was to characterise the fungal communities present in six locations in eastern England over one month. With continuous Burkard sampling and 24 hour sample collections the plan was to accumulate a dataset which could be used to both track the abundance of known plant pathogens important to the UK (those present on the Plant Health Risk Register) and to investigate the potential for such a dataset to be used to uncover new species which may be present in abundance but not yet described. The data analysis proceeded as planned but the results were not as expected. The presence of sequences showing high identity to EPPO listed plant pathogens - in particular those not known to be

present in the United Kingdom - introduced the suspicion of one or more sample contamination events during DNA extraction, PCR amplification or library preparation in the laboratory. Contamination of samples has been determined in many clinical samples where environmental microorganisms would not be expected and in those samples the reagents or equipment is typically the source of contamination. Salter et al., (2014) was the first study to uncover reagent contamination during the metabarcoding processing of clinical samples, with varying levels of environmental microorganisms discovered in different kits and different batches. Contamination has also been shown to have a much greater effect in low biomass samples (Weiss et al., 2014). However, in environmental samples it is extraordinarily difficult to detect contamination of samples with species which may be naturally present in the sample to begin with. In the results presented here, the contamination was overt and able to be identified as certain quarantine species were known to be absent from the United Kingdom and had been present in the laboratories used to process the samples.

This work has determined the significance that should be placed upon preventing laboratory contamination when metabarcoding is to be used for environmental surveillance and monitoring studies. A one-off contamination event in a large dataset may be easy to identify, isolate and remove from or flag in a large long-term monitoring dataset where the samples are processed in batches. In shorter term monitoring datasets such as the short-read fungal dataset (samples over one month) the samples were processed in only two batches and separate contamination was introduced in each batch. The laboratory staff involved were experienced but had processed the samples in a laboratory where fungal plant pathogens had been grown previously and where cultures were present on the open bench. The Illumina libraries were prepared in a dedicated ‘clean’ laboratory used only for next-generation library preparation, which is in agreement with the current recommendations for processing eDNA samples with low biomass (Goldberg et al., 2016) and the technical and negative controls we included allowed for the point at which the contamination had been introduced to be identified rapidly. While there were certainly sequences which represented clear contamination of pathogens which should not have been there, a proportion of the sequences identified as being risk register pathogens were clear misidentifications. The combined reference database comprising UNITE (Nilsson

et al., 2018) and additional risk register pathogens was not comprehensive enough to cover all species in known genera and so some sequences were assigned to closely related species which we would not expect to be present, adding to the complication within the dataset. Despite an estimate of 2.2 to 3.8 million fungal species, only an estimated 3-8% have been described (Hawksworth and Lücking, 2017) and the gaps in the barcode databases resulted in some Illumina sequences within the dataset being identified as risk register pathogens with much lower percent identity alignments than would be expected from an accurate identification given intraspecific variation (Nilsson et al., 2008).

This project identified that while >99% Illumina sequences were assigned to a taxon, significantly fewer nanopore amplicon sequences obtained from the same DNA extract could be identified. Given the short- and long-read sequences were searched using an identical reference sequence database and local alignment search parameters. In community studies where the species identification is of less significance to the overall difference between samples, and where all sequences will be misidentified equally, this would not be a hindrance. In biomonitoring and surveillance this could introduce a number of errors which could easily lead to spurious conclusions about the ecology of the location sampled. The long-read nanopore sequences were more likely to not have an identification and while this may be due, in part, to the larger number of small insertions and deletions affecting the BLAST algorithm (Laver et al., 2015) it could equally be due to the design of the long PCR for the amplification of the whole ribosomal tandem repeat, with primers placed close to taxonomically informative regions.

In recent years, suggestions have been made by various groups to extend the DNA barcoding region to either organellar genome or full genome references to enable larger regions of sequence to be used for species identification (Coissac et al., 2016; Dodsworth, 2015; Straub et al., 2012). The recently announced Earth BioGenome Project (EBP) aims to “sequence, catalog[ue], and characterize the genomes of all Earth’s eukaryotic biodiversity over a period of 10 years” (Lewin et al., 2018). This project may provide much needed genome sequence data to expand the availability of metabarcoding markers to larger nuclear regions, or to realise the full promise of shotgun metagenomics for the detection and

characterisation of new species. After all, the use of metabarcoding for monitoring and surveillance only determines that a species is present in an environment, not what it is doing, how it is living and what its capabilities are.

Glossary

ASV	Amplicon Sequence Variant
BLAST	Basic Local Alignment Search Tool
COI	Cytochrome oxidase I
DARLEQ2	Diatoms for Assessing River and Lake Ecological Quality
DEFRA	Department for Environment, Food and Rural Affairs
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxyribonucleic acid
EA	Environment Agency
eDNA	Environmental DNA
eRNA	Environmental RNA
ELISA	Enzyme-Linked Immunosorbent Assay
EPPO	European Plant Protection Organisation
EQR	Ecological Quality Ratio
EU	European Union
IGS	Intergenic spacer
ITS	Internal Transcribed Spacer
LCA	Lowest Common Ancestor
LM	Light Microscopy
LSU	Large subunit (ribosomal)
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
NMDS	Non-metric Multidimensional Scaling
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate analysis of Variance
QIIME	Quantitative Insights Into Microbial Ecology
rDNA	Ribosomal DNA

RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
SSU	Small subunit (ribosomal)
TDI	Trophic Diatom Index
TDI4	Trophic Diatom Index when used with Light Microscopy weightings
TDI5	Trophic Diatom Index when used with Next-Generation Sequencing weightings.
UNITE	Database of fungal ITS sequences
UKTAG	United Kingdom Technical Advisory Group
WIMP	What's In My Pot
WFD	Water Framework Directive

Bibliography

Afshinnekoo, E., Meydan, C., Levy, S., Mason, C.E., 2015. Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics Article Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. Cell Systems 1–15.

Ahire, Y.R., Sangale, M.K., 2012. Survey of aeromycoflora present in vegetable and fruit market. Elixir Appl. Botany 52, 11381–11383.

Akulenko, R., Merl, M., Helms, V., 2016. BEclear: Batch Effect Detection and Adjustment in DNA Methylation Data. PLoS One 11, e0159921.

Almeida, S.F.P., Elias, C., Ferreira, J., Tornés, E., Puccinelli, C., Delmas, F., Dörflinger, G., Urbanič, G., Marcheggiani, S., Rosebery, J., Mancini, L., Sabater, S., 2014. Water quality assessment of rivers using diatom metrics across Mediterranean Europe: a methods intercalibration exercise. Sci. Total Environ. 476-477, 768–776.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol.

Arfi, Y., Buée, M., Marchand, C., Levasseur, A., Record, E., 2012. Multiple markers pyrosequencing reveals highly diverse and host-specific fungal communities on the mangrove trees *Avicennia marina* and *Rhizophora stylosa*. FEMS Microbiol. Ecol. 79, 433–444.

Artiola, J., Pepper, I.L., Brusseau, M.L., 2004. Environmental Monitoring and Characterization. Elsevier.

Bakker, M.G., 2018. A fungal mock community control for amplicon sequencing experiments. Mol. Ecol. Resour. 18, 541–556.

Baird, D.J., Hajibabaei, M. 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. Molecular Ecoloyg, 21 (8), 2039-2044.

Balint, M., Marton, O., Schatz, M., During, R., Grossart, H. 2018. Proper experimental design requires randomization/balancing of molecular ecology experiments. Ecology and Evolution 8(3), 1786-1793.

Bell, K.L., Burgess, K.S., Okamoto, K.C., Aranda, R., Brosi, B.J. 2016. Review and future prospects for DNA barcoding methods in forensic palynology. Forensic Science International: Genetics 21, 110-116.

Begerow, D., Nilsson, H., Unterseher, M., Maier, W., 2010. Current state and perspectives of fungal DNA barcoding and rapid identification procedures. Appl. Microbiol. Biotechnol. 87, 99–108.

Benítez-Páez, A., Portune, K.J., Sanz, Y., 2016. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. Gigascience 5, 4.

Bergeron, J., Drouin, G., 2008. The evolution of 5S ribosomal RNA genes linked to the rDNA units of fungal species. Curr. Genet. 54, 123–131.

Bokulich, N. a., Mills, D. a., 2013. Improved selection of internal transcribed spacer-specific primers enables quantitative, ultra-high-throughput profiling of fungal communities. Appl. Environ. Microbiol. 79, 2519–2526.

Bray, J.R., Curtis, J.T., 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol. Monogr. 27, 325–349.

Brittain, I., Selby, K., Taylor, M., Mumford, R., 2013. Detection of plant pathogen spores of economic significance on pollen trap slides. Journal of Phytopathology 161, 855–858.

Bulman, S.R., McDougal, R.L., Hill, K., Lear, G., 2018. Opportunities and limitations for DNA metabarcoding in Australasian plant-pathogen biosecurity. Australas. Plant Pathol.

Calderon, C., Ward, E., Freeman, J., Mccartney, A., 2002. Detection of airborne fungal spores sampled by rotating-arm and Hirst-type spore traps using polymerase chain reaction assays. J. Aerosol Sci. 33, 283–296.
Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J.
Callahan, B.J., Mcmurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., Holmes, S.P., 2015. DADA2 : High resolution sample inference from amplicon data. bioRxiv 13, 0–14.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.
Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Jeremy, E., Ley, R.E., Lozupone, C.A., Mcdonald, D., Muegge, B.D., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336.
Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A. 108 Suppl 1, 4516–4522.
CEN, 2014b. Water quality. Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters. EN14407:2014.
CEN, 2014a. Water quality–Guidance standard for the routine sampling and pre-treatment of benthic diatoms from rivers. EN13946:2014.
Chen, W., Hambleton, S., Seifert, K.A., Carisse, O., Diarra, M.S., Peters, R.D., Lowe, C., Chapados, J.T., Lévesque, C.A., 2018. Assessing Performance of Spore Samplers in Monitoring Aeromycobiota and Fungal Plant Pathogen Diversity in Canada. Appl. Environ. Microbiol. 84.
Christensen, J., 2001. Epidemiological concepts regarding disease monitoring and surveillance. Acta Vet. Scand. Suppl. 94, 11–16.
Clarke, K.R., 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. 18, 117–143.
Cleary, M., Nguyen, D., Marčiulynienė, D., Berlin, A., Vasaitis, R., Stenlid, J., 2016. Friend or foe? Biological and ecological traits of the European ash dieback pathogen *Hymenoscyphus fraxineus* in its native environment. Sci. Rep. 6, 21895.
Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423.
Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., Pawlowski., J. 2017. Predicting the Ecological Quality Status of marine environments from eDNA metabarcoding data using supervised machine learning. Environ. Sci. Technol. 51, 16, 9118-9126.
Cordier, T., Forster, D., Dufresne, Y., Martins, C.I.M., Stoeck, T., Pawlowski, J. 2018. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Mol. Ecol. Res. 18 (6), 1381-1391.
Coissac, E., Hollingsworth, P.M., Lavergne, S., Taberlet, P., 2016. From barcodes to genomes: extending the concept of DNA barcoding. Mol. Ecol. 25, 1423–1428.
CropMonitor, 2017. Crop Monitor [WWW Document]. URL http://www.cropmonitor.co.uk (accessed 17).
DEFRA, 2014. Protecting Plant Health - A Plant Biosecurity Strategy for Great Britain. Office of National Statistics.
de Goffau, M.C., Lager, S., Salter, S.J., Wagner, J., Kronbichler, A., Charnock-Jones, D.S., Peacock, S.J., Smith, G.C.S., Parkhill, J., 2018. Recognizing the reagent microbiome. Nat Microbiol 3, 851–853.
Deiner, K., Fronhofer, E.A., Mächler, E., Walser, J.-C., Altermatt, F., 2016.

Environmental DNA reveals that rivers are conveyer belts of biodiversity information. Nat. Commun. 7, 12544.
Deiner, K., Renshaw, M.A., Li, Y., Olds, B.P., Lodge, D.M., Pfrender, M.E., 2017. Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. Methods Ecol. Evol. 1–11.
De Jonge, M., Van de Vijver, B., Blust, R., Bervoets, L., 2008. Responses of aquatic organisms to metal pollution in a lowland river in Flanders: a comparison of diatoms and macroinvertebrates. Sci. Total Environ. 407, 615–629.
Dodsworth, S., 2015. Genome skimming for next-generation biodiversity analysis. Trends Plant Sci. 20, 525–527.
Douglas, W.Y., 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 Ecol. Evol. 3, 613–623.
Eckert, I.M.K., Littlefair, J.E., Zhang, G.K., Chain, F.J.J., Crease, T.J., Cristescu, M.E., 2018. Chapter One - Bioinformatics for Biomonitoring: Species Detection and Diversity Estimates Across Next-Generation Sequencing Platforms. In: Bohan, D.A., Dumbrell, A.J., Woodward, G., Jackson, M. (Eds.), Advances in Ecological Research. Academic Press, pp. 1–32.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.
Edwards, A., Debbonaire, A.R., Sattler, B., Mur, L.A.J., Hodson, A.J., 2016. Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N. bioRxiv.
Eland, L.E., Davenport, R., Mota, C.R., 2012. Evaluation of DNA extraction methods for freshwater eukaryotic microalgae. Water Res. 46, 5355–5364.
Elbrecht, V., Leese, F., 2016. Validation and development of freshwater invertebrate metabarcoding COI primers for Environmental Impact Assessment (No. e2044v3). PeerJ Preprints.
EPPO, 2012. First report of Chalara fraxinea in the United Kingdom. EPPO Rep. Serv. 4, 7–8.
Erland, S., Henrion, B., Martin, F., Glover, L.A., Alexanders, I.J., 1994. Identification of the ectomycorrhizal basidiomycete *Tylospora fibrillosa* Donk by RFLP analysis of the PCR-amplified ITS and IGS regions of ribosomal DNA. New Phytol. 126, 525–532.
Feio, M.J., Almeida, S.F.P., Craveiro, S.C., Calado, A.J., 2009. A comparison between biotic indices and predictive models in stream water quality assessment based on benthic diatom communities. Ecol. Indic. 9, 497–507.
Flückiger, B., Koller, T., Monn, C., 2000. Comparison of airborne spore concentrations and fungal allergen content. Aerobiologia 16, 393–396.
Flynn, J.M., Brown, E.A., Chain, F.J.J., MacIsaac, H.J., Cristescu, M.E., 2015. Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. Ecol. Evol. 5, 2252–2266.
Fourment, M., Holmes, E.C., 2016. Seqotron: a user-friendly sequence editor for Mac OS X. BMC Res. Notes 9, 106.
Galan, M., Razzauti, M., Bard, E., Bernard, M., Brouat, C., Charbonnel, N., Dehne-Garcia, A., Loiseau, A., Tatard, C., Tamisier, L., Vayssier-Taussat, M., Vignes, H., Cosson, J.-F., 2016. 16S rRNA Amplicon Sequencing for Epidemiological Surveys of Bacteria in Wildlife. mSystems 1.
Ganley, A.R.D., Kobayashi, T., 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Res. 17, 184–191.
Gardes, M., Bruns, T.D., 1996. ITS-RFLP matching for identification of fungi. Methods Mol. Biol. 50, 177–186.
Ghannoum, M.A., Jurevic, R.J., Mukherjee, P.K., Cui, F., Sikaroodi, M., Naqvi, A.,

Gillevet, P.M., 2010. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. PLoS Pathog. 6, e1000713.
Gibbons, S., Duvallet, C., Alm, E.J., 2017. Correcting for batch effects in case-control microbiome studies. bioRxiv.
Gilbert, J.A., Stephens, B., 2018. Microbiology of the built environment. Nat. Rev. Microbiol. 16, 661–670.
Glassing, A., Dowd, S.E., Galandiuk, S., Davis, B., Chiodini, R.J., 2016. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 8, 24.
Glover, R., 2018. enviropore: simple taxonomy identification and relative abundance calculation for MinION sequencing. Can be accessed at https://github.com/rachelglover/enviropore.
Goodwin, S., McPherson, J.D., McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 17, 333-351.
Gross, A., Hosoya, T., Queloz, V., 2014. Population structure of the invasive forest pathogen Hymenoscyphus pseudoalbidus. Mol. Ecol. 23, 2943–2960.
Guo, L., Sui, Z., Zhang, S., Ren, Y., Liu, Y., 2015. Comparison of potential diatom “barcode” genes (the 18S rRNA gene and ITS, COI, rbcL) and their effectiveness in discriminating and determining species taxonomy in the Bacillariophyta. Int. J. Syst. Evol. Microbiol. 65, 1369–1380.
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., 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 Ecol. 12, 28.
Hamsher, S.E., Evans, K.M., Mann, D.G., Poulíčková, A., Saunders, G.W., 2011. Barcoding diatoms: exploring alternatives to COI-5P. Protist 162, 405–422.
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. Mol. Ecol. 25, 3101–3119.
Hausner, G., Wang, X., 2005. Unusual compact rDNA gene arrangements within some members of the Ascomycota: evidence for molecular co-evolution between ITS1 and ITS2. Genome 48, 648–660.
Hawksworth, D.L., Lücking, R., 2017. Fungal Diversity Revisited: 2.2 to 3.8 Million Species. Microbiol Spectr 5.
Hebert, P.D.N., Cywinska, A., Ball, S.L., DeWaard, J.R., 2003a. Biological identifications through DNA barcodes. Proceedings of the Royal Society B 270, 313–321.
Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W., 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. PNAS (USA) 101, 14812–14817.
Hebert, P.D.N., Ratnasingham, S., DeWaard, J.R., 2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. Biol. Sci. 270 Suppl 1, S96–9.
Henrion, B., Tacon, F.L.E., Martin, F., 1992. Rapid identification of genetic variation of ectomycorrhizal fungi by amplification of ribosomal RNA genes. New Phytol. 122, 289–298.
Hering, D., Johnson, R.K., Kramm, S., Schmutz, S., Szoszkiewicz, K., Verdonschot, P.F.M., 2006. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. Freshw. Biol. 51, 1757–1785.
Heuch, J., 2014. What lessons need to be learnt from the outbreak of Ash Dieback Disease, Chalara fraxinea in the United Kingdom? Arboric. J. 36, 32–44.
Hildebrand, M., 2008. Diatoms, biomineralization processes, and genomics. Chem.

Rev. 108, 4855–4874.
Hirst, H., Jüttner, I., Ormerod, S.J., 2002. Comparing the responses of diatoms and macro-invertebrates to metals in upland streams of Wales and Cornwall. Freshw. Biol. 47, 1752–1765.
Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., {van Der Bank}, M., 2009. A DNA barcode for land plants. PNAS (USA) 106, 12794–12979.
Hugerth, L.W., Muller, E.E.L., Hu, Y.O.O., Lebrun, L.A.M., Roume, H., Lundin, D., Wilmes, P., Andersson, A.F., 2014. Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. PLoS One 9, e95567.
Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., Tappu, R., 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput. Biol. 12, e1004957.
IPPC, 2011. International Plant Protection Convention.
Iwen, P.C., Hinrichs, S.H., Rupp, M.E., 2002. Utilization of the internal transcribed spacer regions as molecular targets to detect and identify human fungal pathogens. Med. Mycol. 40, 87–109.
Jackson, C.J., Barton, R.C., Evans, E.G., 1999. Species identification and strain differentiation of dermatophyte fungi by analysis of ribosomal-DNA intergenic spacer regions. J. Clin. Microbiol. 37, 931–936.
Jain, M., Tyson, J.R., Loose, M., Ip, C.L.C., Eccles, D.A., O'Grady, J., Malla, S., Leggett, R.M., Wallerman, O., Jansen, H.J., Zalunin, V., Birney, E., Brown, B.L., Snutch, T.P., Olsen, H.E., MinION Analysis and Reference Consortium, 2017. MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. F1000Res. 6, 760.
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. Ecol. Lett. 16, 1245–1257.
Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127.
Jones, D.R., Baker, R.H.A., 2007. Introductions of non-native plant pathogens into Great Britain, 1970--2004. Plant Pathol. 56, 891–910.
Jones, H.M., Simpson, G.E., Stickle, A.J., Mann, D.G., 2005. Life history and systematics of *Petroneis* (Bacillariophyta), with special reference to British waters. Eur. J. Phycol. 40, 61–87.
Joshi, N., Fass, J., 2016. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33). 2011. URL https://githubcom/najoshi/sickle.
Juul, S., Izquierdo, F., Hurst, A., Dai, X., Wright, A., Kulesha, E., Pettett, R., Turner, D.J., 2015. What's in my pot? Real-time species identification on the MinION. bioRxiv 030742.
Kaczmarska, I., Reid, C., Moniz, M., 2007. Diatom taxonomy: morphology, molecules and barcodes. In: Proceedings of the 1st Central-European Diatom Meeting. Botanic Garden and Botanical Museum Berlin-Dahlem FU-Berlin, pp. 69–72.
Kahlert, M., Kelly, M.G., Albert, R.L., Almeida, S., Bešta, T., Blanco, S., Denys, L., Ector, L., Fránková, M., Hlúbiková, D., Others, 2012. Identification is a minor source of uncertainty in diatom-based ecological status assessments on a continent-wide scale: results of a European ring-test. Hydrobiologia 695, 109–124.
Karlsson, E., Lärkeryd, A., Sjödin, A., Forsman, M., Stenberg, P., 2015. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci. Rep. 5, 11996.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
Kelly M, Boonham N, Juggins S, Kille P, Mann D, Pass D, Sapp M, Sato S, Glover R, 2018. A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers. Environment Agency Report SC140024.

Kelly, M.G., 1999. Progress towards quality assurance of benthic diatom and phytoplankton analyses in the UK. Use of Algae for Monitoring Rivers III. Agence de l'Eau Artois-Picardie, Douai, France 208–215.
Kelly, M.G., Whitton, B.A., 1995. The Trophic Diatom Index: a new index for monitoring eutrophication in rivers. J. Appl. Phycol. 7, 433–444.
Kelly, M., Juggins, S., Guthrie, R., Pritchard, S., Jamieson, J., Rippey, B., Hirst, H., Yallop, M., 2008. Assessment of ecological status in U.K. rivers using diatoms. Freshw. Biol. 53, 403–422.
Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., Blomberg, S.P., Webb, C.O., 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26, 1463–1464.
Kerkhof, L.J., Dillon, K.P., Häggblom, M.M., McGuinness, L.R., 2017. Profiling bacterial communities by MinION sequencing of ribosomal operons. Microbiome 5, 116.
Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.-M., Humbert, J.-F., Bouchez, A., 2014. A next-generation sequencing approach to river biomonitoring using benthic diatoms. Freshw. Sci. 33, 349–363.
Kim, D., Hofstaedter, C.E., Zhao, C., Mattei, L., Tanes, C., Clarke, E., Lauder, A., Sherrill-Mix, S., Chehoud, C., Kelsen, J., Conrad, M., Collman, R.G., Baldassano, R., Bushman, F.D., Bittinger, K., 2017. Optimizing methods and dodging pitfalls in microbiome research. Microbiome 5, 52.
Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.
Klymus, K.E., Marshall, N.T., Stepien, C.A., 2017. Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. PLoS One 12, e0177643.
Kõljalg, U., Larsson, K.-H., Abarenkov, K., Nilsson, R.H., Alexander, I.J., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., Others, 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytol. 166, 1063–1068.
Konstantinova, P., Yli-Mattila, T., 2004. IGS-RFLP analysis and development of molecular markers for identification of *Fusarium poae, Fusarium langsethiae*, *Fusarium sporotrichioides* and *Fusarium kyushuense*. Int. J. Food Microbiol. 95, 321–331.
Korpelainen, D.H., Pietilainen, M.M., 2015. Diversity of indoor fungi as revealed by DNA metabarcoding. Genome 0, null.
Korpelainen, H., Pietiläinen, M., Huotari, T., 2015. Effective detection of indoor fungi by metabarcoding. Ann. Microbiol. 66, 495–498.
Kowalski, T., 2006. *Chalara fraxinea* sp. nov. associated with dieback of ash (*Fraxinus excelsior*) in Poland. For. Pathol. 36, 264–270.
Krehenwinkel, H., Pomerantz, A., Henderson, J.B., Kennedy, S.R., Lim, J.Y., Swamy, V., Shoobridge, J.D., Nipam, H.P., Gilespie, R.G., Prost, S. 2019. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. GigaScience 8(5)
Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.
Laver, T., Harrison, J., O'neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., Studholme, D.J. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular Detection and Quantification 3(2015), 1-8.
Lawrence, R., Cheffings, C.M., 2014. A summary of the impacts of ash dieback on UK biodiversity, including the potential for long-term monitoring and further research on management scenarios. Joint Nature Conservation Committee 501.
Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., Goldstein, M.M., Grigoriev, I.V., Hackett, K.J., Haussler, D., Jarvis, E.D., Johnson, W.E., Patrinos, A.,

Richards, S., Castilla-Rubio, J.C., van Sluys, M.-A., Soltis, P.S., Xu, X., Yang, H., Zhang, G., 2018. Earth BioGenome Project: Sequencing life for the future of life. Proc. Natl. Acad. Sci. U. S. A. 115, 4325–4333.
Lievens, B., Thomma, B.P.H.J., 2005. Recent Developments in Pathogen Detection Arrays : Implications for Fungal Plant Pathogens and Use in Practice. Phytopathology 95, 1374–1380.
Li, W., Godzik, A., 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.
Li, W., Wang, M.M., Wang, X.G., Cheng, X.L., Guo, J.J., Bian, X.M., Cai, L., 2016. Fungal communities in sediments of subtropical Chinese seas as estimated by DNA metabarcoding. Sci. Rep. 6, 26528.
Loman, N.J., Quinlan, A.R., 2014. Poretools: A toolkit for analyzing nanopore sequence data. Bioinformatics 30, 3399–3401.
Manimaran, S., Selby, H.M., Okrah, K., Ruberman, C., Leek, J.T., Quackenbush, J., Haibe-Kains, B., Bravo, H.C., Johnson, W.E., 2016. BatchQC: interactive software for evaluating sample and batch effects in genomic data. Bioinformatics 32, 3836–3838.
Mann, D.G., Droop, S.J.M., 1996. Biodiversity, biogeography and conservation of diatoms. In: Kristiansen J. (eds) Biogeography of Freshwater Algae. Developments in Hydrobiology, vol 188. Springer, Dordrecht.
Mann, D.G. 1996. Chloroplast morphology, movements and inheritance in Diatoms. In: Chaudhary & Agrawal (eds) Cytology, Genetics and Molecular Biology of Algae. SPB Academic Publichsing, Amsterdam.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12.
Ma, X., Baron, J.L., Vikram, A., Stout, J.E., Bibby, K., 2015. Fungal diversity and presence of potentially pathogenic fungi in a hospital hot water system treated with on-site monochloramine. Water Res. 71, 197–206.
McCune, B., Grace, J.B., Urban, D.L., 2002. Analysis of ecological communities. MjM software design Gleneden Beach, OR.
McGuire, K.L., Payne, S.G., Palmer, M.I., Gillikin, C.M., Keefe, D., Kim, S.J., Gedallovich, S.M., Discenza, J., Rangamannar, R., Koshner, J. a., Massmann, A.L., Orazi, G., Essene, A., Leff, J.W., Fierer, N., 2013. Digging the New York City Skyline: soil fungal communities in green roofs and city parks. PLoS One 8, e58020.
Menegon, M., Cantaloni, C., Rodriguez-Prieto, A., Centomo, C., Abdelfattah, A., Rossato, M., Bernardi, M., Xumerle, L., Loader, S., Delledonne, M., 2017. On site DNA barcoding by nanopore sequencing. PLoS One 12, e0184741.
Minamoto, T., Yamanaka, H., Takahara, T., Honjo, M.N., Kawabata, Z. 'ichiro, 2012. Surveillance of fish species composition using environmental DNA. Limnology 13, 193–197.
Morales, E.A., Siver, P.A., Trainor, F.R., 2001. Identification of diatoms (Bacillariophyceae) during ecological assessments: Comparison between Light Microscopy and Scanning Electron Microscopy techniques. Proceedings of the Academy of Natural Sciences of Philadelphia 151, 95–103.
Morin, S., Gómez, N., Tornés, E., Licursi, M., Rosebery, J., 2016. Benthic diatom monitoring and assessment of freshwater environments: standard methods and future challenges. Aquatic Biofilms 111.
Mosier, A.C., Miller, C.S., Frischkorn, K.R., Ohm, R.A., Li, Z., LaButti, K., Lapidus, A., Lipzen, A., Chen, C., Johnson, J., Lindquist, E.A., Pan, C., Hettich, R.L., Grigoriev, I.V., Singer, S.W., Banfield, J.F., 2016. Fungi Contribute Critical but Spatially Varying Roles in Nitrogen and Carbon Cycling in Acid Mine Drainage. Front. Microbiol. 7, 238.
Moss, T., Stefanovsky, V.Y., 1995. Promotion and Regulation of Ribosomal Transcription in Eukaryotes by RNA Polymerase1. In: Progress in Nucleic Acid Research and Molecular Biology. Elsevier, pp. 25–66.
Mullineux, T., Hausner, G., 2009. Evolution of rDNA ITS1 and ITS2 sequences and

RNA secondary structures within members of the fungal genera Grosmannia and Leptographium. Fungal Genet. Biol. 46, 855–867.
Nazar, R.N., 2004. Ribosomal RNA processing and ribosome biogenesis in eukaryotes. IUBMB Life 56, 457–465.
Nguyen, N.H., Smith, D., Peay, K., Kennedy, P., 2015. Parsing ecological signal from noise in next generation amplicon sequencing. New Phytol. 205, 1389–1393.
Nicolaisen, M., West, J.S., Sapkota, R., Canning, G.G.M., Schoen, C., Justesen, A.F., 2017. Fungal Communities Including Plant Pathogens in Near Surface Air Are Similar across Northwestern Europe. Front. Microbiol. 8, 1729.
Nilsson, R.H., Kristiansson, E., Ryberg, M., Hallenberg, N., Larsson, K., 2008. Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. Evol. Bioinform. Online 4, 193–201.
Nilsson, R.H., Taylor, A.F.S., Bates, S.T., Thomas, D., Bengtsson-palme, J., Callaghan, T.M., Douglas, B., Griffith, G.W., Ucking, R.L., Suija, A.V.E., Taylor, D.L.E.E., Teresa, M., 2013. Towards a unified paradigm for sequence-based identification of fungi. Mol. Ecol. 22, 5271–5277.
Nyamundanda, G., Poudel, P., Patil, Y., Sadanandam, A., 2017. A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies. Sci. Rep. 7, 10849.
Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M.H.H., Oksanen, M.J., Suggests, M., 2007. The vegan package. Community ecology package 10, 631–637.
Pantou, M.P., Mavridou, A., Typas, M.A., 2003. IGS sequence variation, group-I introns and the complete nuclear ribosomal DNA of the entomopathogenic fungus *Metarhizium*: excellent tools for isolate detection and phylogenetic analysis. Fungal Genet. Biol. 38, 159–174.
Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289–290.
Parker, H.S., Leek, J.T., 2012. The practical effect of batch on genomic prediction. Stat. Appl. Genet. Mol. Biol. 11, Article 10.
Peccia, J., Hernandez, M., 2006. Incorporating polymerase chain reaction-based identification, population characterization, and quantification of microorganisms into aerosol science: A review. Atmos. Environ. 40, 3941–3961.
Peres-Neto, P.R., Jackson, D.A., 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia 129, 169–178.
Pomerantz, A., Penafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L.A., Barrio-Amoros, C.L., Salazar-Valenzuela, D., Prost, S., 2017. Real-time DNA barcoding in a remote rainforest using nanopore sequencing. bioRxiv.
Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L.A., Barrio-Amorós, C.L., Salazar-Valenzuela, D., Prost, S., 2018. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. Gigascience 7.
Potter, C., Urquhart, J., 2016. Tree disease and pest epidemics in the Anthropocene: A review of the drivers, impacts and policy responses in the UK. For. Policy Econ.
Prygiel, J., Carpentier, P., Almeida, S., Coste, M., Druart, J.-C., Ector, L., Guillard, D., Honoré, M.-A., Iserentant, R., Ledeganck, P., Lalanne-Cassou, C., Lesniak, C., Mercier, I., Moncaut, P., Nazart, M., Nouchet, N., Peres, F., Peeters, V., Rimet, F., Rumeau, A., Sabater, S., Straub, F., Torrisi, M., Tudesque, L., Van de Vijver, B., Vidal, H., Vizinet, J., Zydek, N., 2002. Determination of the biological diatom index (IBD NF T 90–354): results of an intercomparison exercise. J. Appl. Phycol. 14, 27–39.
Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Neal, K., Nye, K., Peters, T., De Pinna, E., Robinson, E., Struthers, K., Webber, M., Catto, A., Dallman, T.J., Hawkey, P., Loman, N.J., 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biol.

16, 114.
Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J.H.J., Becker-Ziaja, B., Boettcher, J.P., Cabeza-Cabrerizo, M., Camino-Sánchez, Á., Carter, L.L., Doerrbecker, J., Enkirch, T., García-Dorival, I., Hetzelt, N., Hinzmann, J., Holm, T., Kafetzopoulou, L.E., Koropogui, M., Kosgey, A., Kuisma, E., Logue, C.H., Mazzarelli, A., Meisel, S., Mertens, M., Michel, J., Ngabo, D., Nitzsche, K., Pallasch, E., Patrono, L.V., Portmann, J., Repits, J.G., Rickett, N.Y., Sachse, A., Singethan, K., Vitoriano, I., Yemanaberhan, R.L., Zekeng, E.G., Racine, T., Bello, A., Sall, A.A., Faye, O., Faye, O., Magassouba, N., Williams, C.V., Amburgey, V., Winona, L., Davis, E., Gerlach, J., Washington, F., Monteil, V., Jourdain, M., Bererd, M., Camara, A., Somlare, H., Camara, A., Gerard, M., Bado, G., Baillet, B., Delaune, D., Nebie, K.Y., Diarra, A., Savane, Y., Pallawo, R.B., Gutierrez, G.J., Milhano, N., Roger, I., Williams, C.J., Yattara, F., Lewandowski, K., Taylor, J., Rachwal, P., Turner, D.J., Pollakis, G., Hiscox, J.A., Matthews, D.A., O'Shea, M.K., Johnston, A.M., Wilson, D., Hutley, E., Smit, E., Di Caro, A., Wölfel, R., Stoecker, K., Fleischmann, E., Gabriel, M., Weller, S.A., Koivogui, L., Diallo, B., Keïta, S., Rambaut, A., Formenty, P., Günther, S., Carroll, M.W., 2016. Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228–232.
Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.
Rampersad, S.N., 2014. ITS1, 5.8S and ITS2 secondary structure modelling for intra-specific differentiation among species of the *Colletotrichum gloeosporioides* sensu lato species complex. Springerplus 3, 684.
Rang, F.J., Kloosterman, W.P., de Ridder, J., 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19, 90.
Ranjard, L., Poly, F., Lata, J.C., Mougel, C., Thioulouse, J., Nazaret, S., 2001. Characterization of bacterial and fungal soil communities by automated ribosomal intergenic spacer analysis fingerprints: biological and methodological variability. Appl. Environ. Microbiol. 67, 4479–4487.
Rauwolf, U., Golczk, H., greiner, S., Hermann, R.G., 2010. Variable amounts of DNA related to the size of chloroplasts III: biological determinations of DNA amounts per organelle. Molecular Genetics and Genomics, 283 (1), 35-47.
R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rees, H.C., Maddison, B.C., Middleditch, D.J., Patmore, J.R.M., Gough, K.C., 2014. REVIEW: The detection of aquatic animal species using environmental DNA--a review of eDNA as a survey tool in ecology. J. Appl. Ecol. 51, 1450–1459.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.
Ryberg, M., Nilsson, R.H., 2018. New light on names and naming of dark taxa. MycoKeys 31–39.
Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., Walker, A.W., 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 12, 87.
Sato, Y., Miya, M., Fukunaga, T., Sado, T., Iwasaki, W., 2018. MitoFish and MiFish Pipeline: A Mitochondrial Genome Database of Fish with an Analysis Pipeline for Environmental DNA Metabarcoding. Mol. Biol. Evol. 35, 1553–1555.
Schattner, P., 2002. Searching for RNA genes using base-composition statistics. Nucleic Acids Res. 30, 2076–2082.
Schirmer, M., D'Amore, R., Ijaz, U.Z., Hall, N., Quince, C., 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics 17, 125.

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., Quince, C., 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 1–16.
Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R. a., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.
Schmidt, P.-A., Bálint, M., Greshake, B., Bandow, C., Römbke, J., Schmitt, I., 2013. Illumina metabarcoding of a soil fungal community. Soil Biol. Biochem. 65, 128–132.
Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Fungal Barcoding Consortium, Fungal Barcoding Consortium Author List, 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. U. S. A. 109, 6241–6246.
Schweigkofler, W., Donnell, K.O., Garbelotto, M., 2004. Detection and Quantification of Airborne Conidia of Fusarium circinatum , the Causal Agent of Pine Pitch Canker , from Two California Sites by Using a Real-Time PCR Approach Combined with a Simple Spore Trapping Method Detection and Quantification of Airb. Appl. Environ. Microbiol. 70, 3512–3520.
Shin, J., Lee, S., Go, M.-J., Lee, S.Y., Kim, S.C., Lee, C.-H., Cho, B.-K., 2016. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. Sci. Rep. 6, 29681.
Slatko, B.E., Gardner, A.F., Ausubel, F.M. 2018. Overview of Next-Generation Sequencing Technologies. Current Protocols in Molecular Biology 122(1)
Somervuo, P., Yu, D.W., Xu, C.C.Y., Ji, Y., Hultman, J., Wirta, H., Ovaskainen, O., 2017. Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. Methods in Ecology and Evolution. 8, 398-407
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. Proc. Natl. Acad. Sci. USA. 105 (36) 13486-13491.
Staats, M., Arulandhu, A.H., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., Prins, T.W., Kok, E. 2016. Advances in DNA metabarcoding for food and wildlife forensic species identification. Analytical and Bioanalytical Chemistry. 408 (17) 4615-4630.
Stevenson, J., 2014. Ecological assessments with algae: a review and synthesis. J. Phycol. 50, 437–461.
Stevenson, R.J., Pan, Y., van Dam, H., 2010. Assessing environmental conditions in rivers and streams with diatoms. The Diatoms: Applications for the Environmental and Earth Sciences, 2nd ed. Cambridge University Press, Cambridge 5785.
Straub, S.C.K., Parks, M., Weitemier, K., 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. American Journal of.
Sugiyama, A., Vivanco, J.M., Biology, R., 2010. Pyrosequencing Assessment of Soil Microbial Communities. Plant Dis. 94, 1329–1335.
Taberlet, P., Coissac, E., Pompanon, F., 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular.
Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., Willerslev, E., 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21, 2045–2050.
Takahara, T., Minamoto, T., Doi, H., 2013. Using environmental DNA to estimate the distribution of an invasive fish species in ponds. PLoS One 8, e56584.
Targetti, S., Herzog, F., Geijzendorffer, I.R., Wolfrum, S., Arndorfer, M., Balàzs, K., Choisis, J.P., Dennis, P., Eiter, S., Fjellstad, W., Friedel, J.K., Jeanneret, P., Jongman, R.H.G., Kainz, M., Luescher, G., Moreno, G., Zanetti, T., Sarthou, J.P., Stoyanova, S., Wiley, D., Paoletti, M.G., Viaggi, D., 2014. Estimating the cost of different strategies for measuring farmland biodiversity: Evidence from a Europe-

wide field evaluation. Ecol. Indic. 45, 434–443.
Taylor, D.L., McCormick, M.K., 2008. Internal transcribed spacer primers and sequences for improved characterization of basidiomycetous orchid mycorrhizas. New Phytol. 177, 1020–1033.
Tedersoo, L., Anslan, S., Bahram, M., Põlme, S., Riit, T., Liiv, I., Kõljalg, U., Kisand, V., Nilsson, H., Hildebrand, F., Bork, P., Abarenkov, K., 2015. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MycoKeys 10, 1–43.
Tedersoo, L., Bahram, M., Puusepp, R., Nilsson, R.H., James, T.Y., 2017. Novel soil-inhabiting clades fill gaps in the fungal tree of life. Microbiome 5, 42.
Tedersoo, L., Tooming-Klunderud, A., Anslan, S., 2018. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytol. 217, 1370–1385.
The Water Environment (Water Framework Directive) (England and Wales) Regulations, 2008.
Timmermann, V., Børja, I., Hietala, A.M., Kirisits, T., Solheim, H., 2011. Ash dieback: Pathogen spread and diurnal patterns of ascospore dispersal, with special emphasis on Norway. EPPO Bulletin 41, 14–20.
Toju, H., Tanabe, A.S., Yamamoto, S., Sato, H., 2012. High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples. PLoS One 7, e40863.
Toudjani, A.A., Çelekli, A., Yonca Gümüş, E., Kayhan, S., Ömer Lekesiz, H., Çetin, T., 2017. A new diatom index to assess ecological quality of running waters: a case study of water bodies in western Anatolia. Annales de Limnologie - International Journal of Limnology 53, 333–343.
Tremblay, J., Singh, K., Fern, A., Kirton, E.S., He, S., Woyke, T., Lee, J., Chen, F., Dangl, J.L., Tringe, S.G., 2015. Primer and platform effects on 16S rRNA tag sequencing. Front. Microbiol. 6, 771.
UK Plant Health Risk Register [WWW Document], n.d. URL https://secure.fera.defra.gov.uk/phiw/riskRegister/ (accessed 8.7.18).
United Kingdom Technical Advisory Group (WFD-UKTAG), 2014. UKTAG River Assessment Methods: Phytobenthos - Diatoms for Assessing River and Lake Ecological Quality (River DARLEQ2). Water Framework Directive—United Kingdom Technical Advisory Group (WFD-UKTAG).
Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G., 2012. Primer3—new capabilities and interfaces. Nucleic Acids Res. 40, e115–e115.
Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., Bouchez, A., 2017a. Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? Freshw. Sci. 36, 162–177.
Vasselon, V., Rimet, F., Tapolczai, K., Bouchez, A., 2017b. Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). Ecol. Indic. 82, 1–12.
Vilgalys, R., Barns, S.M., Gonzalez, D., Hibbett, D.S., 1992. Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. Mol. Phylogenet. Evol.
Visco, J.A., Apothéloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L., Pawlowski, J., 2015. Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data. Environ. Sci. Technol. 49, 7597–7605.
Walsh, K., Korimbocus, J., Boonham, N., Jennings, P., Hims, M., 2005. Using Real-time PCR to Discriminate and Quantify the Closely Related Wheat Pathogens Oculimacula yallundae and Oculimacula acuformis. Journal of Phytopathology 153, 715–721.
Walter, M.C., Zwirglmaier, K., Vette, P., Holowachuk, S.A., Stoecker, K., Genzel, G.H., Antwerpen, M.H., 2017. MinION as part of a biomedical rapidly deployable laboratory. J. Biotechnol. 250, 16–22.
Ward, R.D., Hanner, R., Hebert, P.D.N., 2009. The campaign to DNA barcode all

fishes, FISH-BOL. J. Fish Biol. 74, 329–356.
Water Framework Directive, D., 2000. 60/EC of the European Parliament and of the Council. Official Journal of the European Communities L 327, 1–72.
Weiss, S., Amir, A., Hyde, E.R., Metcalf, J.L., Song, S.J., Knight, R. 2014. Tracking down the sources of experimental contamination in microbiome studies. Genome Biol. 15: 564
Wen, C., Wu, L., Qin, Y., Van Nostrand, J.D., Ning, D., Sun, B., Xue, K., Liu, F., Deng, Y., Liang, Y., Zhou, J., 2017. Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. PLoS One 12, e0176716.
Wick, R.R., Judd, L.M., Gorrie, C.L., Holt, K.E., 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microbial Genomics.
Wicker, T., Schlagenhauf, E., Grander, A., Close, T.J., Keller, B., Stein, N. 2006. 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7:275.
Williams, R.H., Ward, E., Mccartney, H.A., 2001. Methods for Integrated Air Sampling and DNA Analysis for Detection of Airborne Fungal Spores Methods for Integrated Air Sampling and DNA Analysis for Detection of Airborne Fungal Spores. Appl. Environ. Microbiol. 67, 2453–2459.
Wurzbacher, C., Larsson, E., Bengtsson-Palme, J., Van den Wyngaert, S., Svantesson, S., Kristiansson, E., Kagami, M., Henrik Nilsson, R., 2018. Introducing ribosomal tandem repeat barcoding for fungi. bioRxiv.
Zgrundo, A., Lemke, P., Pniewski, F., Cox, E.J., Latala, A. 2013. Morphological and molecular phylogenetic studies on *Fistulifera saprophila*. Diatom Research, 28 (4), 431-443
Zhang, J., Kobert, K., Flouri, T., Stamatakis, A., 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30, 614–620.
Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., Gemeinholzer, B., 2015. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Mol. Ecol. Resour. 15, 526–542.

Appendix I: Full report to the Environment Agency.

Environment Agency

Evidence

A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers

SC140024/R

We are the Environment Agency. We protect and improve the environment.

Acting to reduce the impacts of a changing climate on people and wildlife is at the heart of everything we do.

We reduce the risks to people, properties and businesses from flooding and coastal erosion.

We protect and improve the quality of water, making sure there is enough for people, businesses, agriculture and the environment. Our work helps to ensure people can enjoy the water environment through angling and navigation.

We look after land quality, promote sustainable land management and help protect and enhance wildlife habitats. And we work closely with businesses to help them comply with environmental regulations.

We can't do this alone. We work with government, local councils, businesses, civil society groups and communities to make our environment a better place for people and wildlife.

This report is the result of research commissioned and funded by the Environment Agency.

Published by:
Environment Agency, Horizon House, Deanery Road, Bristol, BS1 5AH

www.environment-agency.gov.uk

ISBN: 978-1-84911-406-6

Further copies of this report are available from our publications catalogue:
http://www.gov.uk/government/publications

or our National Customer Contact Centre:
T: 03708 506506

Email: enquiries@environment-agency.gov.uk

Author(s):
Martyn Kelly, Neil Boonham, Steve Juggins, , Peter Kille, David Mann, Daniel Pass, Melanie Sapp, Shinya Sato, Rachel Glover

Dissemination Status:
Publicly available

Keywords:
Diatoms, metabarcoding, NGS, TDI, DNA, ecological, assessment, sequencing, barcoding, barcode

Research Contractors:
Bowburn Consultancy, 11 Monteigne Drive, Durham, DH6 5QB

Fera Science Ltd, National Agri-Food Innovation Campus, Sand Hutton, York, YO41 1LZ

Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR

Cardiff University, Cardiff School of Biosciences, Cardiff, CF10 3AT

Environment Agency's Project Manager:
Kerry Walsh, Research, Analysis and Evaluation

Collaborator(s):
Department for Environment, Food and Rural Affairs (Defra)

Project Number:
SC140024

Evidence at the Environment Agency

Scientific research and analysis underpins everything the Environment Agency does. It helps us to understand and manage the environment effectively. Our own experts work with leading scientific organisations, universities and other parts of the Defra group to bring the best knowledge to bear on the environmental problems that we face now and in the future. Our scientific work is published as summaries and reports, freely available to all.

This report is the result of research commissioned by the Environment Agency's Research, Analysis and Evaluation group.

You can find out more about our current science programmes at https://www.gov.uk/government/organisations/environment-agency/about/research

If you have any comments or questions about this report or the Environment Agency's other scientific work, please contact research@environment-agency.gov.uk.

Professor Doug Wilson
Director, Research, Analysis and Evaluation

Executive summary

The UK currently uses diatoms as part of a suite of ecological methods to inform decision-making associated with EU directives (Water Framework Directive, Urban Wastewater Treatment Directive, Habitats Directive) on water quality in rivers and lakes. When used alongside evaluations of other components of the aquatic biota, these provide a measure of the health of aquatic ecosystems. This in turn supports decision-making within catchments to ensure the delivery of critical ecosystem services. Current methods are based on light microscopy (LM), underpinned by European standards and producing outcomes that have been verified via the EU's intercalibration exercise.

Current biological assessment is a time-consuming process requiring highly skilled individuals to analyse and interpret data. There are several sources of uncertainty in the pathway from sample collection to data interpretation; one of these is the process of identification and enumeration of the organisms. In the case of diatoms, uncertainty associated with this stage can be controlled by training and quality control but, when combined with the time required to analyse a sample, and multiplied by the number of sites for which data are required, these add up to a substantial resource commitment. Alternative approaches that offer a similar level of precision at a lower cost would, therefore, be very attractive.

Another complication in the use of diatoms for ecological assessment is that their widespread adoption, particularly for assessments associated with the Water Framework Directive, has taken place alongside a paradigm shift in understanding of their taxonomy and phylogenetics. There is now known to be considerable taxonomic diversity within aggregates formerly thought to be single species. This diversity often pushes the capabilities of optical microscopy and analysts to the limit, and there is a real possibility that the use of a molecular approach may help to unlock taxonomic information in a form that can be used for ecological assessments.

Molecular techniques offer a potentially more cost-effective alternative and complementary approach to ecological assessment, with scope for improved efficiency and reduced analytical error through automation and standardisation. Recent developments combining DNA barcoding with next generation sequencing (NGS) enable DNA from whole communities of organisms to be sequenced simultaneously ('metabarcoding') in an assessment.

The overall aim of the project was to develop a high-throughput, cost-effective method for identifying and quantifying diatom taxa from environmental samples in a manner suitable for calculating the Trophic Diatom Index (TDI) and associated metrics using NGS for Water Framework Directive classifications.

This report presents the results of the first large-scale proof of concept to establish the suitability of metabarcoding – combining DNA barcodes (targeting the chloroplast rbcL gene) of diatoms with NGS – for the quantitative ecological assessment of diatoms.

- A 'gold standard' diatom rbcL barcode reference database of known diatom species was produced by isolating and culturing diatom species from water bodies of different ecological quality. Although the barcode database currently contains only 176 species or less than 10% of the diatom species that have been described from the UK,[1] it includes representatives of most of the commonly encountered taxa. It was demonstrated that this is sufficient to account for most of the variation in TDI analyses. Occasional

[1] This number is increasing through the addition of barcodes from online databases.

misclassifications may occur when taxa that are absent from the barcode database are abundant in a sample.

- A good quality barcode reference database is the backbone to any metabarcoding approach that requires taxonomy assignment. Culturing diatom species is a specialised, resource intensive exercise. An unexpected outcome of this project was the ability to 'discover' new barcodes by inferring species using NGS and bypassing the need to culture strains. Additional species were added from other online databases.
- A field sampling strategy for the collection and preservation of diatom samples was developed.
- A protocol for the extraction and amplification of DNA from environmental samples, suitable for high-throughput automation, was produced.
- A short rbcL barcode has been evaluated that allows the simultaneous amplification of a DNA fragment from a large number of diatom taxa while retaining taxonomic resolution. To the project team's knowledge, this is the first report of the use of this region of the rbcL gene for metabarcoding diatoms.
- The fragment is of a size (340 base pairs) that enables it to be analysed using the most cost-effective sequencing platform currently available (MiSeq™ from Illumina).
- A bioinformatics pipeline was developed to match NGS outputs with the relevant species in the barcode reference database. The pipeline is also capable of screening out non-diatom algae at an early stage and includes routines to manipulate data and produce an output in a form suitable for use by the Environment Agency to calculate diatom metrics for water body classification.
- The relationship (similarities, differences, uncertainties) between NGS and LM has been evaluated and a new variant of the current TDI (TDI4) for NGS (TDI5) has been developed. Despite an incomplete rbcL barcode reference database and observed variability in the relative abundance of certain taxa evaluated using LM and NGS, significant correlation between the current LM TDI4 and a new NGS TDI5 has been shown.

Overall, the outcomes of this study are very positive and a method that is compatible with the latest NGS technologies has been developed. The intention was to develop a molecular 'mirror' of the existing diatom assessment method, and although not a 1:1 relationship, significant correlation between the 2 approaches has been demonstrated. The aspiration of producing a molecular 'mirror' of the existing LM approach is a sensible starting point as it forces close examination of the relationship between the NGS and 'traditional' data. It should also be borne in mind that the traditional LM approach is itself a constrained approach which is used to generate a summarised view of reality. Therefore, the 2 approaches offer alternative views of the river ecosystem that need to be reconciled; it is rarely as simple as deciding that one method is 'right or that it is 'better' than the alternative.

Given this understanding of the relationship between the 2 approaches, it will be possible to begin to consider how to provide added value to that contained within the NGS data, exploiting the intrinsic information on diversity using operational taxonomic unit information in combination with species assessments. So long as these metrics can be linked to legislative drivers such as the Water Framework Directive, then an NGS metric may be effective.

Acknowledgements

We are extremely grateful to Tim Jones (Environment Agency) for his invaluable and continuous technical support throughout the project and to Rosetta Blackman (formerly Environment Agency, now University of Hull) for co-ordinating the collection of diatom samples for molecular analysis. Sincere thanks also go to operational staff from the Environment Agency for carrying out the light microscopy analysis on the calibration dataset samples and to the Scottish Environmental Protection Agency, Natural Resources Wales and Northern Ireland Environment Agency for providing reference samples. We also thank Sarah Pritchard (Beacon Biological) for help with preparing diatom samples.

Contents

List of tables and figures

1 Introduction

1.1 Background to UK diatom assessment

The UK currently uses diatoms as part of a suite of ecological methods (Box 1) to classify the quality of water bodies (rivers and lakes) in line with EU directives (Water Framework Directive, Urban Wastewater Treatment Directive, Habitats Directive). When used alongside evaluations of other components of the aquatic biota, these provide a measure of the health of aquatic ecosystems. This, in turn, supports decision-making within catchments to ensure the delivery of critical ecosystem services. Current methods are based on light microscopy (LM), underpinned by European standards (CEN 2014a, 2014b), and producing outcomes that have been verified via the EU's intercalibration exercise (European Commission 2008, 2013).

Box 1: Ecological assessment using diatoms

Diatoms are a group of microscopic plant-like organisms that are widespread in aquatic habitats throughout the world. Along with other algae, they play an important role in natural ecosystems and make a major contribution to global primary productivity. Those algae that are found attached to submerged surfaces such as stones and plant stems are referred to as 'phytobenthos'; European legislation requires that these are examined as part of assessments of the health (ecological status) of lakes and rivers.

In the UK, this was achieved using the Trophic Diatom Index (TDI). The first version (Kelly and Whitton 1995) has been updated several times and the version currently used by UK agencies is TDI4. The Water Framework Directive required that the condition of water bodies was expressed as a ratio – the Ecological Quality Ratio (EQR) – using the index value expected under conditions of no or minimal human impact as the denominator (Kelly et al. 2008, Bennion et al. 2014). This led to the development of a new tool, DARLEQ (Diatoms for Assessing River and Lake Ecological Quality), which calculated the EQR as the observed TDI divided by the expected TDI for any lake or river. This, too, has been updated, as a result of extensive testing and comparisons with macrophyte assessments; the current tool is DARLEQ2. For the Water Framework Directive, the results from DARLEQ2 are combined with those from macrophyte assessments (LEAFPACS 2) to give an overall assessment for the biological quality element 'macrophytes and phytobenthos'.

The TDI is based on a weighted average equation. Diatom taxa are each assigned a score from 1 (nutrient sensitive) to 5 (nutrient tolerant). The average sensitivity of all the taxa in the sample, each weighted by the number of individuals for that taxon, determines the final value of the TDI. The TDI scores range from 0 (very low nutrients) to 100 (very high nutrients). The EQR is calculated based on observed data and predicted reference values, resulting in a scale which ranges from 0 to 1 and which is itself divided to give 5 ecological status classes: High, Good, Moderate, Poor or Bad.

More information on the UK methods can be found from the website of the Water Framework Directive UK Technical Advisory Group (UK TAG):

- Rivers – phytobenthos (www.wfduk.org/resources/rivers-phytobenthos)
- Lakes – phytobenthos (www.wfduk.org/resources/lakes-phytobenthos)

The current method of biological assessment is a time-consuming process, requiring highly skilled individuals to identify the diatoms at the species level and interpret the data. There are also several sources of uncertainty in the pathway from sample

collection to data interpretation, one of which is the process of identification and enumeration of the organisms.

In the case of diatoms, the uncertainty associated with this stage can be controlled by training and quality control. When combined with the time required to analyse a sample and multiplied by the number of sites for which data are required, this adds up to a substantial resource commitment. Alternative approaches that offer a similar level of precision at a lower cost would therefore be very attractive.

Another complication to the use of diatoms for ecological assessment is that their widespread adoption –, particularly for assessments associated with the Water Framework Directive (Kelly 2013) – has taken place alongside a paradigm shift in understanding of their taxonomy and phylogenetics. Several workers have shown that there is considerable taxonomic diversity within aggregates formerly thought to be single species (see, for example, Mann et al. 2008, Trobajo et al. 2009, Kermarrac et al. 2013, Rovira et al. 2015). This diversity often pushes the capabilities of optical microscopy and analysts to the limit. There is also a real possibility that the use of a molecular approach may help to unlock taxonomic information in a form that can be used for ecological assessments (Mann et al. 2010).

1.2 Molecular approach to diatom assessment

Molecular techniques have the potential to overcome many of the hurdles facing the UK and other European regulators in monitoring our environments. They offer an alternative to traditional approaches, with the scope for improved efficiency and reduced analytical error through automation and standardisation.

Molecular techniques use the variation in the genetic code – deoxyribonucleic acid (DNA) – to distinguish between individuals of the same species or to identify specific species. A variety of techniques are available, each with their own strengths and limitations; there is no single ‘one-size-fits-all’ solution (Environment Agency 2011).

Many of the techniques have been around for a number of years. However, it is only recently that the science has developed to a level where complex species assemblages can be identified and given a semi-quantitative enumeration. Two advances have made this possible. The first advance is the development of DNA barcoding (Box 2). The second is that technology now allows high-throughput next generation sequencing (NGS) to be performed at a fraction of the cost that previously precluded advances in the field of ecological monitoring. NGS is a new technology that enables automated high-throughput DNA sequencing that can produce thousands or millions of DNA sequences at the same time.

Combining DNA barcoding with NGS as a rapid method for multiple species identification from a complex environmental sample is termed ‘metabarcoding’. This has been shown to have great potential when applied to the ecological assessment of diatoms (Kermarrec et al. 2014, Visco et al. 2015), as it has the potential to replace the labour-intensive stages of species identification.

Genetic markers used as DNA barcodes need to be specific for the target organism. Taxonomic resolution to discriminate at the species level is highly desirable; the marker should have a well understood pattern of molecular evolution and be ideally linked to a comprehensive taxonomic database. Numerous gene markers have been investigated as potential DNA barcode targets for diatom identification (Evans et al. 2007, Moniz and Kaczmarska 2009, Moniz and Kaczmarska 2010). These included:

- classical cytochrome c oxidase subunit 1 (COI) gene
- small ribosomal subunit (SSU)

- second ribosomal internal transcribed spacer (ITS) region together with 5.8S gene (ITS-2 + 5.8S)

Although SSU had the highest amplification success, it required a significantly longer fragment to be amplified and sequenced before species resolution was attained. COI showed substantial heterospecific divergence and was readily aligned, but its amplification efficacy was low, which was a potential limiting factor in its use. In contrast, the 300–400 base pair (bp) ITS-2 + 5.8S fragment provided a high success rate of amplification together with good species level resolution. A further supposed advantage of ITS2 (Moniz and Kaczmarska 2010) was that compensatory base changes in helix regions may give insights into the limits of biological species, since their presence is claimed to correlate with sexual incompatibility (Coleman 2009). However, this idea has not been supported by critical studies (Caisová et al. 2011). Following identification of the ITS-2 + 5.8S region as a suitable fragment for barcoding diatoms, Moniz and Kaczmarska (2009) then exploited this region to genotype 114 diatom species (Moniz and Kaczmarska 2010). The technique enabled the separation of morphologically defined species with a success rate of 99.5%.

Box 2: DNA barcoding

DNA barcoding is based on the principle that a defined DNA sequence can be used to represent a specific species. A DNA sequence of a specified marker gene becomes a unique 'tag' or 'DNA barcode' for a particular organism. A gene region that is commonly used in plants, for instance, is a gene region in the chloroplast called the ribulose-1,5-bisphosphate carboxylase/oxygenase large (rbcL) chain gene.

Fundamental to DNA barcoding is a sound knowledge base where the DNA sequence is anchored to a known species that has been identified using classical morphology (voucher specimen). Linking DNA sequences to known voucher specimens has benefitted in recent years from international DNA barcoding campaigns, though these have been largely restricted to animals and land plants. These campaigns have created large online reference databases that link species taxonomies to diagnostic DNA sequences such as the Barcode of Life Data System (www.boldsystems.org) and the International Nucleotide Sequence Database Collaboration (www.insdc.org). These DNA barcode reference databases can be augmented by user-created DNA databases for particular taxa, combining the skills of molecular biologists and traditional taxonomists. Figure 1.1 shows an example of the steps involved in creating the DNA barcode reference database for diatoms.

Creating these databases can be a resource intensive exercise. Development starts with a specimen either obtained from the field or from a specimen collection. In the laboratory, the specimen is cultured and the DNA extracted. The barcode region on the marker gene is isolated using an amplification process called polymerase chain reaction (PCR). A DNA sequencer is used to read the nucleotides – cytosine (C), guanine (G), thymine (T) and adenine (A) – along the barcode region. Once the DNA sequence has been determined, it can be added to the reference database along with images of the voucher specimen and other specimen metadata.

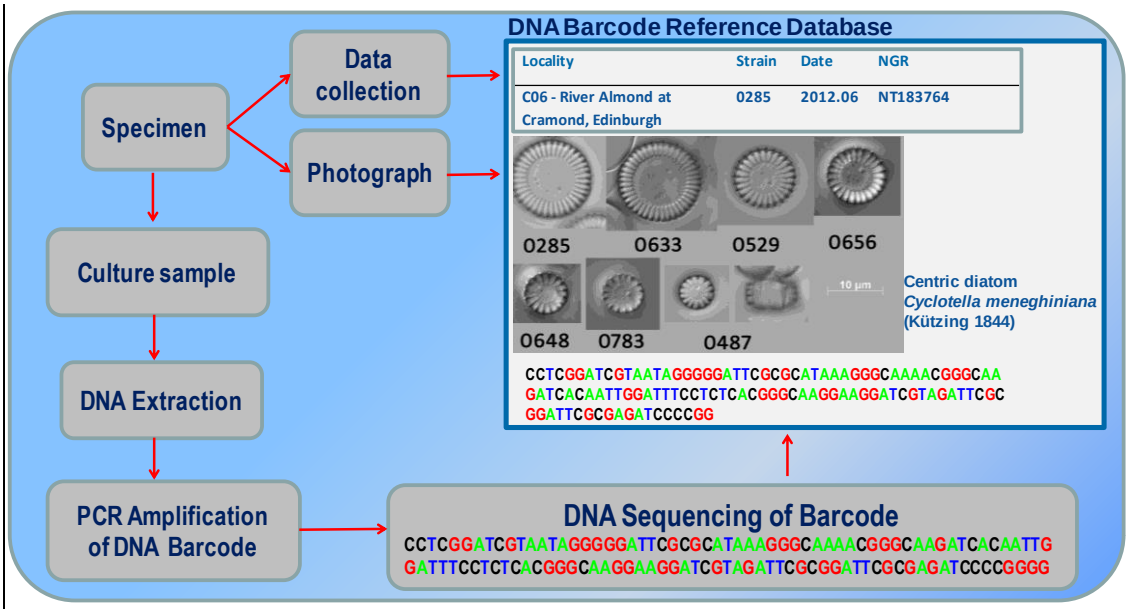


Figure 1.1 Steps involved in creating a DNA barcode reference database

Subsequent workers, however, have questioned the focus on ITS-2 (for example, because of the significant intra-individual heterogeneity in ITS), preferring to look either at the SSU (Zimmerman et al. 2011) or the rbcL) gene (Mann et al., 2010). Mann et al. (2010) argue that protein-encoding genes such as COI and rbcL pose fewer practical problems than rDNA, once they have been obtained. Benefits include that there is rarely any intragenomic variation and they are very easily aligned and compared. Sequencing errors can often be detected by frame shifts and unlikely amino acid changes such as exchange of one type of amino acid by a different one (for example, polar by non-polar, or basic by acidic). The rbcL gene, in particular, has been exploited for taxonomy (Trobajo et al. 2009) and ecological assessment (Kermarrec et al. 2014).

The chloroplast-based rbcL gene provides a very practical advantage over its nuclear SSU counterpart in the context of characterisation of real-world community analysis of water bodies related to targeting of the amplicon to chloroplast-containing ecosystem constituents. On the other hand, a number of environmental DNA (eDNA) studies have used SSU to describe the extensive complement of macro and micro fauna in rivers and lakes (Barnes et al. 2014, Liang and Keeley 2013). So although deployment of 18S would reduce the signal observed for the targeted diatom taxa, it could potentially open the way to integrated assessment of organism groups to provide more of a holistic overview.

1.3 About the project

This report describes the development of a DNA metabarcoding approach to ecological assessment based on diatoms using the NGS of a fragment of the rbcL gene. Although some have advocated abandoning traditional taxonomic approaches (Biomonitoring 2.0; Baird and Hajibabaei 2012, Woodward et al. 2013), this research tried to construct a molecular ‘mirror’ of the current approach based on LM. This ensures continuity with existing methods while, at the same time, complying with the normative definitions of the Water Framework Directive, which refer to ‘taxonomic composition’. While there is support for the claim by Baird and Hajibabaei (2012) that there is potential within DNA based approaches to explore aspects of diversity and ecosystem function that are difficult to measure using traditional approaches, it is still useful from a practical point of

view to understand the relationship between molecular evidence and traditional biological methods.

The twin foundations for this study are a calibration dataset of samples, analysed by both current LM and NGS approaches, along with a reference database of rbcL DNA barcodes which link to Linnaean taxonomy. The samples span a wide range of ecological quality encountered primarily in England, but also across other parts of the UK. They also provide a 'bridge' between current approaches to analysing and interpreting ecological quality using diatoms and new methods based on outputs from NGS.

1.3.1 Aims and objectives

The Environment Agency is looking to improve the efficiency and effectiveness of the way in which it carries out environmental monitoring. Fundamental to this are new ways of working, and using new and more effective approaches to ecological assessment. This project was developed in direct response to an initiative to identify recent developments in DNA-based methods that could potentially deliver novel, operationally valid monitoring approaches and at the same time provide efficiency savings and improvements in data quality within the Environment Agency's routine monitoring programme, focusing on the identification of diatoms followed by classification of the water body ecological status (Environment Agency 2011).

The overall aim of the project was to develop a high-throughput, cost-effective method for identifying and quantifying diatom taxa from environmental samples in a manner suitable for calculating the TDI and associated metrics using NGS for the Water Framework Directive. Although one objective was to develop a cost-effective method, a comparison of the costs and benefits are not presented within this report.

The work was conducted in 2 phases. Phase 1 was a proof of concept, an overview of which is presented in Appendix 1.

Specific objectives of the project were to:

- develop a reference database of rbcL DNA barcodes from known diatom species, isolated and cultured from water bodies of different ecological quality
- optimise DNA extraction and PCR protocols for the amplification of diatom DNA barcodes that will enable resolution of diatoms to an appropriate taxonomic level to enable TDI calculation using NGS
- optimise a bioinformatics pipeline for the routine analysis of diatom taxa from the NGS metabarcoding data
- perform a validation study comparing diatom species composition metrics acquired using NGS metabarcoding data with data produced using LM
- calibrate the estimation of TDI calculated from NGS metabarcoding data against matched samples analysed by LM
- quantify the performance characteristics of the work flow in terms of sources of uncertainty and variability compared with current LM in both the laboratory and the field

2 Development of diatom rbcL DNA barcode reference database

2.1 Introduction

Ecological assessments based on an examination of community structure require organisms present in a sample to be assigned to the appropriate Linnaean binomial so as to provide a link with autecological and habitat information for that species from which ecological quality can be inferred. For conventional microscope-based analyses, morphological criteria are matched by eye to descriptions in identification guides. For molecular analyses, the identification guide is replaced by a database of DNA barcodes of known provenance to which DNA sequences can be matched using bioinformatics algorithms.

As over 2,500 diatom species have been recorded in UK freshwaters (Whitton et al. 1998), effort was focused in this project on ensuring that those taxa most likely to influence the outcome of ecological assessments were included in the barcode database. Taxa were prioritised from an analysis of existing datasets.

As diatom assessments are based on a weighted average equation, the primary focus was on taxa that were both often abundant (defined as ≥10% of the total) and commonly encountered (that is, found in ≥10% of samples). Secondary considerations included whether the taxon was a good indicator of either high/good status or poor/bad status, and was not well represented in existing barcode libraries. A third category used in the primary screening was taxa closely related to those selected by the first 2 steps to ensure that the method could discriminate closely related species.

This screening exercise produced a list of taxa from which likely locations for obtaining them were identified, again using existing databases. As many as possible of these locations were visited and samples obtained provided the raw materials for culturing and isolation described below. Once barcodes had been obtained, permanent slides were made from the cultures and digital images collected to enable the taxa to be identified.

2.2 Methods

2.2.1 Isolation, culture and harvesting for DNA extraction and voucher preparation

Samples were collected as described in 2.1 from the locations listed in Appendix 3, and kept cool to avoid decay and deoxygenation. Within 1–3 days, samples were placed in 50mm Petri dishes, sometimes diluted with Woods Hole culture (WC) medium (Table 2.1). Individual cells of diatoms were isolated by micropipette or by streaking on 2–3% agar plates. Micropipette isolations were made with either a Zeiss inverted microscope or a stereomicroscope. With the inverted microscope, higher magnifications (of up to 400×) were possible and identifications to genus could often be made (from a combination of cell shape and chloroplast arrangement) but rarely to species, though in some cases even the genus could not be determined with any certainty.

Selected cells (or, in the case of plated material, discrete small colonies of clonal cells) were transferred into small volumes of freshwater medium in the wells of 96-well

plates. Initially a general purpose freshwater medium was used (WC medium with silicate, adjusted to pH 7) (Guillard and Lorenzen 1972). However, trials during the first couple of months indicated that this was unsuitable for diatoms from oligotrophic and acid habitats. For these, modified WC media were used containing less nitrogen (N) and/or phosphorus (P) (one-tenth of the usual WC additions) and modified Grundgloeodinium II medium (von Stosch and Fecher 1979), replacing the silicon dioxide (SiO_2) with the sodium metasilicate addition of WC medium. After a few days of incubation, the health and clonality of each culture was confirmed under an inverted microscope. Successfully established clonal cultures were then grown in 90mm Petri dishes for DNA extraction and preparation for a voucher slide. All the clones were grown at 15–22°C under cool white fluorescent light on a 14:10 (light: dark; L:D) photoperiod at a photon flux density of 5–20 µmol photons m^{-2} s^{-1}.

Cells were harvested by either pipetting (for species forming visible colonies, for example, Fragilaria and Staurosira) or scraping them from the bottom of the dish using pieces of silicone tubing (for benthic species, for example, Nitzschia and Navicula). The resulting slurries of cells were collected in 1.5ml test tubes and centrifuged at 2,000*g* for 10 minutes. Most of each pellet was transferred into a 1.5µl tube and kept at –20°C until DNA extraction, leaving a small amount which was resuspended with distilled water and dried onto one 18mm square coverslip and one 10mm diameter circular coverslip. The square coverslip was used to prepare a voucher slide for LM; the circular coverslip was retained in case of the need to examine material with scanning electron microscopy (SEM).

For both the LM and SEM vouchers, cells were cleaned in situ on cover slips by adding nitric acid to the cover slip on a hotplate and heating to oxidise organic material. After oxidation the diatom cell walls, still on the cover slips, were washed with distilled water several times to remove digestion products and then dried again on a hotplate. For LM, voucher cells were mounted in the high refractive index resin Naphrax, whereas SEM specimens were stored in Petri dishes at the Royal Botanical Gardens Edinburgh.

Table 2.1 Composition of algal growth media used in this study

Compound	Concentration (mg l^{-1})	Weight per litre of element	µM
WC medium (Guillard and Lorenzen 1972)1			
$CaCl_2.2H_2O$	36.76		250
$MgSO_4.7H_2O$	36.97		150
$NaHCO_3$	12.60		150
KH_2PO_4	8.71		50
$NaNO_3$	85.01		1,000
$Na_2SiO_3.9H_2O$	28.42		100
Trace metals			
Disodium EDTA	4.36	–	c. 11.7 (EDTA)
$FeCl_3.6H_2O$	3.15	0.65 mg Fe	c. 11.7
$CuSO_4.5H_2O$	0.01	2.5 µg Cu	c. 0.04
$ZnSO_4.7H_2O$	0.022	5.0 µg Zn	c. 0.08

Compound	Concentration (mg l^{-1})	Weight per litre of element	µM
$CoCl_2.6H_2O$	0.01	2.5 µg Co	c. 0.05
$MnCl_2.4H_2O$	0.18	0.05 mg Mn	c. 0.9
$NaMoO_4.2H_2O$	0.006	2.5 µg Mo	c. 0.03
H_3BO_3	1.0	0.17 mg B	c. 16
Vitamins			
Thiamin hydrochloride		0.1 mg l^{-1}	
Biotin		0.5 µg l^{-1}	
Cyanocobalamin (Vitamin B12)		0.5 µg l^{-1}	
$Na_2SiO_3.9H_2O$	28.42		100
Grundgloeodinium II medium (von Stosch and Fecher 1979)[2]			
KNO_3			500
Na_2HPO_4			10
$MgSO_4$			10
$CaCl_2$			1
$FeSO_4$			1
$Na_2SiO_3.9H_2O$			100
Disodium EDTA			2
Trace elements			As above

Notes: [1] Adjust pH to 6.5–8 with drops of concentrated hydrochloric acid. Stock solutions were prepared at 1,000× concentration and aliquots of 1ml added per litre of final medium. The medium was autoclaved at 120°C for 20 minutes. On standing, a fine brown precipitate often forms in autoclaved medium. This dissolves again with agitation and does not seem to harm cultures.
[2] Adjust pH to 5–7 with drops of concentrated hydrochloric acid. Stock solutions can be prepared at 1,000× concentration and aliquots of 1ml added per litre of final medium.

2.2.2 Imaging and identification of reference strains

Reference strains were photographed using a Zeiss Axio-imager photomicroscope using 100× or 63× oil immersion objectives (nominal NA 1.4) and either bright field or Nomarski interference contrast optics. All images are kept securely as TIFF files. Image metadata were recorded on associated .xml files, which are interpretable using Zeiss Axiovision software. Images were also listed with their microscope configurations in a Microsoft® Excel spreadsheet. Some image processing for montages was performed using Adobe Photoshop v.7 or CS2.

2.2.3 DNA extraction, PCR amplification of rbcL, sequencing and alignment

The enzyme ribulose-1,5-bisphosphate carboxylase (Rubisco) is responsible for carbon fixation. The rbcL gene encoding the large subunit of Rubisco is located in a single copy region of the chloroplast genome, of which there are multiple copies per cell. The rbcL gene provides conserved primer sites that have been shown to be appropriately conserved within the diatom phyla and allow effective amplification of a high proportion of species tested (Hamsher et al. 2011). Both the ~1,400 bp region of rbcL and a ~850bp region of the 3 prime end of rbcL (rbcL-3P; rbcL-3′) have been shown to have the power to discriminate between all species tested (Jones et al. 2005, Hamsher et al. 2011).

Extraction of DNA from each pellet was conducted using a high-throughput genomic DNA extraction instrument QIAxtractor (Qiagen). The forward and reverse primers used were the ones reported by Jones et al. (2005), that is, DPrbcL1: AAGGAGAAATHAATGTCT and DPrbcL7: AARCAACCTTGTGTAAGTCTC, which amplified a region of ~1,400 bp, covering the rbcL gene. The PCR reaction for the amplification of rbcL was in 25µl volumes containing 10ng DNA, 1 mM deoxynucleotides (dNTPs), 1× Roche diagnostics PCR reaction buffer (Roche Diagnostics GmbH, Mannheim, Germany), 1 unit Taq DNA polymerase (Roche) and 0.5 µM of each primer. The PCR cycling comprised an initial denaturing phase for 3 minutes (94ºC), followed by 30–40 cycles of 94°C for 1 minute, 55°C for 1 minute and 72°C for 1.5 minutes, with a final extension of 72°C for 5 minutes.

The quantity and length of the PCR products were examined by agarose gel electrophoresis against known standards. PCR products were purified using ExoSAP-IT (USB Corporation, Ohio, USA). Sequencing was conducted in 10µl volumes using 0.32 µM of PCR primer or sequencing primers NDrbcL5: CTCAACCATTYATGCG and DrbcL11: CTGTGTAACCCATWAC (Jones et al. 2005), 1µl of BigDye v3.1 and 2µl of sequencing reaction buffer (Applied Biosystems). Sequencing PCR conditions were 25 cycles of 95°C for 30 seconds, 50°C for 20 seconds and 60°C for 4 minutes. Excess dye-labelled nucleotides were removed using the Performa DTR V3 clean-up system (EdgeBio) and sequence products were run on an ABI 3730 DNA sequencer (Applied Biosystems) at the University of Edinburgh.

Sequencing reads were edited and assembled using SeqMan (DNASTAR, Madison, WI). Each rbcL region was sequenced by 4 reads (using primers DPrbcL1, DPrbcL7, NDrbcL5 and DrbcL11) and the whole region was sequenced by at least 2 overlapping reads.

The sequence was defined as ‘high quality’ if all the reads were obtained successfully and resulted in no ambiguous bases. ‘Low quality’ reads were those with at least one read having weak signal(s) and/or noise(s), so that not all the sequence region was covered by multiple overlapping reads.

Because rbcL is a translated protein (with almost no variation in sequence length), the gene sequences of different taxa were easily aligned manually in BioEdit 7.0.2 (Hall 1999).

2.2.4 Addition of externally validated barcodes

Until the work of Jones et al. (2005) introduced new primers, there were few rbcL sequences for diatoms available in GenBank® (www.ncbi.nlm.nih.gov/genbank/) and most of these were for planktonic species, for example, Aulacoseira and Thalassiosira. Since 2005, many more sequences have been deposited by a variety of laboratories, so that now there are >2,000 rbcL sequences in GenBank {Nucleotide search

(Bacillariophyta[Primary Organism]) AND rbcL [Gene Name}. However, some of these are of marine or planktonic taxa and are relevant to freshwater ecological assessment only through the phylogenetic context they provide for interpreting unknown NGS sequences. Others are poorly documented (through published images and metric data) for the identification to be trusted, or represent taxa known to need or be undergoing taxonomic revision.

Nevertheless, among the GenBank sequences there were significant numbers that could be added to the reference database for this project, especially the well-documented sequences deposited by Rimet and colleagues at the French National Institute for Agricultural Research (INRA) at Thonon-les-Bains (based on their 'TCC' culture collection) and a number of Nitzschia and Sellaphora sequences already obtained by the Royal Botanical Gardens Edinburgh.

As a prelude to developing the reference dataset, the entries on GenBank were evaluated based on the project's team knowledge of the expertise associated with submitting groups; only those from 'trusted' sources were included. Other external sequences were obtained through the curated, open access barcode database for diatoms at R-SYST (www.rsyst.inra.fr).

In many cases, the taxonomic classification used by GenBank was out-of-step with that currently accepted by diatomists and implemented for diatom analyses. As a consequence, the taxonomic hierarchy for any GenBank sequence would need to be annotated by hand before it was imported to the DNA barcode reference database.

Given the time it takes to check the provenance and documentation of sequences deposited in GenBank, GenBank sequences were evaluated and added only when these offered a closer match to NGS results than any sequences obtained during this study. In future, it may be mutually advantageous to reach agreements with other groups active in developing barcodes for ecological assessment (for example, the INRA group) to share unpublished, well-documented sequences. It will be important to regularly re-inspect external barcode sources for sequences closely related to the NGS molecular operational taxonomic units and import them to the DNA barcode reference database.

2.2.5 Addition of inferred barcodes

Some barcodes were assigned a species identity by inference. This worked by comparing their occurrence frequency between LM and NGS and their relative phylogenetic position using maximum likelihood (see Sections A1.2.3 and A1.2.4 in Appendix 1).

2.2.6 Addition of Xanthophyta (yellow-green algae) contaminants

The preliminary study identified Xanthophyta contaminants in environmental diatom samples that were influencing the proportional representation of diatom species. Xanthophyta sequences were incorporated into the barcode reference database to allow pre-filtering prior to NGS analysis (see Section A1.2.5 in Appendix 1).

2.3 Results

A total of 987 unialgal cultures were obtained from samples collected from 60 locations in England and Scotland. DNA was extracted and sequenced from 554 of these, representing 123 species from 41 genera (Appendix 4).

Multiple strains were sequenced from some genera (as many as 67 and 78 for *Fragilaria gracilis* and *Achnanthidium minutissimum*, both common 'pioneer' species from low nutrient environments). These, in turn, permit broader coverage of cryptic and semi-cryptic variation within species complexes that can be difficult to identify with certainty with LM alone.

In addition, the identities of 8 taxa were inferred directly from the congruence of unassigned NGS reads with LM results (Appendix 5). These included additional barcodes that clustered close to *Achnanthidium minutissimum* and *Eolimna minima*.

Finally, 45 strains were added from GenBank or R-SYST (Appendix 6); 307 sequences for Xanthophyta (yellow-green algae) were also added from GenBank so as to filter out close relatives of the diatoms that would otherwise cause problems during the bioinformatics (Appendix 7).

3 General methods

3.1 Diatom sample collection

Diatom samples were collected from UK rivers using standard Environment Agency sampling techniques for benthic diatoms. This involves placing 5 cobbles in a tray with about 50ml of stream water and then brushing the upper surface of each cobble with a toothbrush to remove the biofilm (Kelly et al. 1998, CEN 2014a). These samples were then transferred to the laboratory in a cool box. Using a Pasteur pipette, 5ml of the suspension of biofilm and water was transferred to a sterile 15ml centrifuge tube containing 5ml nucleic acid preservative (hereafter referred to as diatom preservative) consisting of 3.5 M ammonium sulphate, 17 mM sodium citrate and 13 mM ethylenediaminetetraacetic acid (EDTA) The sample was then frozen at -30^{o}C prior to extraction of the DNA. The remainder of the sample was preserved using Lugol's iodine for morphological analysis by LM (Appendix 2).

Preliminary experiments looked at the possibility of using alternative sampling methods, such as clinical swabs to collect samples, rather than toothbrushes. However, the yield of DNA from such samples was generally much lower than from toothbrush-collected samples, so the latter were retained as the preferred sampling instrument.

3.2 Preparation and analysis of diatoms by LM

Samples for LM were digested either with a mixture of sulphuric and oxalic acids, with potassium permanganate (Environment Agency laboratories) or cold hydrogen peroxide (CEN 2014b).

Following digestion, samples were rinsed several times to remove all traces of oxidising agents. Between rinses samples were either centrifuged at 3,000–5,000 rpm for 4–5 minutes (Environment Agency laboratories) or allowed to stand overnight to ensure that all diatoms settled to the bottom of the tube. Permanent slides were prepared using Naphrax (Brunel Microscopes, Chippenham) as a mountant, following Kelly et al. (2008). At least 300 valves on each slide were identified to the highest resolution possible using a Nikon BX40 microscope with 100× oil immersion objectives with phase contrast and their abundance recorded.

The primary floras and identification guides used were Krammer and Lange-Bertalot (1986, 1997, 2000, 2004), Hartley (1996) and Hofmann et al. (2011). All nomenclature was adjusted to that used by Whitton et al. (1998), which follows the conventions of Round et al. (1990) and Fourtanier and Kociolek (1999).

3.3 Preparation and analysis of diatoms for NGS

DNA was extracted using the enzymatic lysis method described in Appendix 9.

3.3.1 Target amplification

Amplification of rbcL prior to sequencing was carried out with the following method. PCR reactions of 30µl containing 6µl of HF buffer (NEB, USA), 0.3 µM forward and reverse primers (Table 4.1), 0.3 mM dNTPs, 0.3µl Phusion high-fidelity DNA

polymerase (NEB) and 0.5µl of a 1:10 dilution of extracted sample DNA. The final reaction volume was made up with nuclease-free water to 30µl.

The following PCR protocol was followed: amplification started with an initial single denaturation step at 98°C for 2 minutes, followed by 35 cycles of denaturation at 98°C for 20 seconds, annealing at 55°C for 45 seconds and extension at 72°C for 60 seconds, followed by a final extension at 72°C for 5 minutes. All PCR reactions were carried out without replication on a C1000 thermal cycler (Bio-Rad, UK); each run contained a number of negative controls including 'no template' controls, index PCR controls and extraction buffer controls that passed through the whole procedure.

PCR products were visualised on 1% agarose gels. They were then purified using AMPure Beads following the Illumina 16S Metagenomic Sequencing library preparation protocol and were eluted in 50µl nuclease-free water.

3.3.2 Index addition

In order to identify and remove sequences from previous runs (something that happens in small amounts when using MiSeq™ from Illumina even with improved decontamination procedures due to common fluidics that are not changed between runs), 3 sets of indices were used, changing the index set between runs. Experience shows that, after 3 runs, within instrument contamination is no longer detectable and the first index can then be reused. This results in indexes only being used every third MiSeq run, effectively removing the possibility of samples on subsequent runs containing sequences from the previous run.

Illumina Nextera XT sequencing adapters and indices were attached to each sample with a PCR step by combining 10µl HF buffer, 0.3 mM dNTPs, 1 µM $MgCl_2$, 0.5µl Phusion polymerase (NEB, USA), 5µl of each specific 'index 1' and 'index 2' primer, and 5µl of purified sample PCR product. The final reaction volume of 50µl per sample was made up with nuclease-free water.

The PCRs were carried out on a C1000 thermal cycler. Amplification cycling conditions were as follows: 95°C for 3 minutes, followed by 8 cycles of 95°C for 30 seconds, 55°C for 30 seconds and 72°C for 30 seconds, with a final extension of 72°C for 5 minutes. The PCR product was then purified with AMPure Beads following the Illumina 16S Metagenomic library preparation protocol. Final libraries were eluted in 25µl nuclease-free water.

The quality and quantity of each amplicon library was evaluated with Tapestation (Agilent, USA) along with quantification using Qubit (Life Technologies, CA, USA) prior to sequencing.

3.3.3 Illumina sequencing (MiSeq)

All samples, including controls, were quantified using the Qubit method. They were then combined to produce a 20 nM library, which was again quantified and diluted to produce a final 4 nM library for sequencing. Negative controls were water controls for both the PCR amplification and MiSeq library preparation steps. The positive control for PCR reactions was a mock community constructed from cultured extracts (described in more detail in Table 5.3). The library was denatured and combined with 5% PhiX sequencing control DNA and loaded onto a MiSeq instrument following the Illumina 16S Metagenomic Sequencing library preparation protocol.

4 Development of the short rbcL barcode

4.1 Introduction

During the course of the project, significant changes occurred in the availability and performance of NGS technologies. The 2 most important technologies of relevance to this project are GS FLX (Roche) and MiSeq™ (Illumina). The former platform was used initially due to the increased read length (up to 900 bp) compared with the 400 bp achievable using the latter platform (Appendix 1). A further consideration is cost and availability; while the GS FLX costs remained high, the MiSeq costs have fallen continuously, resulting in the GS FLX being withdrawn from sale in 2016. As a result, a short barcode was required of a length appropriate for sequencing on the MiSeq platform and which provided good taxonomic resolution.

The research to identify suitable primer binding sites and evaluate barcodes of differing lengths and positions enabled the most cost-effective sequencing technology available today to be accessed. It also had the extra advantage that, should new technologies provide the opportunity to use longer barcodes (for example, MinION, Oxford Nanopore), it may be possible to implement their use with minimal extra cost, given that almost full length rbcL sequences have been determined and are available in the barcode reference database.

4.2 Materials and methods

4.2.1 DNA extraction

Field samples were received in diatom preservative and stored at -30ºC until DNA extraction. Two DNA extraction methods were compared:

- the method of Fawley and Fawley (2004) combining homogenisation using glass beads with buffer containing dodecyltrimethylammonium bromide (DTAB) followed by Qiagen DNeasy® column purification using FastDNA buffers (MP-Biomedicals)
- the enzymatic lysis method of Eland et al. (2012), essentially 5 hours of incubation with Proteinase K, followed by column purification using Qiagen DNeasy® Blood and Tissue kit according to the manufacturer's instructions

The quantity of DNA was estimated using a Qubit fluorimeter and dsDNA BR Assay Kit following the manufacturer's instructions (Thermo Fisher Scientific, Cat: Q32850). Genomic DNA was stored at -30ºC prior to PCR and NGS analysis.

4.2.2 PCR amplification

Amplifications were performed in 20µl volumes containing 4µl of HF buffer, 0.3 µM of forward and reverse primers (Table 4.1), 0.3 mM of dNTPs, 0.4 units Phusion high-fidelity DNA polymerase (New England Biolabs, UK). The final reaction volume was made up with nuclease-free water (Severn Biotech, UK). All PCRs were carried out on a C1000 thermal cycler.

The PCR cycling conditions were one cycle of 98°C for 2 minutes, followed by 35 cycles of denaturation at 98°C for 20 seconds, annealing at temperatures ranging from 60 to 50°C for 45 seconds and extension at 72°C for 60 seconds, and a final extension at 72°C for 5 minutes.

The quantity and length of the PCR products were examined following electrophoresis on 1% agarose gels compared with DNA standards of known sizes, stained using ethidium bromide and visualised on an ultraviolet (UV) transilluminator.

Table 4.1 Sequences of primers used for amplifying rbcL barcodes

Primer name	Sequence (5′ to 3′)	Experiment	Reference
rbcL-39F	TGWCCGTTACGAATCTGGTG	Short barcode evaluation	This study
rbcL-404F	CWGCDTTACGTTTAGAAGATATGCG	Short barcode evaluation	This study
rbcL-404R	CGCATATCTTCTAAACGTAAHGCWG	Short barcode evaluation	This study
rbcL-646F[1]	ATGCGTTGGAGAGARCGTTTC	Short barcode evaluation	This study
rbcL-646R	GAAACGYTCTCTCCAACGCAT	Short barcode evaluation	This study
rbcL-998F	CAGTTGTWGGTAAATTAGAAGGTGATC	Short barcode evaluation	This study
rbcL-998R [1]	GATCACCTTCTAATTTACCWACAACTG	Short barcode evaluation	This study
rbcL-3P_640F [2]	CCRTTYATGCGTTGGAGAGA	Proof of concept (Appendix 1)	Hamsher et al. 2011
rbcL-3P_1538R [3]	AARCAACCTTGTGTAAGTCT	Proof of concept (Appendix 1)	Hamsher et al. 2011

Notes: [1] Primers used to amplify the short barcode for subsequent NGS.
[2] Formerly known as Cfd F
[3] Formerly known as DPrbcL7

4.2.3 Determination of conserved regions and primer design

To establish a short barcode from the rbcL gene, it was necessary to identify regions of diverse (informative) sequence flanked by regions of low diversity sequence where primers could be designed to amplify the barcode region from a large number of diatom species.

A total of 390 diatom sequences from the rbcL barcode reference database were used to develop a short rbcL barcode suitable for high-throughput NGS analysis. The diatom sequences were aligned using MAFFT (Katoh and Stanley 2013) using default settings. The diatom alignments were analysed using primer design software currently under development (https://github.com/rachelglover/diatom-analysis).

The settings applied to identify conserved regions of the alignment were 96% similarity with a maximum of 4 gaps in the alignment at that position. A sliding window of 25 nucleotides and a threshold of 5% of an alignment column differing from the most prevalent base were used to identify degenerate bases. Primers were designed to the identified regions using Primer3 (Undergasser et al. 2013) with default settings. When multiple primers were identified for a region, the best individual primer was selected based on the lowest number of degenerate nucleotides and the highest percentage sequence identity for that primer against the original diatom alignment.

4.2.4 Estimation of the resolving power of the short rbcL barcode

Potential rbcL barcode regions were independently assessed for taxonomic coverage, using the following protocol in QIIME (Quantitative Insights into Microbial Ecology) v1.5 (Caporaso et al. 2010). Firstly, operational taxonomic units (OTUs) were picked with UCLUST (Edgar 2010) from all the sequences in that alignment region with a similarity level set at 100% in order to create distinct OTUs from identical sequences. A representative sequence for each OTU was then selected. The OTU representative sequences were then assigned taxonomy using BLAST® (Basic Local Alignment Search Tool; http://blast.ncbi.nlm.nih.gov/Blast.cgi) (Zhang et al. 2000) based on the diatom reference database.

The raw OTU counts and taxonomic assignments for each OTU were then used to calculate the number of sequences in the region that had been assigned to the correct taxonomic level for the sequence used in the alignment (which has known taxonomy). This processing step was carried out using a custom script (processOTUs.py; python code for taxonomic assignments), (Appendix 8). The counts for each region were plotted using the statistical computing package R v3.0.2 (www.r-project.org).

4.2.5 Testing primer amplification

Using DNA extracted from *Tabellaria* sp. (Culture Collection of Algae and Protozoa, number 1081/7) as the PCR template, the performance of the different primer sets was compared experimentally. Criteria for comparison were the amplification of fragments of the correct length, with no amplification of secondary bands.

The robustness of amplification was assessed by comparing results following amplification at different annealing temperatures.

4.3 Results

4.3.1 DNA extraction

The performance of the 2 methods was tested on environmental diatom samples. The results (Table 4.2, Figure 4.1) show that the Proteinase K method gave higher average and more consistent amounts of purified DNA. A further consideration was that the Proteinase K method could be applied to high-throughput extraction using robotics, while the DTAB method was lengthy and complex to complete. The Proteinase K

method (Appendix 9) was therefore used to extract DNA from all diatom samples and optimised for use on a robotic DNA extraction system (BioRobot, Qiagen).

Table 4.2 Average, minimum and maximum amounts of DNA purified from 8 diatom samples

	DTAB (ng per µl)	Proteinase K (ng per µl)
Average	14.8	20.3
Maximum	54.6	32.8
Minimum	1.22	10.4

Notes: Samples were vortexed and split into 2 prior to extraction using the 2 methods.

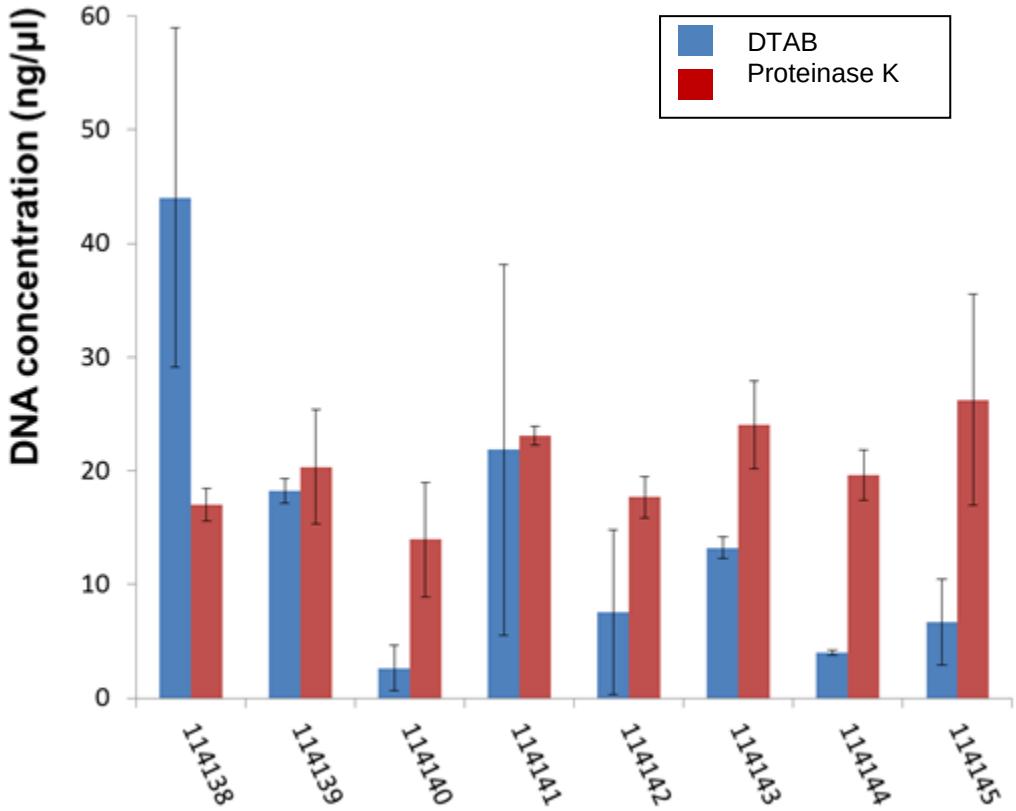


Figure 4.1 DNA concentrations from the extracts of 8 diatom samples

Notes: Samples were vortexed and split into 2 prior to extraction using the 2 methods.

4.3.2 Determination of conserved regions and primer design

A total of 11 regions along the rbcL gene were identified as having >96% sequence identity suitable for primer design (Figure 4.2, Table 4.3). A small number of the regions identified were immediately adjacent to each other and for primer design were considered to be one region only.

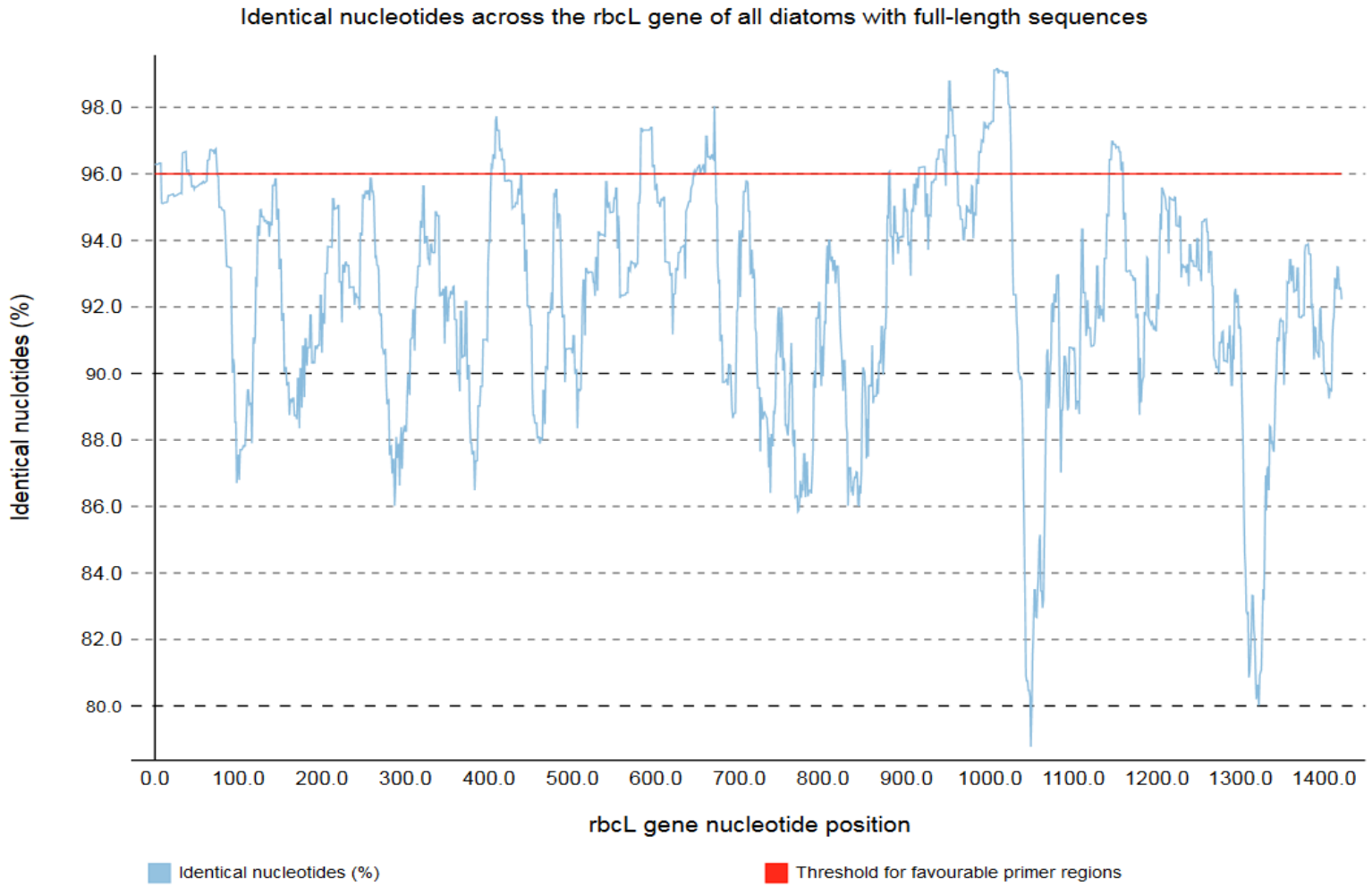


Figure 4.2 Percentage of identical nucleotides plotted along the length of an alignment of full length diatom rbcL sequences

Notes: The red line is a threshold used by the software to select regions which are most suitable for primer design.
In this alignment, 11 rbcL regions were suitable for conserved primer design.

Table 4.3 Location of regions identified as suitable for primer design for Illumina amplicon sequencing

Alignment location	Approximate sequence conservation	Sequence (5′–3′)
0–32	96.3%	ATGTCTCAATCTGTAWCAGAACGGACTCGAAT
33–65	96.6%	AAAAGTGACCGTTACGAATCTGGTGTAATYCC
63–99	96.7%	CCWTAYGCTAAAATGGGTTACTGGGATGCTKCATAY
403–443	96.7%	CWGCDTTACGTTTAGAAGATATGCGTATTCCWCAYTCWTA
582–623	97.3%	GAAGGTTTAAAAGGTGGTTTAGAYTTCTTAAAAGATGAYGA
645–696	96.3%	ATGCGTTGGAGAGARCGTTTCTTAWACTGTATRGAAGSTATY AACCGTGCW
879–905	96.0%	TTACAYTTACAYCGTGCDGGTAACTC
915–947	96.5%	CGTCAAAARAAYCAYGGTATYAAYTTCCGTGT
936–986	97.0%	AAYTTCCGTGTWATYTGTAAATGGATGCGTATGKCWGGTGT WGAYCAYAT
987–1,050	96-99%	CAYGCWGGTACAGTTGTWGGTAAATTAGAAGGTGATCCTTT AATGATTAAAGGTTTCTAYGA
1,143–1,184	96.4%	TCWGGTGGTATYCAYTGTGGTCAAATGCACCAATTAVTWCA

Primers were designed (Table 4.1) to amplify 4 regions along the rbcL gene that showed good potential for species discrimination. The location on the rbcL gene of the 4 hypothetical amplicons (I-L), along with the already validated longer amplicon (B) used by Hamsher et al. (2011) and tested in Appendix 1, are shown in Figure 4.3. The amplicons varied in size from 213 bp (L) to 344 bp (I).

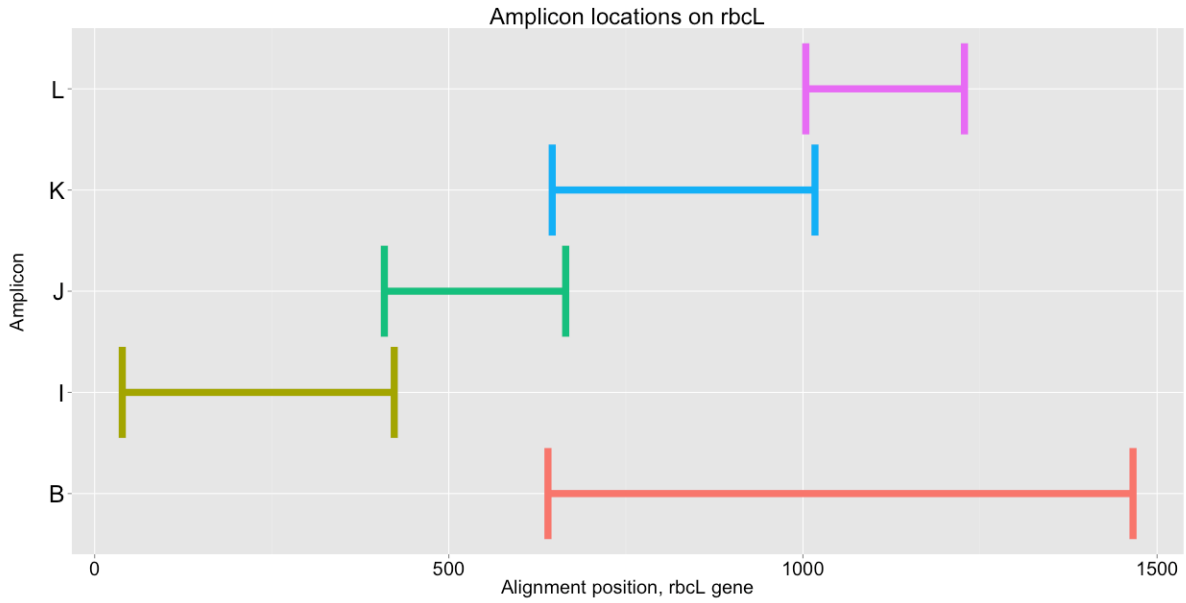


Figure 4.3 Locations of each hypothetical amplicon region (fragment) along the length of the rbcL gene

4.3.3 Estimation of the resolving power of the short rbcL barcode

The number of sequences that could be correctly assigned to class, family, genus, species and isolate were calculated for each amplicon (Table 4.4). For example, a count of 1 for the taxonomic level 'class' means that 'class' was the lowest taxonomic level at which an accurate taxonomic classification could be made for that sequence using this amplicon. In this way, it was possible to use the sum of the 'species' and 'isolate' counts as an assessment of the efficacy of a particular amplicon to be used for species level assignments.

Table 4.4 Amplicons assessed for their ability to place sequences to species level identifications

Amplicon	Forward primer	Reverse primer	Length (bp)	Lowest taxonomic level where the taxonomic assignment was correct					
				Class	Family	Genus	Species	Isolate	No identification
B	rbcL-3P_640F	rbcL-3P_1538R	786	2	0	12	177	199	0
I	rbcL-39F	rbcL-404R	344	2	0	21	204	156	7
J	rbcL-404F	rbcL-646R	216	2	0	37	202	142	7
K	rbcL-646F	rbcL-998R	331	2	0	22	201	165	0
L	rbcL-998F	rbcL-3P-1229R	213	2	0	26	202	151	9

Notes: 'No identification' is due to missing sequence coverage in that region for those sequences.

From the taxonomic assignments in Table 4.4, all amplicon regions could be used to provide a respectable number of species level assignments for the 390 sequences present in the original dataset. But because diatom metric TDI estimation is based on species level discrimination, the number of correct species level identifications was plotted against the length of the fragment (Figure 4.4).

As expected, the longest fragment (B) produced the largest number of correct species level identifications. However, it is unsuitable for Illumina sequencing, given the requirement for a good overlap between the paired end reads to maintain quality (and therefore accurate species level identification).

Amplicons J and L, while suitably short, appear to flank sequence that is too conserved and cannot be used to provide an adequate number of correct species level identifications in comparison to the longer fragment.

Amplicons I and K are both of an appropriate length (344 and 331 bp respectively) to give accurate sequences using the Illumina MiSeq platform. They can also be used to provide a satisfactory number of correct species level identifications. The forward primer for amplicon K also spans a region of the rbcL gene, which is 99% conserved across all 390 sequences.

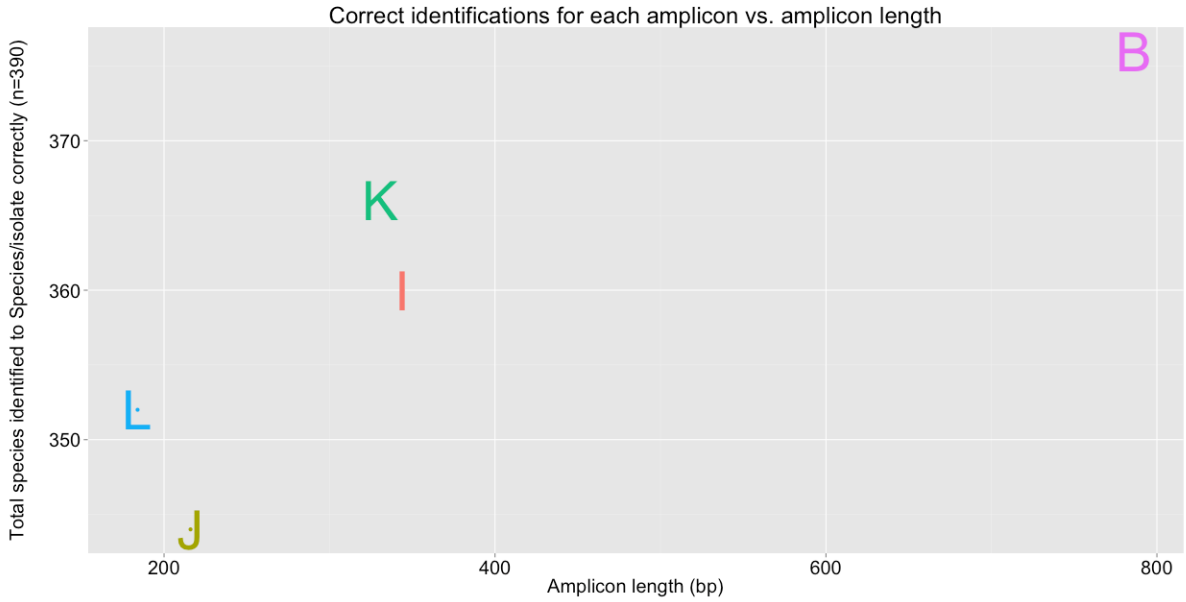


Figure 4.4 Correct species level taxonomic assignments plotted against the length of the amplicon fragment

Notes: The centre of the text is the exact point plotted.

4.3.4 PCR amplification of the short barcode

Overall, the primer pair rbcL-646F/rbcL-998R (amplicon K) gave the best performance, consistently giving an intense band of the correct size across the full range of annealing temperatures tested (Figure 4.5). The other primer sets performed less well; rbcL-39F/rbcL-404R (amplicon I) gave miss-priming at temperatures below 58°C, while primer pairs rbcL-404F/rbcL-646R and rbcL-998F/rbcL-3P-1229R (amplicons J and L respectively) amplified DNA less efficiently giving faint bands at all temperatures tested.

Based on its taxonomic coverage, amplicon length, primer conservation and robust performance, amplicon K (331 bp) was selected for use in all downstream Illumina analyses for benthic diatoms.

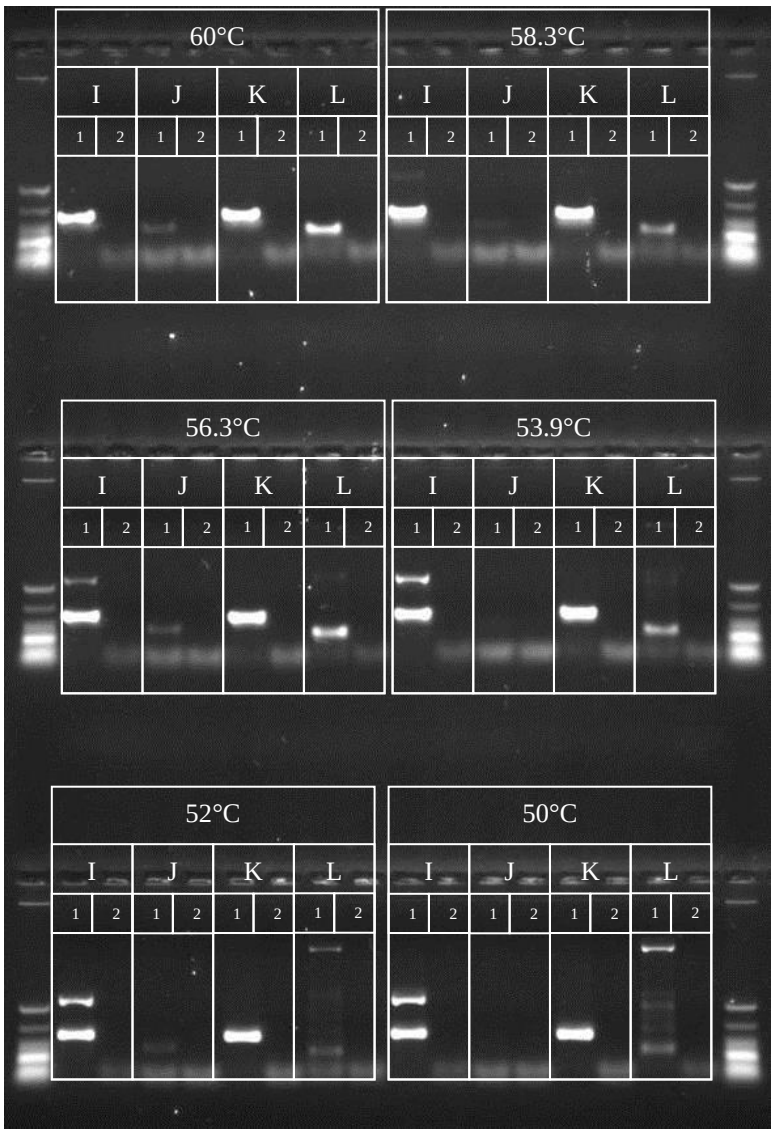


Figure 4.5 Gel electrophoresis of PCR products post amplification performed at different annealing temperatures (between 50 and 60°C) using newly designed primer sets (I, J, K and L) tested on DNA from a diatom sample and a no template control

Notes: The diatom sample is (1) and the no template control is (2) in each pair of tracks. The PCR products are flanked on the gel by low molecular weight markers (New England Biolabs, UK).

5 Development of NGS workflow and data analysis

5.1 Introduction

An NGS workflow based on the use of PROMpT (Primary Rapid Overview of Metagenomic Taxonomy) software (https://github.com/passdan/prompt) was developed during the early developmental phase of this project (Appendix 1). However, because the preferred operating system for PROMpT is Biolinux, this limits its utility, especially since government agencies are unable to install and run Biolinux due to Public Services Network restrictions. As a result the project changed to the QIIME platform (www.qiime.org), which is considered the industry gold standard. It can be scaled up to the data produced by larger NGS platforms such as Illumina's MiSeq and HiSeq, and has the potential to incorporate a high-throughput pipeline that would automate the analysis.

5.2 Bioinformatic analysis

The data from each instrument run was analysed independently to mitigate against any intra- and inter-run variation that may have been introduced during PCR or library preparation. A mock community sample, extraction and PCR controls for each run were analysed alongside the samples in the respective run (Section 3.3). The analysis pipeline developed is in 2 parts: quality control and taxonomic assignment (Figure 5.1).

5.2.1 Quality control

Any errors incorporated into the DNA sequences generated – even single nucleotide polymorphisms – have the potential to create additional taxa (false positives) in the downstream analysis. As a result a very stringent quality control procedure was implemented consisting of the following 4 steps:

1. Removal of PCR amplification primers from both sequenced strands of DNA using Cutadapt v1.9.1 (Martin 2011)
2. Sliding window trimming of poor quality 3′ ends of sequences from both strands (this is a typical Illumina artefact) was achieved using Sickle v1.33 (Joshi and Fass 2011) in paired end mode
3. Joining of the trimmed, paired end reads to form one consensus strand using PEAR v0.9.6 (Zhang et al. 2014)
4. Further round of quality assessment for the removal of any sequences with an overall accuracy of less than 99.9% using Sickle v1.33 (Joshi and Fass 2011) in single-read mode

Following quality control, each sample was independently prepared for analysis using QIIME and the taxonomic assignment pipeline described in the next section applied.

5.2.2 Taxonomic assignment

The downstream analysis was completed in 4 main steps as follows:

OTU picking

Taxonomic assignment of each individual sequence in the dataset would be very computationally intensive. As a result, OTU picking is used to make the number of sequences requiring taxonomic assignment much smaller.

Because there are varying levels of intra-species variation in most genes used for amplicon metabarcoding studies, the first step is to cluster sequences into OTUs which are then used for downstream analysis. The diatom pipeline uses UCLUST (Edgar 2010) within QIIME to carry out this step and clusters the sequences based on 97% similarity. The similarity percentage was a mid-range value chosen partly because it is the same used in bacterial and fungal studies, and partly because the nucleotide identities between most invertebrate species range between 95% and 99%. As the identities between species can vary by genus, it is difficult to pick a de novo clustering value that will accurately cluster all species separately as individual OTUs.

OTU representative sequence selection

A representative sequence from the OTU cluster must be chosen for downstream analysis. The diatom pipeline uses the most abundant unique sequence in the cluster for this purpose.

Assigning taxonomy to OTUs

Once a representative sequence is available for each OTU it can be used to assign taxonomy to the sequences within the OTU cluster defined above. This is carried out with the QIIME package using BLASTn to search against the diatom reference database (see Section 2) for sequences with >90% sequence identity (now amended to 95%, see Section 5.3).

If a sequence match is found in the database, the OTU is assigned the taxonomy of the sequence with the highest identity.

Reporting of assignments and abundance

QIIME is used to calculate the abundance of each species. In the OTU picking stage, the total sequences in each OTU cluster are retained. These values are then summed by species (as each OTU will have a taxonomy assignment), giving the total number of sequences per species. A percentage of the overall sample is then calculated in order to report the relative abundance for each species present in the sample.

All species detected are reported, even in low abundance, as no minimum abundance threshold value is set. Potential false positives, including chimeras, are likely to occur at a rate of less than 2%, which is the threshold used to calculate the TDI values. During testing of the pipeline, low numbers of Xanthophyta contaminants were identified from each sample in correlation with that reported in Appendix 1. It was decided that, as QIIME reports the relative abundances of the algae present as well as diatoms (due to their inclusion in the diatom reference database constructed in the original PROMpT software, Appendix 1), they would be reported in the final pipeline but with the option remaining to remove them during the analysis in the future. Xanthophyta contaminants were observed in very low quantities, suggesting that the PCR primers used for amplifying the new short fragment either did not amplify Xanthophyta efficiently or Xanthophyta were not present in those samples.

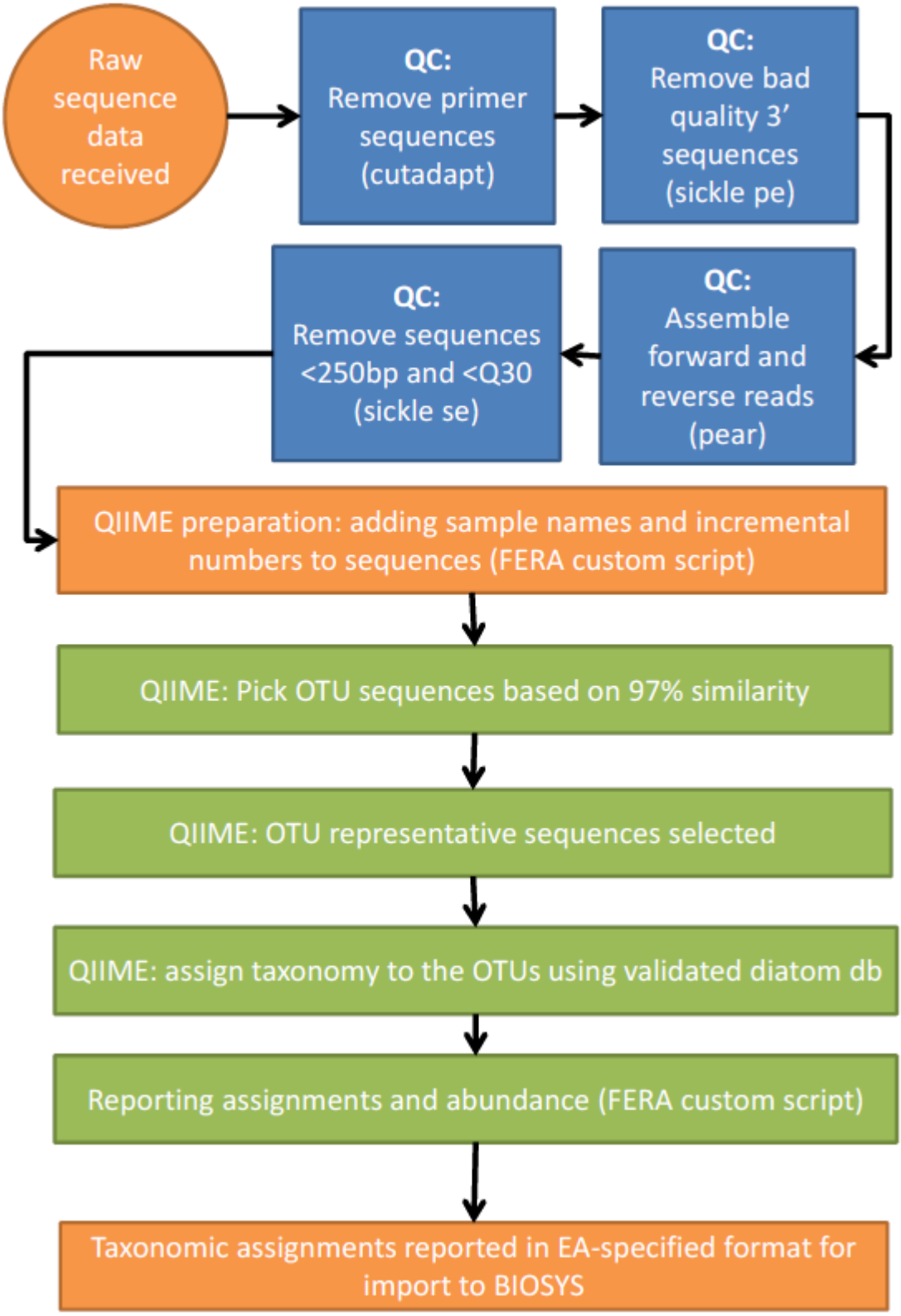


Figure 5.1 Quality control and QIIME pipeline for analysis of diatom NGS data

Notes: db = database; EA = Environment Agency; FERA = Food and Environment Research Agency; pe = paired end; QC = quality control; se = single end; BIOSYS = EA database for storing, manipulating and reporting data from freshwater and marine biological surveys

5.3 Validation

The NGS procedure developed was assessed for robustness by estimating its reproducibility, repeatability, sensitivity and specificity. The assessment of reproducibility and repeatability were completed using field samples. Sensitivity was estimated using a mock community constructed from cultured diatoms with a decreasing amount of one species. Specificity was estimated using a mock community constructed from cultured diatoms.

5.3.1 Reproducibility and repeatability

Four field samples (114061, 114078, 114092 and 114161) were used for the reproducibility and repeatability experiments. Inter-individual reproducibility was tested by 2 different staff members carrying out the PCR and clean-up steps of the sequencing protocol on all 4 samples. Each sample was amplified in triplicate to test the repeatability of amplification from the same DNA extract. To test for inter-instruments reproducibility, the sequencing was completed with the same library preparation split between 2 MiSeq instruments: one at the Food and Environment Research Agency in York and one at NewGene Ltd in Newcastle.

The data from both sequencing runs were passed through the quality control pipeline as described above, and sequences which passed quality control were prepared for further analysis using QIIME. OTUs were constructed by clustering with UCLUST at 97% nucleotide similarity, and the most abundant sequence was chosen as the representative sequence for each cluster. A Biological Observation Matrix (BIOM) table (www.biom-format.org) was constructed to store the individual OTU composition of each sample.

To statistically assess the differences between different groupings of samples, a distance matrix of beta (inter-sample) diversity was calculated using the Bray–Curtis dissimilarity metric. The Bray–Curtis matrix was used for each of the reproducibility and repeatability experiments. Two statistical methods, ANOSIM and adonis (Anderson 2001), were used to assess the variance between the OTU composition of the 4 field samples (totalling 56 sequencing samples) for each experiment.

The statistics in Table 5.1 can be used to draw a number of conclusions about the reproducibility experiments. The low R^2 values from adonis and low R values from ANOSIM, paired with the very high *p* values, lead to the conclusion that there are no significant differences between the samples when split by staff member and by different MiSeq instrument. In contrast, when the same test is applied to split the samples themselves as a control, the R and R^2 values are high and the differences are significant ($p = 0.001$).

Table 5.1 Inter-individual and inter-machine reproducibility statistics, as tested using adonis and ANOSIM

Experiment	adonis result, R^2 (*p* value)	ANOSIM result, R (*p* value)
Inter-individual reproducibility	0.00539 (0.994)	0.01786 (0.774)
Inter-machine reproducibility	0.00405 (0.997)	0.02586 (0.969)
Control	0.79659 (0.001)	0.97838 (0.001)

Notes: A control (a sample containing a large diatom diversity) is included to show the difference in variation between diatom samples when calculated using the same methods

Table 5.2 shows the data from the same adonis and ANOSIM analysis applied to the replicates in each of the diatom samples, split by staff member. The low number of PCR replicates (n = 3) is affecting the ability of the statistical methods to detect and measure differences. The data are also plotted in a stacked bar chart to visually represent the different taxa identified in the replicates (Figure 5.2).

Table 5.2 Differences detected between 3 replicates of each PCR carried out for each sample, split by staff member

Diatom sample	Staff member E		Staff member I	
	adonis result R^2 (p value)	ANOSIM result R (p value)	adonis result R^2 (p value)	ANOSIM result R (p value)
114061	0.49478 (0.667)	–	0.75697 (0.333)	–
114078	0.28906 (1.000)	–	0.75519 (0.167)	–
114092	0.63066 (0.167)	–	0.65475 (0.333)	–
114161	0.29529 (0.833)	–	0.55793 (0.333)	–

Notes: ANOSIM was unable to detect any differences due to the low number of PCR replicates (n = 3) and the extremely high similarity.

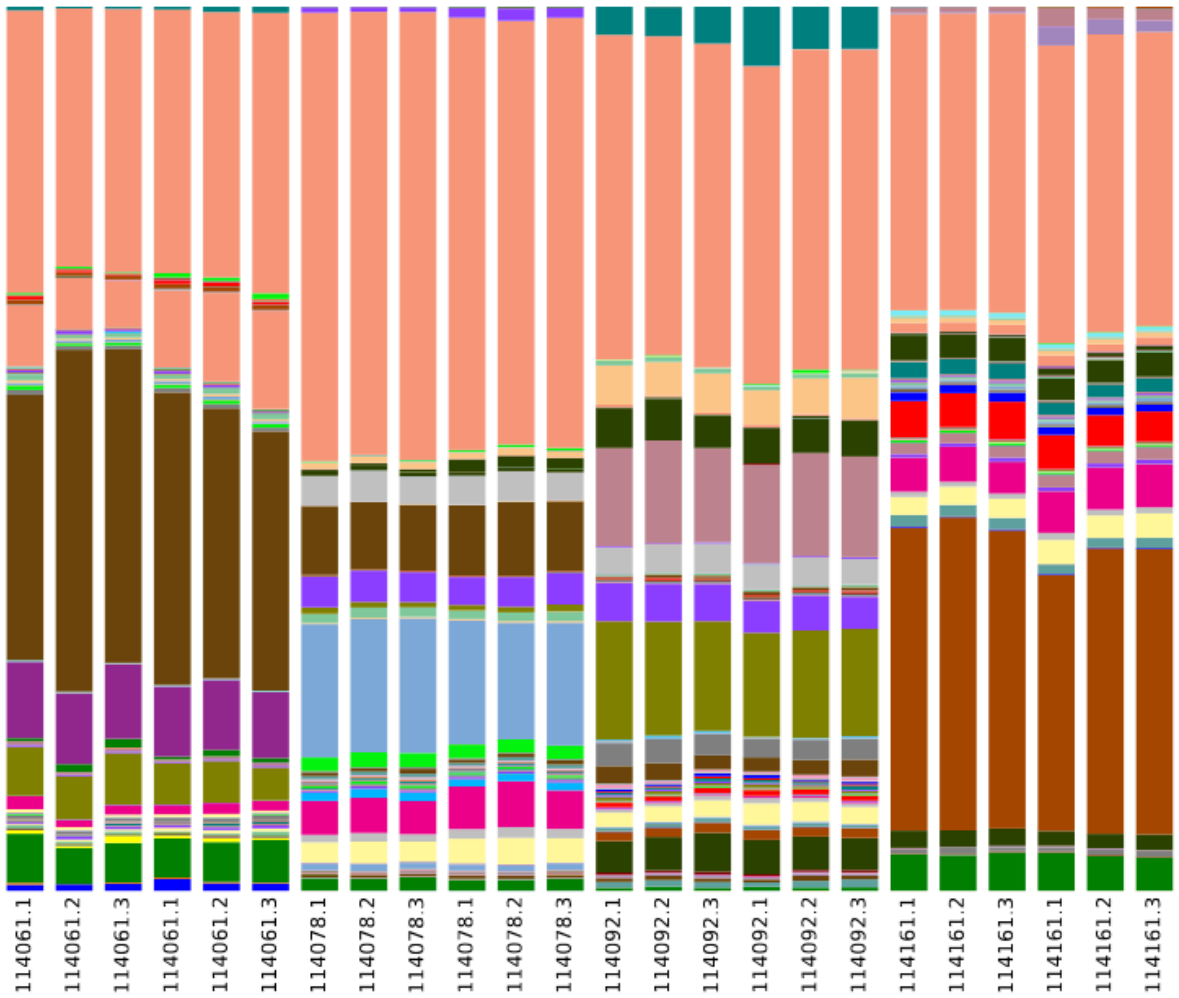


Figure 5.2 Stacked bar chart showing each of the 4 samples with 6 PCR replicates

Notes: The colour-blocks in each bar represent single species and the relative proportion in the sample.
Very little variation is seen between the 6 replicates of each sample.

Summary result

No significant differences were detected between staff members, PCR replicates or separate sequencing instruments when the same diatom samples were processed.

5.3.2 Sensitivity

To test the sensitivity of the protocol, a mock community (from extracted DNA) was constructed containing each of the 11 species listed in Table 5.3 to provide a background of diatom DNA. The species *Gomphonema parvulum* was added in different dilutions (1:10 to 1:1,000,000) to the background mock community. Each of the mock community samples were taken through the PCR, sequencing and pipeline protocol, and the results are shown in Figure 5.3. The relative abundance of *G. parvulum* is seen to drop in response to dilution within the mock community. Changes in relative abundance are also observed in response to the reduction of *G. parvulum* in the 1:10 and 1:100 dilutions.

In order to initially check the identity of each culture used in the sensitivity experiment, DNA was extracted from each culture separately and the rbcL gene (amplicon K) was

amplified. Sanger sequencing was performed on each culture to ensure the correct identification of the cultures prior to the mock community construction. Table 5.3. shows the results of the sanger sequencing for each culture and details the revised names of the species used in the mock-community and the following discussion.

Table 5.3 Cultured species obtained from culture collections, indicating their references and revised identify following Sanger sequencing

Mock community species	Culture collection	Culture collection ID	Revised identification following Sanger sequencing
Melosira nummuloides	Bigelow	CCMP482	*Melosira nummuloides*
Cyclotella cryptica	CCAP	CCAP 1070/6	*Cyclotella meneghiniana*
Eucocconeis sp.	Bigelow	CCMP2525	*Nitzschia inconspicua* (98% identity match)
Stephanodiscus hantzschii	CCAP	CCAP 1079/4	*Cyclotella cryptica*
Tabellaria sp.	CCAP	CCAP 1081/7	*Tabellaria flocculosa*
Asterionella formosa	CCAP	CCAP 1005/7	*Asterionella formosa*
Fragilaria crotonensis	SAG Goettingen	28.96	*Fragilaria crotonensis* and *Fragilaria bidens* (99% identity match)
Gomphonema parvulum	SAG Goettingen	1032-1	*Gomphonema parvulum*
Navicula pelliculosa [1]	SAG Goettingen	1050-3	*Mayamaea permitis* (97% identity match)
Nitzschia palea	SAG Goettingen	1052-3a	*Nitzschia palea*
Sellaphora capitata	Ugent	*Sellaphora capitata* D.G. Mann and S. Droop (03x38) F1-9	*Sellaphora capitata*

Notes: [1] Identification by morphology of these small naviculoid diatoms is problematic and *Navicula pelliculosa* has long been known to be a widely misapplied name.
CCAP = Culture Collection of Algae and Protozoa (www.ccap.ac.uk)

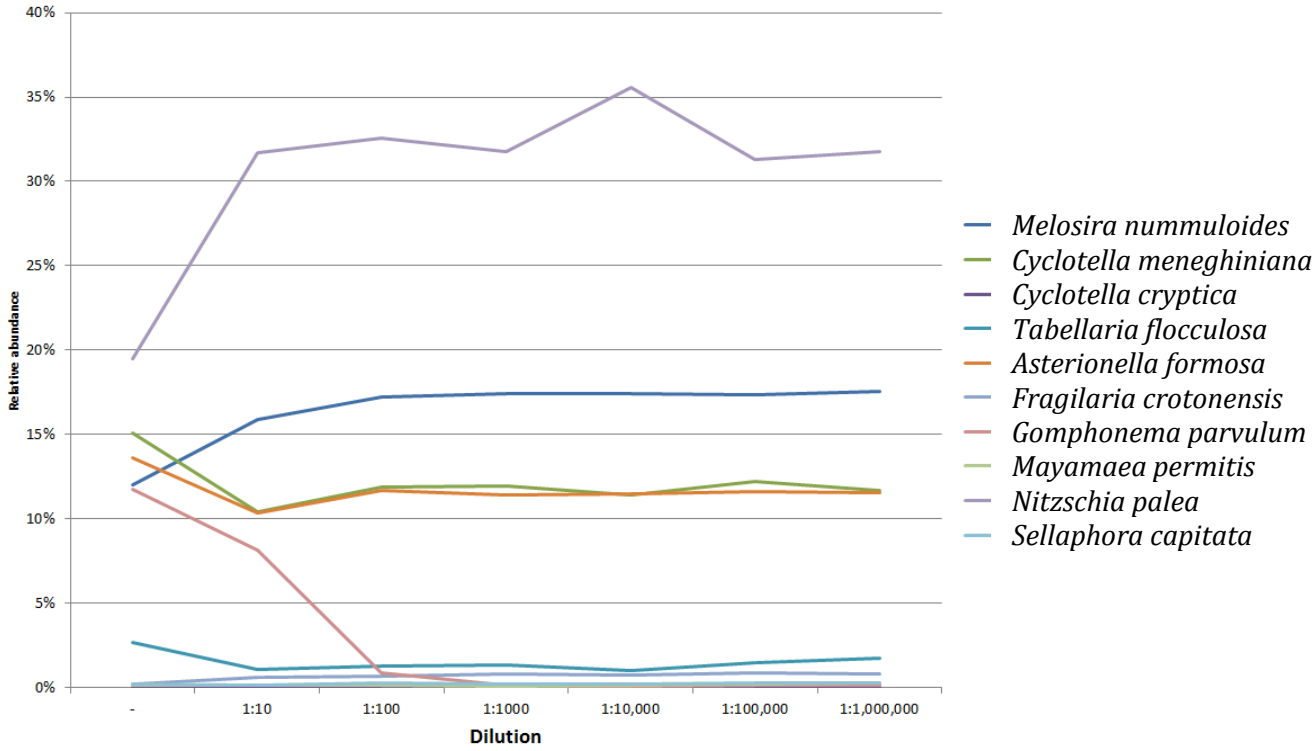


Figure 5.3 Relative abundance of each species in the mock community

Notes: *Gomphonema parvulum* (pink) is observed in reducing relative abundance within each community sample as its input amount is reduced by the serial dilution. As the relative abundance of *G. parvulum* decreases, the relative abundance of other species in the mock community increases and stabilises.

5.3.3 Specificity

To assess the specificity of the taxonomic assignments being made by the pipeline, the neat mock community sample from the sensitivity experiments was analysed in depth (Figure 5.4). As described earlier, the taxonomic assignments were made using the most abundant sequence from each OTU cluster following clustering at 97% nucleotide similarity. Each of the representative sequences were searched against the reference diatom database using BLASTn and the sequence with the highest BLAST score was used to assign taxonomy to that OTU. The BLAST step of the pipeline was carried out independently on the representative sequences and the percentage similarity to the best diatom match in the database was recorded and imported into R 3.1.1 for further analysis.

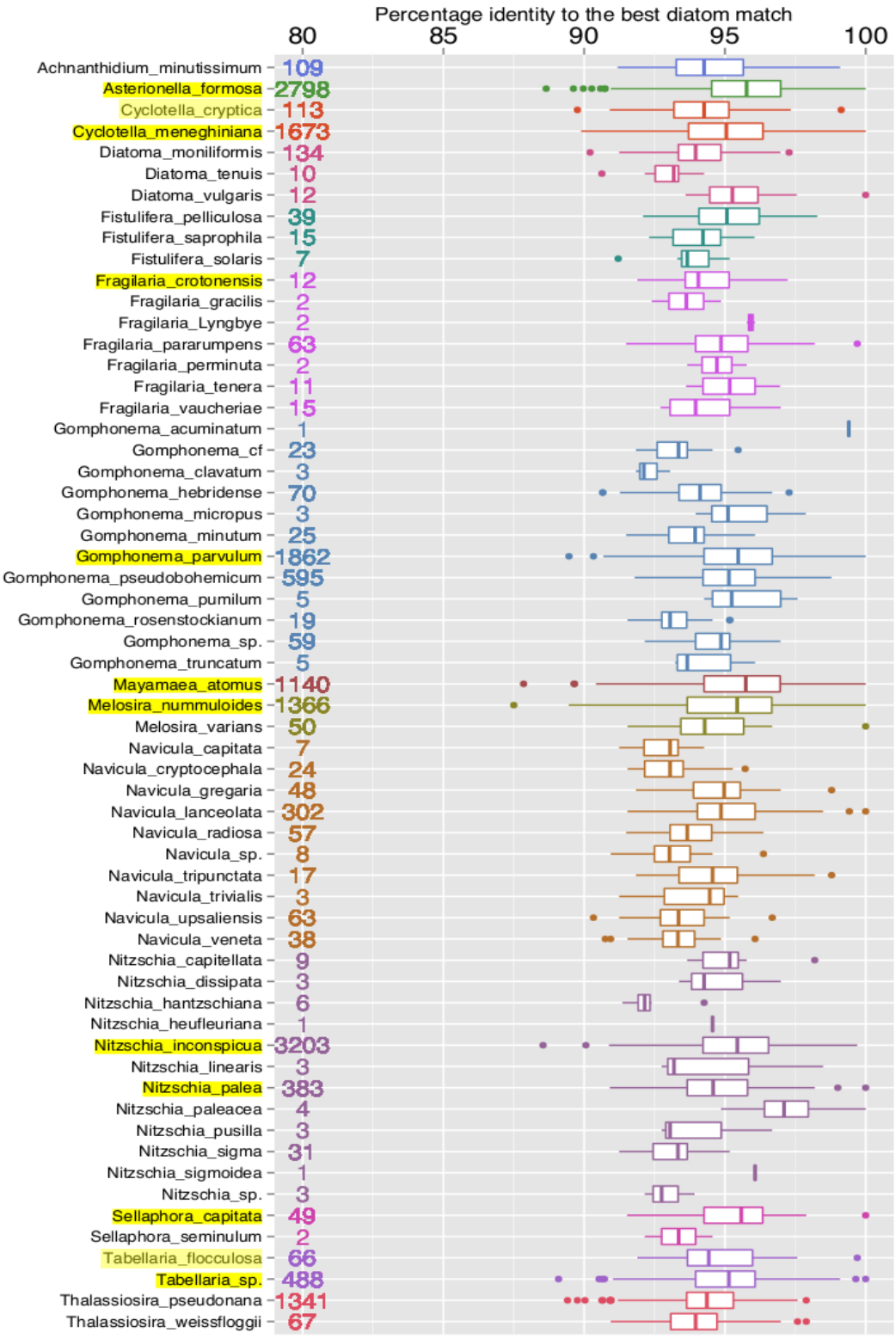


Figure 5.4 Box and whisker plots for all species detected in the mock community sample, showing the number of OTUs assigned to the species (right of the name) and boxplots showing the percentage similarities of all the representative sequences to the best match in the database that resulted in assignment to the species

Notes: [1] Different colours are used (for the number of OTUs/box plots) to represent different genera.
[2] Particular taxa of interest in the analysis are highlighted in yellow.

The results of the mock community analysis (Figure 5.4) show that, of the 11 species included in the mock community (Table 5.3, as identified by the Sanger sequencing of DNA extracted from the cultures – final column), six were identified accurately with high numbers of OTUs (*Asterionella formosa, Cyclotella meneghiniana, Gomphonema parvulum, Melosira nummuloides, Nitzschia inconspicua, Nitzschia palea* with reads of 2798, 1673, 1862, 1366, 3203, 383 respectively). Four species (*Cytotella cryptica, Sellaphora capitata*, *Fragillaria crotnensis and Tabellaria flocculosa*) were also identified in the analysis, but with a low number of OTUs (113, 49, 12 and 66 respectively) assigned to them. *Cyclotella cryptica* and *Cyclotella meneghiniana* share 99% sequence identity between their rbcL barcodes. Since the OTUs are clustered with 97% similarity, the numbers of OTUs assigned to each species may not be accurate. Similalrly a large number of OTUs (488) were assigned to Tabellaria sp. which may belong to *Tabellaria flocculosa*. In particular, the *Cyclotella cryptica / meneghinaina* example highlights some of the challenges faced in this work. The same clone can produce both '*Cyclotella cryptica*' and '*Cyclotella meneghiniana*' morphologies, depending on environmental conditions (Schultz 1971), indicating the problems associated with assigning binomials to barcodes in situations where the underlying taxonomy is still not fully resolved.

The 11th and final species included in the community, *Mayamaea permitis*, was not identified. There were, however, 1,140 OTUs identified as *Mayamaea atomus and* the rbcL barcodes of these 2 species share 97% identity. In addition, *M. atomus* appears to be designated *M. atomus var. permitis* in GenBank records, suggesting some uncertainty in species designation in GenBank. Of the remaining OTUs assigned in a significant number, 595 sequences were identified as *Gomphonema pseudobohemicum*; this species was not knowingly included in the mock community, but may have been a contaminant in the cultures used. Another 567 OTUs were identified as *Navicula* spp., which may have been contaminants introduced with the *M. permitis* which was found by Sanger sequencing to be the predominant species in the *Navicula pelliculosa* culture (Table 5.3).

Although many species in the mock community sample were identified effectively by the large number of OTUs assigned to those species, the analysis in general overestimated the number of taxa present in the sample as well as misidentifying some of the species that are present. However, these represent small relative abundances within the sample when the number of sequences per OTU is investigated and are therefore unlikely to affect the TDI value.

Figure 5.5 shows a brief exploration of one of the potential reasons for this with 36 diatom samples sequenced during this project. The pipeline takes advantage of QIIME's BLAST identification script, which has a (currently) unchangeable threshold for deciding when an identification has been made: if the BLAST hit in the diatom database has at least 90% identity with the OTU being searched then an identification is made. This is less than ideal as OTUs with distant hits could be assigned incorrect taxonomy, rather than left as 'unknown' or investigated further as potential new species. Figure 5.5 demonstrates that, as the identity threshold for BLAST hits is increased, the percentage of the sample given an identification decreases. In the current pipeline, unidentified sequences are searched against GenBank in order to broaden the search for a more accurate identification; however, this is not without risk as the GenBank database contains misidentified sequences and the taxonomy for diatoms is not updated. Currently, with the 90% threshold, very few OTUs are searched against GenBank. A better threshold could be 95%: while only 40% of OTUs would be given an identification, though this would still represent 75% of the sequences in the sample. By increasing the BLAST threshold, there is potential for 25% of the sequences in samples to be left without an identification and deemed 'unknown'; however, the identifications applied to sequences should be more taxonomically

accurate, ultimately leading to a more accurate read across between the LM and NGS methods.

Further work is required to refine the approach for assigning sequences to taxa, a problem that falls into 2 parts. Firstly identifying OTUs by clustering sequences with a strict cut-off (in this case 97%) may not be the most appropriate analysis approach for identifying species. Some species share a sequence identity higher than the cut-off and thus multiple taxa may be combined within a single OTU cluster. Other species are diverse with a larger amount of within species sequence variability; the sequences for these species may be split across multiple clusters. Secondly, it is known that, even when used in conjunction with GenBank, the taxon dictionary massively underrepresents the numbers of species found in the samples. As a result, using a relatively unconstrained criteria (>90% identity) to assign OTUs to taxa may compound the misidentification of the OTUs, as shown in Figure 5.5. Going forward, the current analysis pipeline will be amended to restrict species identifications to those with >95% sequence similarity to the reference database in order to reduce the potential for misidentifications.

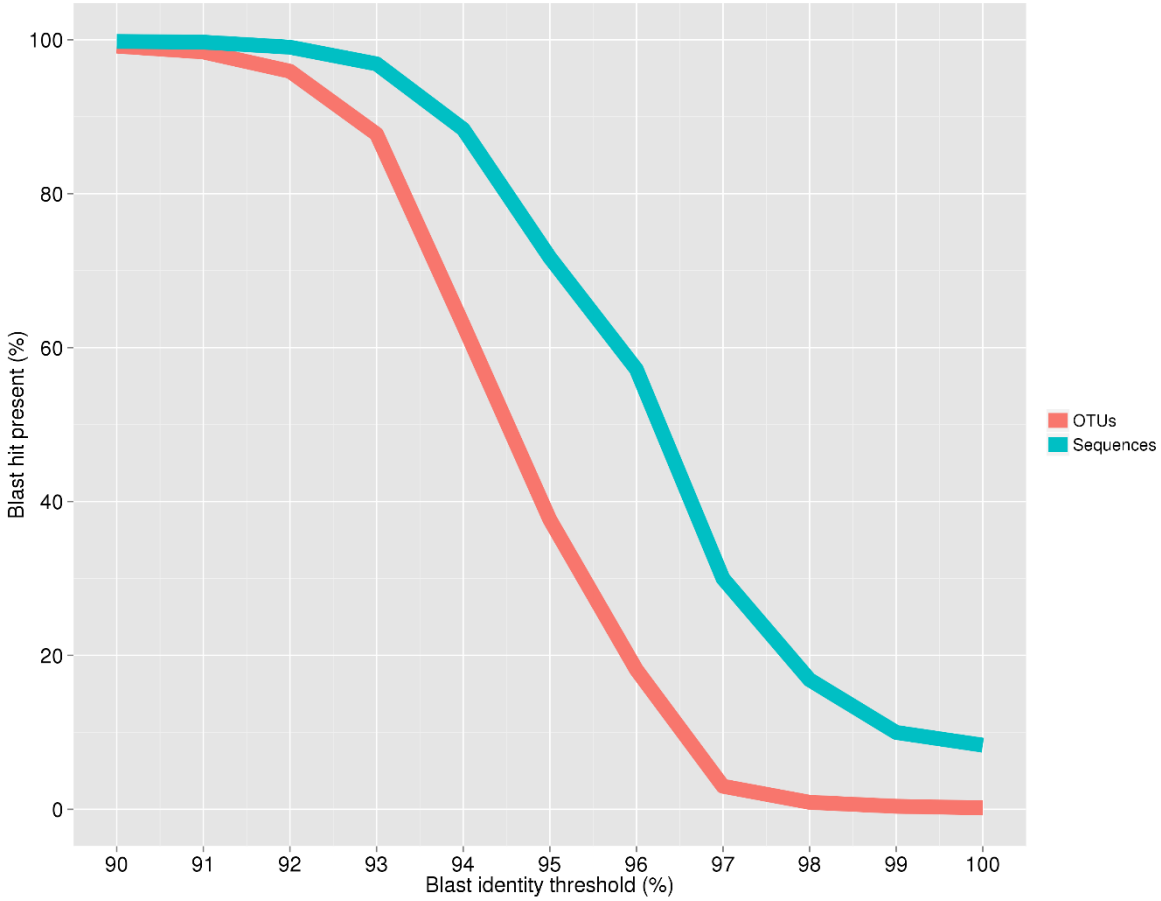


Figure 5.5 Number of OTUs (red) and overall proportion of sequences in samples (blue) having a hit in the diatom database within increasing BLAST identity threshold

Notes: As the threshold is increased, less OTUs/sequences are given an identification. This figure was created from further assessment of 36 diatom samples previously tested during the project.

6 Development and calibration of NGS metric

6.1 Introduction

Section 5 shows that it is possible to use Illumina NGS technology to process short rbcL barcodes from field samples to yield quantitative data. In theory, both LM and NGS approaches yield equivalent data (that is, a list of taxonomic categories, with the abundance of each expressed as a proportion of the total). In practice, however, the 2 approaches count different entities: LM records diatom valves (that is, half of a frustule or complete cell wall), while NGS records rbcL genes. Because rbcL genes are part of the chloroplast rather than the nuclear genome, and because the number of chloroplasts varies between genera, the relationship between LM and NGS data cannot be assumed to be 1:1. This, in turn, would be a potential source of bias if LM methods for data processing were applied to NGS data.

This section therefore begins by examining the relationship between LM and NGS data, before going on to constructing a modification of the existing TDI based method for estimating ecological status.

6.2 Methods

6.2.1 Study design

Development of a metric compatible with Water Framework Directive requirements requires the observed state of a water body to be compared with the reference state (that is, the ecological conditions encountered when anthropogenic disturbance is absent or minimal). Therefore, 2 separate datasets (with both LM and NGS analyses for each sample) were compiled:

- **Calibration dataset** – spans a range of ecological conditions along the primary nutrient/organic gradient to which diatoms are known to be particularly sensitive
- **Reference dataset** – consists only of samples from 'reference sites' (that is, locations where anthropogenic disturbances are absent or minimal)

Calibration dataset

Samples were collected from all of the approximately 1,000 sites scheduled for routine diatom sampling in England during 2014. A subsample of 250 sites were selected to provide as broad a range as possible of Water Framework Directive phosphorus status classes across the range of alkalinity types from low to high alkalinity. This was done because phosphorus concentration alone is insufficient to indicate the degree of pressure over the full alkalinity gradient; phosphorus concentrations are naturally higher in high alkalinity rivers than in low alkalinity rivers (Appendix 10).

Sites were ultimately selected by placing all sites within a matrix categorised by their alkalinity (1–9, 10–19, 20–29 mg $CaCO_3$ per litre and so on) against the 5 Water Framework Directive phosphorus status classes (where 1 = poor and 6 = high). A number of sites were then selected at random from each matrix category depending on

how many sites occurred within that category. Where between 1 and 3 sites occurred in the matrix category, 1 site was selected at random to be used in the validation analysis. For every subsequent 3 sites occurring within the category, a further site was selected at random. This led to around 230 sites being selected. The additional 20 sites were then randomly selected from the matrix categories that contained the highest number of sites within them to bring the total number of sites up to 250. As diatom samples are collected in both spring and autumn, this meant that a total of 500 samples were available for analysis.

Reference dataset

As there are very few reference sites in England, that is, following ECOSTAT criteria (Pardo et al. 2012), samples for this dataset were also collected from Scotland, Wales and Northern Ireland. A total of 232 samples from 113 sites identified as reference or near reference in Environment Agency (2013) were included in this exercise.

6.2.2 Statistical analysis

Non-metric multidimensional scaling (NMDS) (McCune and Grace 2002) was used to investigate the structure of the LM and NGS datasets using the R software package (R Development Core Team 2017) with the vegan package (Oksanen et al. 2007) for multivariate analyses. The aim of NMDS is to produce a low dimensional representation of the dissimilarity between samples, measured across all taxa. The success of NMDS is given by the stress, which quantifies the agreement between the (in our case) two-dimensional (2D) representation and original dissimilarities with (McCune and Grace 2002):

- values <0.1 representing a good ordination from which inferences may be drawn
- values of 0.1–0.2 representing an ordination that is useable with caution
- values >0.3 indicating that the ordination may be misleading

The similarity in structure between the LM and NGS ordinations was tested using a Procrustes analysis and associated permutation test (Peres-Neto and Jackson 2001) in the vegan package, and by scatterplots and computation of the Pearson correlation coefficient.

Calculation of TDI4 values used DARLEQ2 software (www.wfduk.org/resources/category/biological-standard-methods-201). A NGS specific variant of the TDI (TDI5) was derived using weighted averaging to calculate new NGS taxon indicator scores that gave the optimal prediction of LM-TDI4 values for the matched LM/NGS dataset (ter Braak and Barendregt 1996, ter Braak and Looman 1996). That is:

$$NGS_j = \sum_{i=1}^{n} y_{ij} * TDI4_i \Big/ \sum_{i=1}^{n} y_{ij} \qquad 6.1$$

where NGS_j is the NGS indicator score for taxon *j*, $TDI4_i$ is the LM-TDI4 value of sample *i*, y_{ij} is the relative abundance of taxon *j* in sample *i*, and *n* is the total number of samples in the dataset.

When calculating taxon and sample scores using weighted averaging, the range of scores is shrunk with respect to the original values. To correct for this effect, it is standard practice to 'deshrink' the scores using a linear regression of original on weighted averaging predicted sites scores (Birks et al. 1990). For this study, this

regression was applied to the NGS taxon coefficients so that the new TDI5 scores would be deshrunk to the correct range of values. Any taxa with indicator values <1.0 had their values set to 1.0. Weighted averaging calculations were performed in R using the package rioja (Juggins 2015). The indicators values derived in this way are listed in Table 6.1. They can be used to derive TDI5 sample values using the following equations:

$$TDI5\_initial_i = \sum_{j=1}^{m} y_{ij} * NGS_j \Big/ \sum_{j=1}^{m} y_{ij} \quad 6.2$$

$$TDI5 = (TDI5\_initial * 25) - 25 \quad 6.3$$

where $TDI5\_initial_i$ is the new initial NGS derived TDI score for sample *i*, NGS_j is the NGS indicator value for taxon *j*, and *m* is the number of taxa in sample *i*.

The 2 variants of the TDI (TDI4 and TDI5) were compared using Lin's concordance correlation coefficient (Lin 1989). This is a modification of correlation analysis which assesses the deviation from a perfect 1:1 relationship between the 2 variables. It was calculated by means of the epiR package (Stevenson 2010) within R.

EQR values were computed for each sample using expected values (eTDI) derived from alkalinity data from the closest appropriate site. Initial comparisons used site-specific alkalinity applied to 2014 diatom data only. This enabled a direct comparison of EQRs based on LM and NGS data for each site. However, classifications are not necessarily based on a single year's data and this needs to be borne in mind when comparing the NGS based classifications with the formal classification results (Table 6.2). EQR was calculated as:

EQR = (100 – observed TDI) / (100 – expected TDI) 6.4

6.3 Results

6.3.1 Dataset composition

The calibration and reference datasets were combined for the analyses of differences between LM and NGS in order to encompass the widest possible range of habitat conditions, from soft water upland sites in near pristine conditions to highly enriched lowland streams. A total of 628 samples passed NGS quality control, and had both LM and NGS data available for analysis.

Composition, as analysed by LM and NGS, was broadly similar with *Achnanthidium minutissimum* type having the highest maximum relative abundance (RA) in both methods (Figure 6.1) and being the most frequently recorded (Figure 6.2). There were, however, a number of differences in details. *Melosira varians*, for example, was both more frequently recorded and occurred at higher RA in NGS than LM samples, while the opposite was true for *Platessa conspicua*. In some cases, differences may represent gaps in the barcode reference database (that is, *Achnanthidium pyrenaicum, Gomphonema calcifugum*); however, others are harder to explain. For example, *Luticola ventricosa* and *Lemnicola hungarica* occasionally occurred in high numbers in the NGS outputs but are unlikely to be missed by LM analysts. A discrepancy also occurred for the genera Fistulifera and Mayamaea*;* in both cases, the maximum abundance recorded was higher in LM than in NGS, though the number of records was much higher with NGS. This issue is discussed in more detail below.

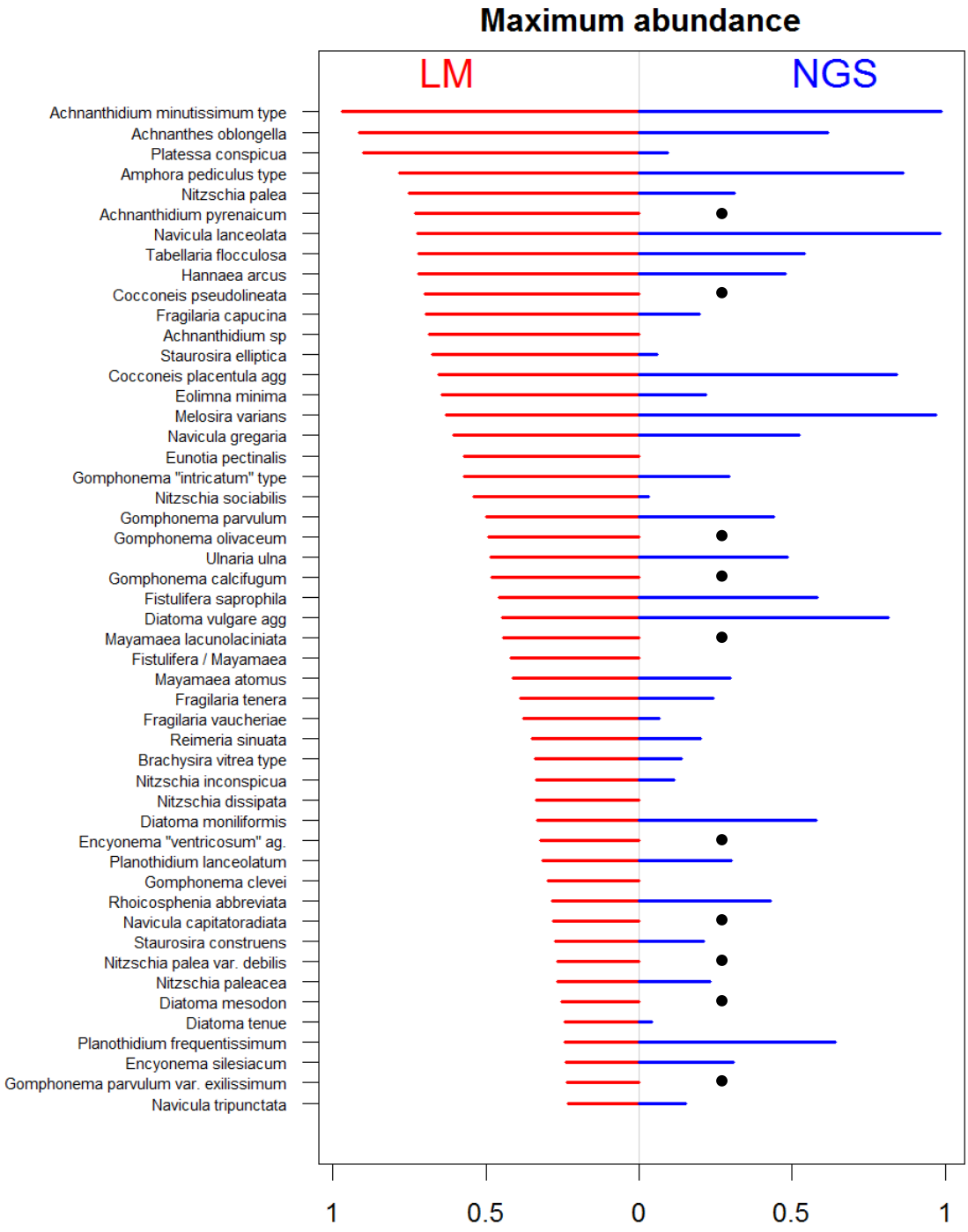


Figure 6.1 Differences in maximum abundance of the 50 most common diatom taxa in 628 samples as recorded by LM to show comparison with NGS data

Notes: ● Barcode for taxa absent from database.
A value of 0.5 means this is the maximum RA at which the taxon in question was recorded.

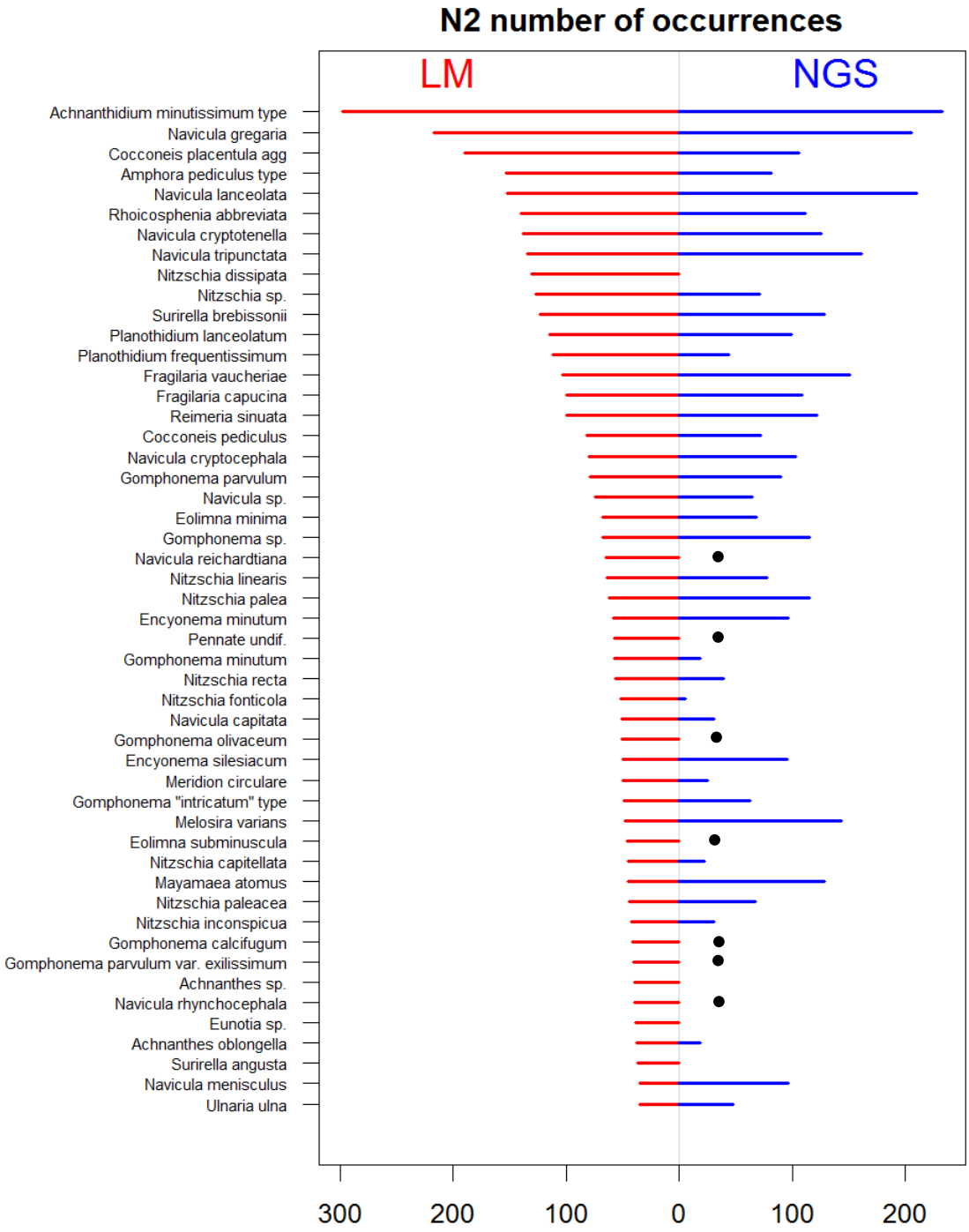


Figure 6.2 Differences in the total number of times that a taxon was recorded for the 50 most frequently occurring diatom taxa in samples as recorded by LM compared to NGS in the 628 sample dataset.

Notes: ● Barcode of taxa absent from database.

Figures 6.1 and 6.2 illustrate 2 separate problems faced in the development of a workable NGS method. The former illustrates fundamental differences in the units counted by the 2 methods (diatom valves for LM, rbcL sequences for NGS), while the latter highlights the ability of the barcode database to detect the full range of variation as understood by current morphology-based taxonomy.

The former issue means that there was rarely 1:1 correspondence between the proportions of individual taxa in LM and NGS. The general tendency was for small,

single-celled species such as *Achnanthidium minutissimum* and *Amphora pediculus* to have lower representation in NGS than LM, while larger cells with 2 chloroplasts (for example, *Navicula lanceolata*) or many chloroplasts (for example, *Melosira varians*) typically had greater representation in NGS compared with LM (Figure 6.3). There was considerable scatter in all the relationships between LM and NGS for individual taxa, reflecting uncertainty in both axes associated with the calculation of proportions of single taxa from a pool of many taxa. So while species A might generally form a higher proportion of the total in LM compared with NGS, this effect might be masked if species A co-exists with species B, which forms a much greater proportion in NGS than in LM. In practice, there are upwards of 20 taxa per sample, all of which will have an individual response, along with components of stochastic and analytical variability.

Particular issues were encountered for the genera *Fistulifera* and *Mayamaea*, both of which are far more prominent in many NGS reads but are absent from corresponding LM analyses (Figure 6.3). Representatives of these genera are tiny (<10µm) with weakly silicified frustules that may not survive the preparation process used in LM. When they were recorded in LM, they were often abundant (Figure 6.1), but there were many fewer records than for NGS.

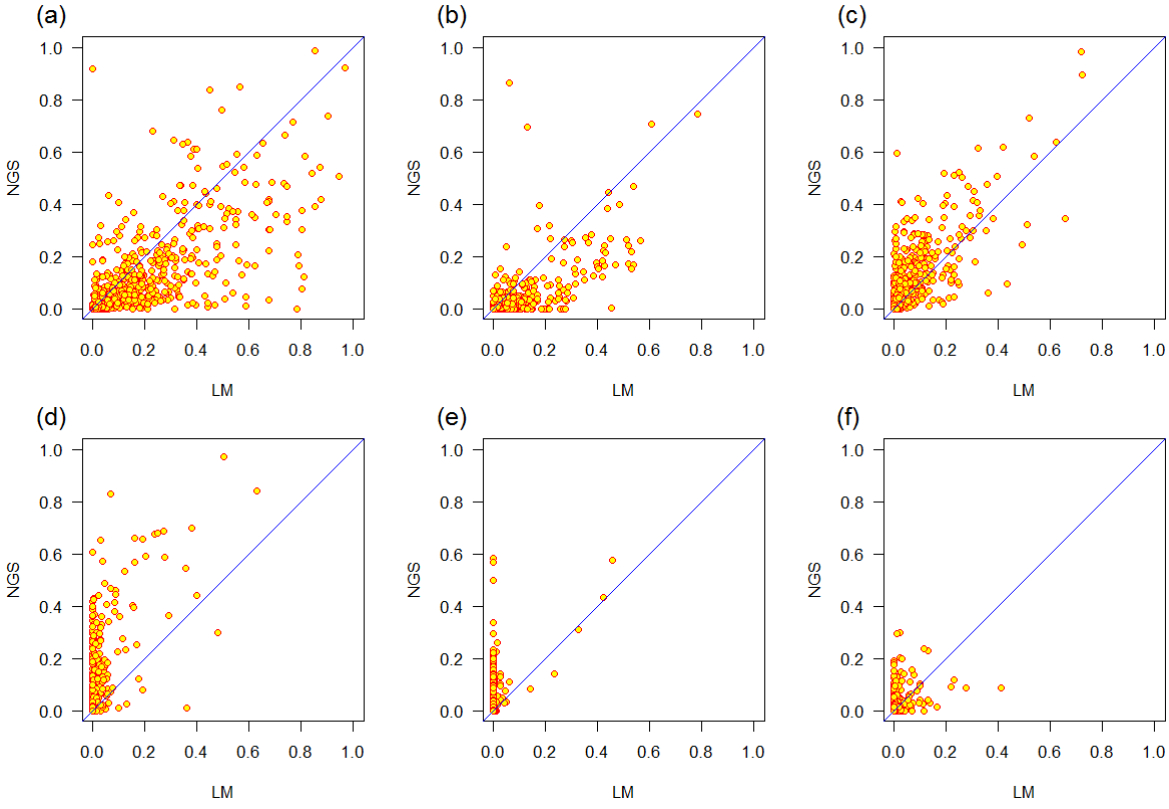


Figure 6.3 Differences between representation of common taxa in LM and NGS analyses of selected diatom species: (a) *Achnanthidium minutissimum* type (small, 1 chloroplast); (b) *Amphora pediculus* (small, 1 chloroplast); (c) *Navicula lanceolata* (medium sized, 2 chloroplasts); (d) *Melosira varians* (large, many chloroplasts); (e) *Fistulifera saprophila* (very small, 4 chloroplasts, weakly silicified); (f) *Mayamaea atomus* including var. *permitis* (very small, possibly 2 chloroplasts, weakly silicified)

Mismatches in Figure 6.2 represent 3 possible situations.

The first is that, because NGS typically produces 2 orders of magnitude more data per sample than LM, the detection limit is much lower. Hence, a species 'missed' by LM may be detected by NGS, albeit as a very small proportion of the total (and, as such, is unlikely to have a significant effect on interpretation). This component will be

exacerbated in situations such as *Melosira varians* (see above), which are typically overrepresented in NGS compared with LM (Figure 6.3).

However, it is also possible that some of the discrepancy within Figure 6.2 reflects limitations in either the barcode database or morphology-based taxonomy. In cases where a true 'biological' species can be summarised by distinct morphological criteria, there should be a good correspondence with the corresponding barcode, as is the case for *Navicula lanceolata* (Figure 6.4a). However, many taxa are known or suspected to be complexes and, in many cases, the limits of species within these complexes are still the subject of debate. In some instances (for example, *Nitzschia palea*), the complex is represented by a number of barcodes and the barcode database can be assumed to reflect much of the genetic diversity (Figure 6.4b). In other cases, the complex may be represented by fewer barcodes (for example, *Amphora pediculus, Cocconeis placentula*), leading to underrepresentation in the NGS data (Figure 6.4c). Finally, in a few instances (for example, *Gomphonema calcifugum*), the absence of a barcode altogether means that taxa will be missed entirely by NGS.

A third reason for discrepancies may be limitations in the LM method. Firstly, the LM method does not distinguish between cells that were alive or dead at the time of sampling, and secondly, the use of strong oxidising agents in the preparation of samples for analysis can lead to the dissolution of weakly silicified valves. Conversely, every record of an rbcL gene does not necessarily originate from a cell that was healthy at the time the sample was collected, and it is best to assume that the 2 types of data offer different perspectives, rather than that one is 'right' while the other is 'wrong'.

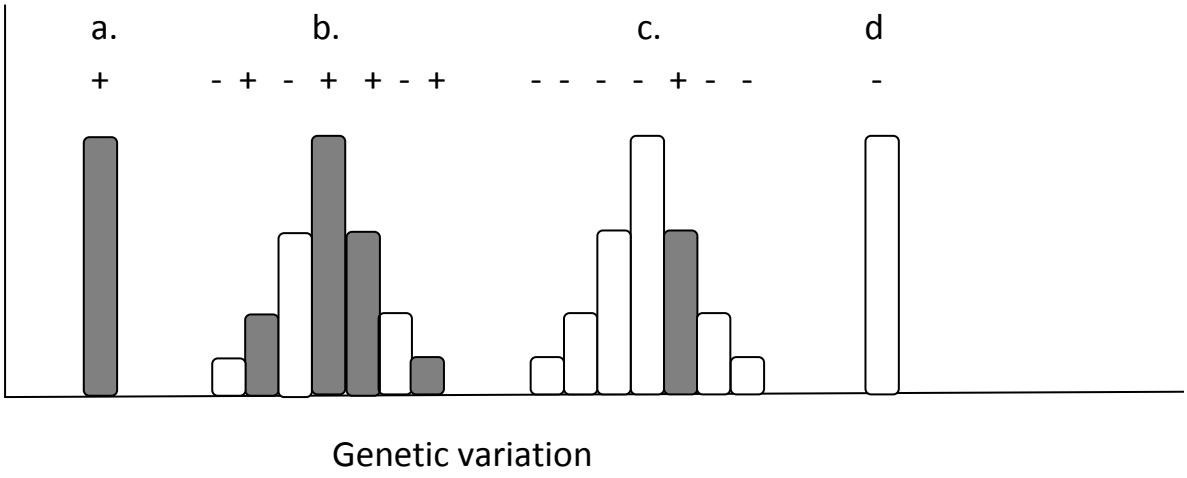


Figure 6.4 Conceptual diagram of relationship between LM and NGS outputs for 4 different scenarios: (a) clearly defined taxon aligns with barcode; (b) species complex with several different barcodes represented in the barcode database; (c) species complex poorly represented in the barcode database; and (d) species (or complex) not represented in the barcode database

Notes: '+' or '-' indicate that a barcode either does or does not exist for a particular genotype within a species complex.

6.3.2 Comparisons of LM and NGS datasets

Following these initial comparisons of the distribution of species within the LM and NGS datasets, both were then subject to NMDS ordinations to examine the consequences of any differences on the structure of the datasets. This, in turn, would indicate whether:

- ecological status concepts developed for LM can be reliably transferred to NGS
- inferences derived from NGS data can be compared with older data based on LM

In both cases, NMDS yielded ordinations with low levels of stress (LM: 0.17, NGS: 0.18) that faithfully represented the original inter-sample dissimilarities. The 2 ordinations showed similar structure in terms of the first axes of each being strongly correlated (Pearson correlation coefficient, r = 0.87) (Figure 6.5) and in terms of the correlation between the first 2 axes assessed by a Procrustes analysis (p = 0.001; 999 permutations). Moreover, the first axis of the NMDS based on LM was strongly (negatively) correlated with TDI4 (Pearson correlation coefficient, r = -0.94) (Figure 6.6)

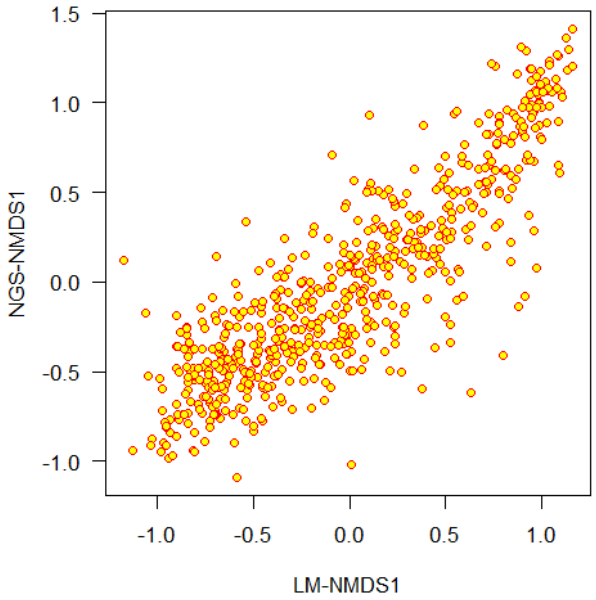


Figure 6.5 Comparison of the first axes of NMDS ordinations performed using LM and NGS data (r = 0.87)

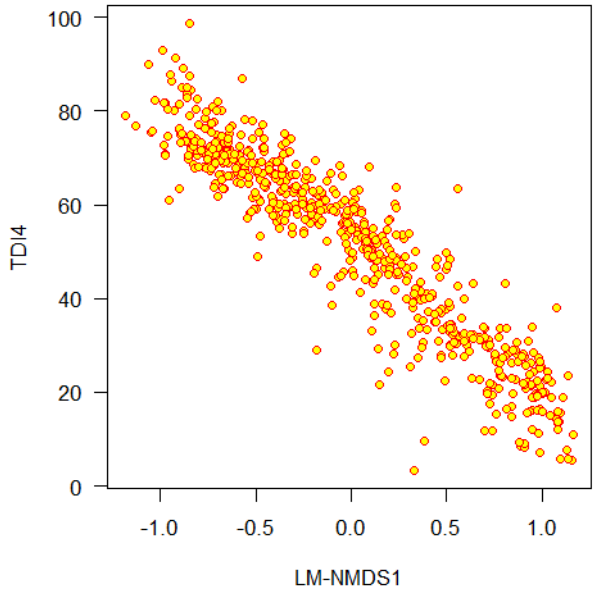


Figure 6.6 Axis 1 of NMDS of LM data versus TDI4 (r = -0.94)

TDI4, calculated using the current version from NGS data was strongly correlated with the TDI4 calculated using LM data (Figure 6.7a; Pearson correlation coefficient, *r* = 0.86), but the line deviated from 1:1 (Lin's concordance correlation coefficient: 0.81), with many NGS analyses returning higher values for the same sample than LM when the TDI (LM) was low and moderate. This may reflect the generally high numbers of *Achnanthidium minutissimum*, which has a high LM to NGS ratio (Figure 6.3a), in low nutrient (low TDI) sites and higher numbers of taxa such as *Navicula lanceolata* and, in particular, *Melosira varians*, which have much lower LM:NGS ratios (Figure 6.3c, Figure 6.3d).

These initial results suggested the need to recalibrate the TDI for use with NGS data. Figure 6.7b shows the outcome when NGS specific weights are calculated by weighted averaging, using the LM TDI as the explanatory variable.

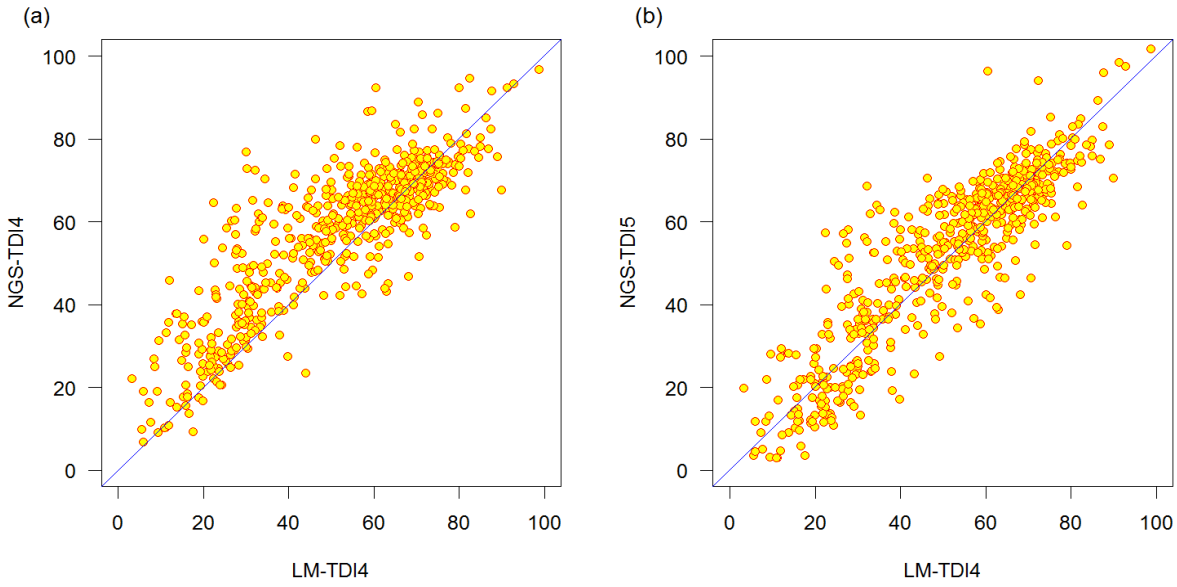


Figure 6.7 Comparison between the TDI calculated on LM and NGS data for 628 samples from UK rivers: (a) using TDI4 (LM) weights to calculate TDI for NGS data (Pearson's *r* = 0.86, Lin's *r* = 0.81; and (b) using NGS specific weights ('TDI5', Pearson's *r* = 0.90, Lin's *r* = 0.89; RMSE = 9.3)

Notes: RMSE = root mean square error

However, early attempts at this exercise revealed a continuing strong influence of *Melosira varians* and RAs of this taxon were down weighted (multiplied by 0.5) to reduce this effect. The Lin's concordance correlation coefficient rose from 0.81 to 0.89 as a result of these changes. There was, in addition, a strong correlation between this NGS based variant of the TDI and the first axis of an ordinations based on the NGS data (Figure 6.8; *r* =-0.95), indicating that the TDI5 captured the main ecological gradient in the data. Using the NGS specific variant of the TDI (referred to henceforth as 'TDI5'), 78% of all samples fell within 10 TDI units of the current LM based TDI4, compared with 68% when the TDI4 was applied to NGS data (Figure 6.9). For context, 10 TDI units represent 10% of the total TDI scale; acceptable variation for replicate analyses of the same sample by LM is ±8 TDI units. Species weights for TDI5 are given in Table 6.1.

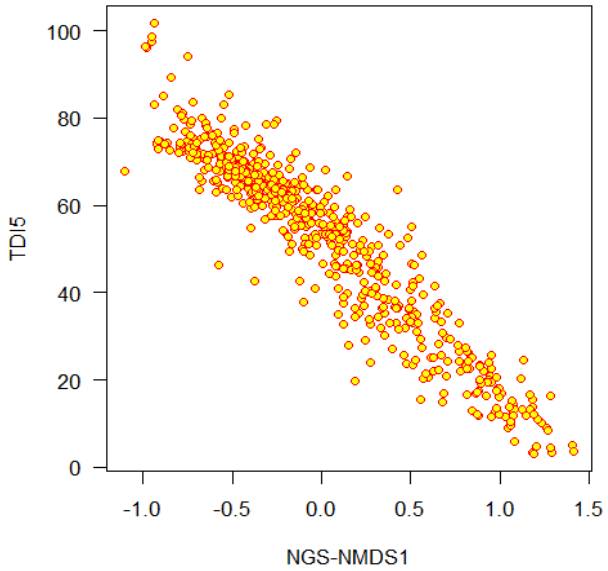


Figure 6.8 Axis 1 of NMDS of NGS data versus TDI5 (*r* = -0.95).

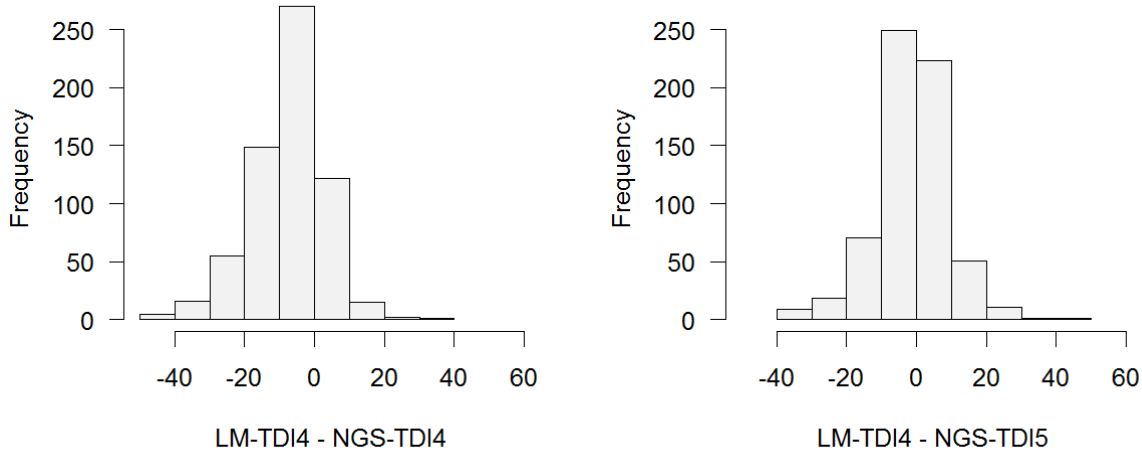


Figure 6.9 Histograms showing agreement between TDI calculated with LM and NGS data for 628 samples from UK rivers, calculated using NGS data and TDI4 weights (left) and calculated using NGS specific weights (right)

The above evaluation of TDI5 is derived using the full NGS dataset and the model tested using the same dataset. Consequently the correlation between TDI5 and TDI4 (derived from NGS and LM data respectively) may be over optimistic. Since there was no independent dataset with which to evaluate the model, bootstrap cross-validation was used to estimate the correlation between TDI4 and TDI5 likely when TDI5 is applied to independent data. Results using 1,000 bootstrap samples demonstrated the relationship to be robust; the bootstrap correlation coefficient is 0.89, with a 95% confidence interval of 0.86–0.91. Corresponding values for Lin's concordance correlation were also 0.89, with a 95% confidence interval of 0.86–0.91.

Table 6.1 Species coefficients [1]

Taxon ID	Taxon	Coefficient	Note
AC143A	*Achnanthes oblongella*	1.00	
AC004A	*Achnanthes pseudoswazi*	1.47	
XX0021	*Achnanthidium coarctatum*	2.06	
ZZZ835	*Achnanthidium minutissimum*	1.63	RA upweighted by 1.5
AT9999	*Actinocyclus* sp.	3.36	
ADLA-01	*Adlafia bryophila*	2.96	
ADLA-03	*Adlafia minuscula*	3.28	
XX0004	*Amphora berolinensis*	3.94	
AMPH-05	*Amphora pediculus*	5.24	
BA005A	*Bacillaria paxillifer*	3.97	
XX0023	*Berkeleya* sp.	4.11	
BR010A	*Brachysira neoexilis*	1.00	
BRAC-02	*Brachysira vitrea*	1.00	
CA9999	*Caloneis* sp.	4.50	
COCO-01	*Cocconeis euglypta*	2.86	
CO005A	*Cocconeis pediculus*	3.69	
CI002A	*Craticula accomoda*	3.17	
YH001A	*Ctenophora pulchella*	1.80	
CL001A	*Cymatopleura solea*	2.04	
CM007A	*Cymbella cymbiformis*	1.82	
CM9999	*Cymbella* sp.	2.36	
CYMB-01	*Cymbopleura naviculiformis*	1.82	
DE003A	*Denticula kuetzingii*	5.34	
DT022A	*Diatoma moniliformis*	1.85	
DT9999	*Diatoma* sp.	3.99	
DT004A	*Diatoma tenuis*	2.64	
DIAT-01	*Diatoma vulgaris*	4.01	
DD001A	*Didymosphenia* cf *geminata*	1.48	See note 2
DP9999	*Diploneis subovalis*	4.97	
EL001A	*Ellerbeckia* sp. TN-2014 isolate 12	4.82	
EY011A	*Encyonema minutum*	2.32	
EY016A	*Encyonema silesiacum*	2.21	

Taxon ID	Taxon	Coefficient	Note
EY9999	*Encyonema* sp.	2.82	
ENCS-07	*Encyonopsis falaisensis*	1.99	
ENCS-01	*Encyonopsis microcephala*	2.10	
EOLI-01	*Eolimna minima*	3.34	
XX0008	*Eolimna* sp.	4.80	
EP003A	*Epithemia argus*	3.86	
EP001A	*Epithemia sorex*	1.44	
EUCO-01	*Eucocconeis laevis*	1.00	
EU013A	*Eunotia arcus*	1.00	
EU070A	*Eunotia bilunaris*	1.00	
EU018A	*Eunotia* cf *formica*	1.00	
EU009A	*Eunotia exigua*	1.00	
EU107A	*Eunotia implicata*	1.00	
EU110A	*Eunotia minor*	1.00	
FA001A	*Fallacia pygmaea*	3.73	
FIST-02	*Fistulifera pelliculosa*	3.90	
FIST-01	*Fistulifera saprophila*	3.63	
FIST-03	*Fistulifera solaris*	3.88	
FR009A	*Fragilaria capucina*	1.19	
FR040B	*Fragilaria mesolepta*	1.55	
FRAG-03	*Fragilaria pararumpens*	1.00	
ZZZ842	*Fragilaria perminuta*	2.54	
ZZZ939	*Fragilaria radians*	3.73	
FR9999	*Fragilaria* sp.	1.89	
SY013A	*Fragilaria tenera*	1.00	
FR007A	*Fragilaria vaucheriae*	2.26	
FRUS-03	*Frustulia crassinervia*	1.00	
GEIS-02	*Geissleria decussis*	3.12	
ZZZ834	*Gomphonema 'intricatum'* type	2.69	
GO006A	*Gomphonema acuminatum*	1.00	
GO003E	*Gomphonema angustatum*	2.90	
GO029A	*Gomphonema clavatum*	2.65	
GO074A	*Gomphonema hebridense*	1.00	

Taxon ID	Taxon	Coefficient	Note
GO050A	*Gomphonema minutum*	2.36	
GO013A	*Gomphonema parvulum*	1.45	
XX0006	*Gomphonema pseudobohemicum*	2.83	
GO9999	*Gomphonema* sp.	2.45	
GO023A	*Gomphonema truncatum*	2.75	
AM084A	*Halamphora montana*	3.89	
HN001A	*Hannaea arcus*	1.00	
KARA-03	*Karayevia ploenensis*	5.09	
ZZZ900	*Lemnicola hungarica*	2.36	
LU9999	*Luticola* sp.	3.23	
LU009A	*Luticola ventricosa*	4.97	
MA9999	*Mastogloia* sp.29x07B	2.12	
MAYA-01	*Mayamaea atomus*	3.83	
ME015A	*Melosira varians*	3.99	RA downweighted by 0.5
MR001A	*Meridion circulare*	1.27	
NA037A	*Navicula angusta*	1.55	
NA066A	*Navicula capitata*	4.98	
NA007A	*Navicula cryptocephala*	3.38	
NA751A	*Navicula cryptotenella*	4.27	
NA023A	*Navicula gregaria*	3.95	
NA009A	*Navicula lanceolata*	3.97	RA downweighted by 0.5
NA030A	*Navicula menisculus*	4.18	
NA003A	*Navicula radiosa*	3.80	
NA9999	*Navicula* sp.	3.61	
NA095A	*Navicula tripunctata*	4.38	
NA063A	*Navicula trivialis*	4.13	
NA054A	*Navicula veneta*	4.04	
NE003A	*Neidium affine*	4.29	
NE007A	*Neidium dubium*	2.20	
NI042A	*Nitzschia acicularis*	3.68	
XX0002	*Nitzschia alicae*	2.91	
NI014A	*Nitzschia amphibia*	5.48	
NI028A	*Nitzschia capitellata*	4.22	

Taxon ID	Taxon	Coefficient	Note
NI024A	*Nitzschia dissipata*	3.86	
NI002A	*Nitzschia fonticola*	1.57	
NI034A	*Nitzschia hantzschiana*	3.33	
NI052A	*Nitzschia heufleuriana*	3.70	
NI043A	*Nitzschia inconspicua*	4.66	
NI031A	*Nitzschia linearis*	4.03	
NI009A	*Nitzschia palea*	3.63	
NI033A	*Nitzschia paleacea*	3.49	
NI005A	*Nitzschia perminuta*	1.52	
NI152A	*Nitzschia pusilla*	4.64	
NI025A	*Nitzschia recta*	4.56	
XX0020	*Nitzschia romana*	3.63	
NI006A	*Nitzschia sigma*	3.63	
NI046A	*Nitzschia sigmoidea*	4.53	
NI166A	*Nitzschia sociabilis*	2.07	
NITZ-03	*Nitzschia soratensis*	3.90	
NI9999	*Nitzschia* sp.	3.17	
XX0022	*Parlibellus hamulifer*	4.25	
PARL-01	*Parlibellus protracta*	3.00	
PE002A	*Peronia fibula*	1.00	
PI006A	*Pinnularia grunowii*	2.79	
PI011A	*Pinnularia microstauron*	1.00	
XX0007	*Pinnularia neomajor*	2.23	
PI9999	*Pinnularia* sp.	1.00	
PI022A	*Pinnularia subcapitata*	1.03	
ZZZ872	*Placoneis clementis*	2.91	
ZZZ896	*Planothidium frequentissimum*	4.26	
ZZZ897	*Planothidium lanceolatum*	3.34	
PLAT-01	*Achnanthes Platessa conspicua*	5.89	
ZZZ910	*Psammothidium bioretii*	2.07	
PS001A	*Pseudostaurosira brevistriata*	4.55	
RE001A	*Reimeria sinuata*	2.87	
RC002A	*Rhoicosphenia abbreviata*	4.46	

Taxon ID	Taxon	Coefficient	Note
RH001A	*Rhopalodia gibba*	1.00	
SELL-01	*Sellaphora joubaudii*	4.66	
SL002A	*Sellaphora seminulum*	4.14	
SA006A	*Stauroneis phoenicenteron*	2.43	
SR001A	*Staurosira construens*	3.28	
SR002A	*Staurosira elliptica*	4.41	
STAS-01	*Staurosirella martyi*	4.31	
SU073A	*Surirella brebissonii*	3.49	
SY003A	*Synedra acus*	1.43	
TA001A	*Tabellaria flocculosa*	1.00	
TU003A	*Tabularia fasciculata*	3.86	
TF9999	*Tryblionella constricta*	2.76	
ZZZ985	*Tryblionella debilis*	4.13	
SY001A	*Ulnaria ulna*	2.66	

Notes [1] Table gives species scores for TDI5 (see Section 6.2.2 for details)
[2] The species epiphet *Didymosphenia geminata* has been applied to all records of the genus *Didymosphenia* assigned to NGS reads. *D. geminata* is not represented in the barcode database. Some records of *D. dentata* were assigned during BLAST searches of GenBank, although this species has not been recorded from the UK.

How much of the observed difference between the TDI calculated with LM and NGS data is likely to be due to gaps in the barcode database? The database currently represents just 176 of over 2,500 species recorded from UK and Ireland freshwaters? Figure 6.10 shows the relationship between the LM TDI calculated with all available taxa (x axis) and the LM TDI calculated with just those taxa included in the barcode database. The high correlation between the 2 variants (Pearson correlation coefficient, $r = 0.991$) suggests that most of the biological variation within diatom assemblages is being captured by the barcode database, although there are still a few samples where the variation is greater. A few ecologically significant taxa – in particular, *Achnanthidium pyrenaicum, Gomphonema calcifugum* and *G. pumilum* – are still absent from the barcode database or are underrepresented.

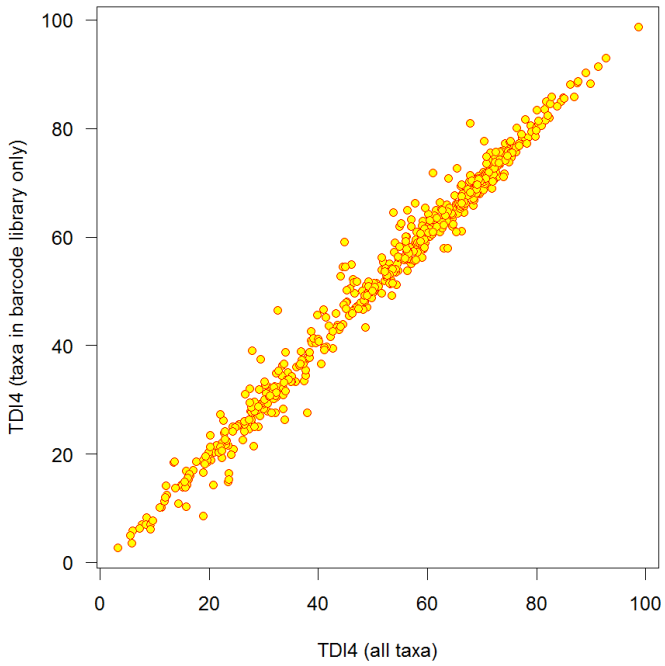


Figure 6.10 Difference between TDI4 based on LM data calculated with all taxa and with just those taxa represented in the barcode database

6.3.3 From metric to classification: calculation of eTDI and EQR

The next step is to transform the raw TDI into an EQR by dividing by a denominator that provides an estimate of the TDI at reference conditions for that site. For the current LM based approach, this is determined by an equation that uses alkalinity to predict the value of the TDI:

$$eTDI4 = 9.933 \times \exp(\log10(Alk) \times 0.81) \quad 6.5$$

where eTDI4 is the expected value of TDI4 and Alk is the average alkalinity at the site.

Figure 6.11a shows a strong correspondence between this equation and LM analyses of 171 samples from this study, which were collected from reference sites throughout the UK ($r^2 = 0.58$).

However, Equation 6.5 appears to under predict eTDI when applied to the NGS data. Therefore the procedure in the derivation of the original TDI4 was followed and a new equation was fitted to the NGS data using least squares regression (Figure 6.11b).

$$eTDI5 = -11.43 + [32.65 \times \log_{10}(Alk)] \quad 6.6$$

where eTDI5 is the expected value of TDI5.

Equation 6.6 provided a means to calculate the site-specific predicted NGS-TDI score from mean site alkalinity (eTDI5) for use in calculating the EQR, and was adopted for subsequent analyses of the NGS data.

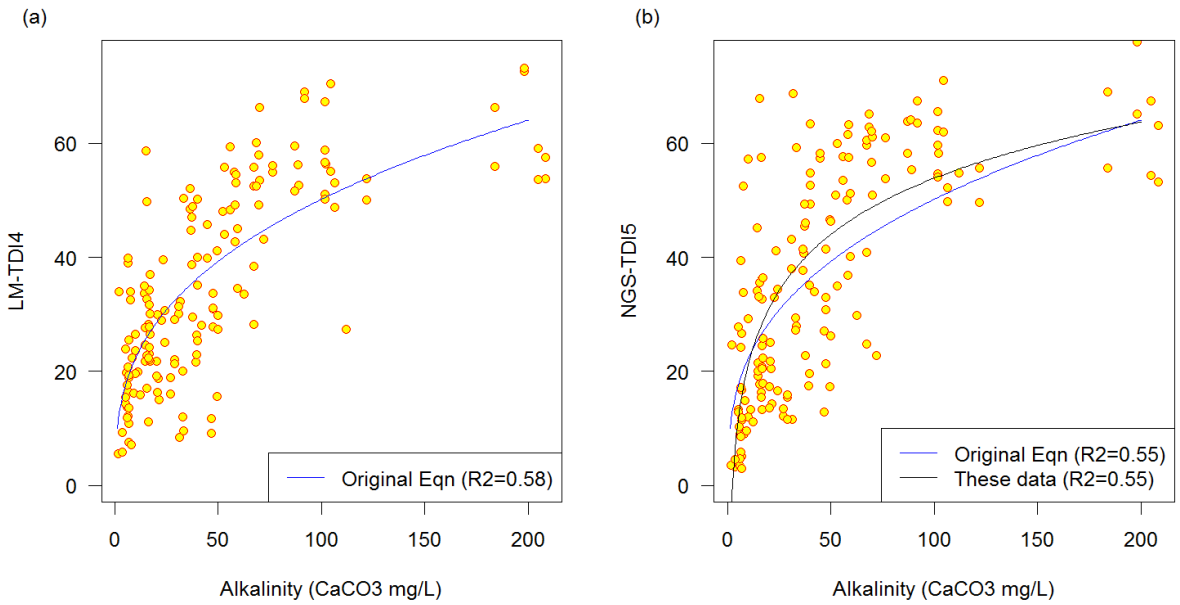


Figure 6.11 Relationship between alkalinity and TDI for 171 samples from reference sites throughout the UK: (a) based on LM results and TDI4 calculation (Equation 6.5); and (b) based on NGS results and TDI5 calculation (see Equation 6.6).

Like the raw metrics, there was a strong relationship between EQRs computed with LM and NGS approaches (Figure 6.12).

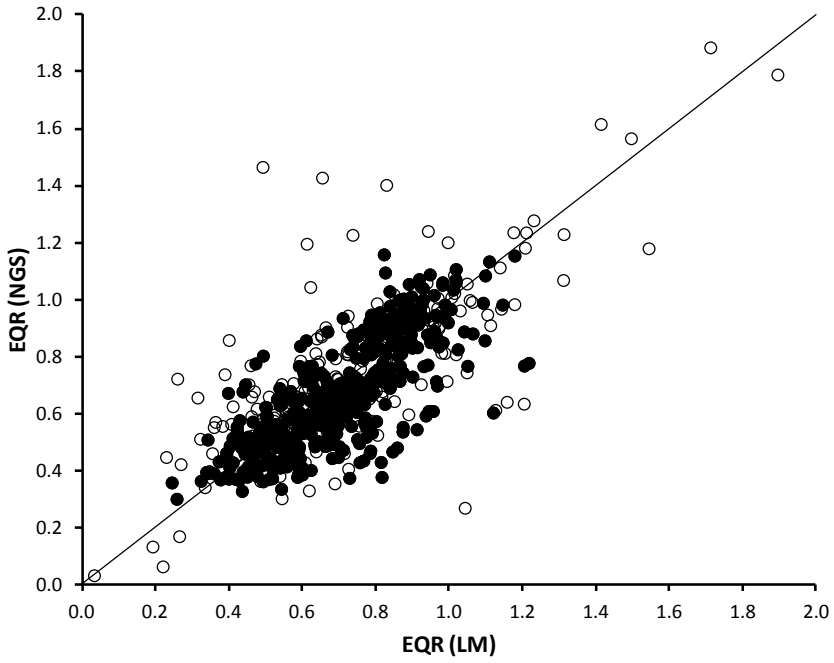


Figure 6.12 Comparison between EQR calculated on LM and NGS data for 620 samples from UK rivers for which alkalinity data were available

Notes: Open circles show samples from the entire alkalinity gradient (1.7–353 mg $CaCO_3$ per litre) (r = 0.75).
Closed circles show samples from sites where alkalinity is <120 mg $CaCO_3$ per litre) (r = 0.77).
Diagonal line shows slope = 1.

Many of the outliers around this relationship are samples from sites with high alkalinity (>120 mg $CaCO_3$ per litre), where it is recognised that the current phytobenthos EQRs do not necessarily reflect the response to nutrient pressure effectively. Excluding these high alkalinity data from the relationship increases the correlation slightly (Pearson correlation coefficient, r = 0.75 for all sites and 0.77 for sites with alkalinity ≤120 mg $CaCO_3$ per litre).

Using normalised versions of the current intercalibrated class boundaries (high: 0.8, good: 0.6, moderate: 0.4, poor: 0.2) and amalgamating all samples from a water body following current Environment Agency classification procedures, 70% of water bodies were assigned to the same class using both LM and NGS. Some 98% agreed to within one class (Table 6.2), with the current LM method showing a tendency (21% of sites) to more stringent classifications than NGS. As a result, no sites currently classified as high or good status would be downgraded to moderate, poor or bad status using NGS. However, this analysis is based on the sub-element phytobenthos only. In practice, final water body status is determined from several biological quality elements, which will further buffer the effect of any changes in status based on phytobenthos alone.

Table 6.2 Comparison between ecological status classes computed by LM and NGS variants of the TDI

	TDI5 (NGS)				
TDI4 (LM)	**H**	**G**	**M**	**P**	**B**
H	105	7	0	0	0
G	31	34	2	0	0
M	3	18	6	0	0
P	0	0	1	0	0
B	0	0	0	0	0

Notes: *n* = 207 water bodies
B = bad status; P = poor status; M = moderate status; G = good status; H = high status
Green shading: identical classification for both LM and NGS
Yellow shading: agreement to within one class between LM and NGS

7 Comparison of uncertainty in LM and NGS analyses

7.1 Introduction

The previous sections have established that there is a strong correspondence between LM and NGS analyses across the nutrient/organic pressure gradient while, at the same time, noting some important differences in the expression of individual species. As most ecological assessment methods involve the conversion of a continuous EQR scale to a categorical classification of status, some mismatch between class (see Table 6.2) is statistically inevitable when 2 classifications are compared, reflecting uncertainties in the underlying model. Other aspects of uncertainty will reflect stochastic and analytical variability introduced during the data gathering phases. Although LM and NGS share the same sampling process, subsequent treatment of samples is different in each case and it is therefore likely that the uncertainties associated with LM and NGS will also differ. This, in turn, will influence the confidence with which water bodies can be assigned to particular status classes.

This section describes experiments on variation at a number of levels, from field sampling through to laboratory analysis, in order to investigate differences in method uncertainty and performance characteristics between the current approach using LM and the NGS analytical process.

7.2 Methods

7.2.1 Study design

The sources of uncertainty investigated during this study are listed in Table 7.1. Background details of the locations from which samples were collected are provided in Table 7.2.

Triplicate samples were collected from one site per water body in spring 2014 to allow within site variation to be estimated; and one of these samples was also subsampled to allow analytical variation to be established (Experiment 1). In addition, samples were collected from this site and 2 others within the same water body on 4 occasions, allowing simultaneous investigation of variation within a water body and between seasons (Experiment 2).

One subsample per location in Experiment 1 was also circulated to a number of experienced analysts as part of the UK/Ireland diatom ring test. The standard deviation of the TDI was used as an estimate of between-operator variation. This was compared with the standard deviation for 2 operators each using 2 machines (see Section 5.1.2) to indicate the scale of between-operator variation for NGS.

Table 7.1 Sources of uncertainty investigated during the study

Source	Investigated by ...
Water body	3 locations within a single water body Stretches chosen to have no major point source inputs along their length
Site	3 samples collected from a site (location within a water body from which routine samples are collected) Samples spaced ~10m apart (upstream–downstream)
Season	4 samples collected over a 12 month period
Analytical (within sample) ('repeatability')	LM: 3 separate slides prepared from individual samples NGS: 3 separate aliquots taken for subsequent DNA extraction, amplification and analysis
Analytical (between-analyst) ('reproducibility')	LM: one sample per water body used for UK/Ireland diatom ring test; results for 'expert panel' (experienced analysts) used as indication of between-analyst variation. NGS: one sample per water body prepared separately by 2 individuals and analysed on 2 separate NGS machines

Table 7.2 Locations and characteristics of sites visited during investigations of uncertainty

Water body/site	NGR	Altitude (m)	Alkalinity (mg $CaCO_3$ per litre)
River Ehen (high status, Special Area of Conservation)			
Scout Camp [1]	NY 087 153	110	<5
Mill, footbridge	NY 081 152	100	–
Oxbow [2]	NY 072 157	95	<5
Upper River Wear (good status)			
Stanhope	NY 991 392	200	74.1
Frosterley	NZ 036 369	160	–
Wolsingham	NZ 075 369	135	84.2
River Derwent (County Durham) (moderate status)			
Ebchester [3]	NZ 101 556	60	44.5
Low Westwood	NZ 111 565	57	–
Blackhall Mill	NZ 122 569	55	85.3
River Team (poor/bad status) [4]			

Water body/site	NGR	Altitude (m)	Alkalinity (mg $CaCO_3$ per litre)
D/S East Tanfield STW [5]	NZ 198 558	120	103.7
Causey Arch	NZ 202 554	100	–
Beamish Hall	NZ 215 549	85	70.2

Notes: Alkalinity is presented as a site average (all records since 1 January 2010) based on routine Environment Agency chemical sampling, except for the River Ehen, where most values are below the routine detection limit (5 mg $CaCO_3$ per litre).
[1] Closest chemical sampling point: Bleach Green Bridge (NY 085 154)
[2] Closest chemical sampling point: Ennerdale Bridge (NY 069 158)
[3] Closest chemical sampling point: Shotley Bridge (NZ 091 527)
[4] The River Team is classified as 'heavily modified'; current ecological potential is defined as 'moderate'. Phytobenthos results are not presented in the latest River Basin Management Plan, but invertebrates are 'poor (very certain)' and phosphate is 'bad (very certain)'
[5] Closest chemical sampling point: u/s East Tanfield STW (NZ 197 553)
NGR = National Grid Reference; STW = sewage treatment works

7.2.2 Statistical analysis

Initial analyses of data structure used NDMS (see Section 6.2.2) on the combined datasets for Experiments 1 and 2. Following this, data for Experiments 1 and 2 were analysed separately, examining the variation in TDI within and between treatments using analysis of variance where initial tests demonstrated homogeneity of variances, or non-parametric alternatives (Kruskal–Wallis test for one-way comparisons, Friedman's test for two-way comparisons). The F_{max} test was used to test for homogeneity of variances.

7.3 Results

7.3.1 Preliminary analysis of data structure

Preliminary analyses investigated the structure of the pooled data from both experiments. For both LM and NGS, NMDS ordinations of the data showed low stress (0.145 and 0.169 respectively), good separation of the 4 sites, and a strong relationship between axis 1 of the ordination and the respective TDI (r = 0.861 for LM and 0.967 for NGS). There were, in addition, strong correlations between the first axis of the LM and NGS ordinations (r = 0.832) and between TDIs (r = 0.887) (Figure 7.1), though there were some interesting patterns within the datasets. In the River Ehen, for example, NGS results were fairly consistent despite variability in the LM results, while in the River Team the opposite is true, with high variability in the NGS results but stable LM results.

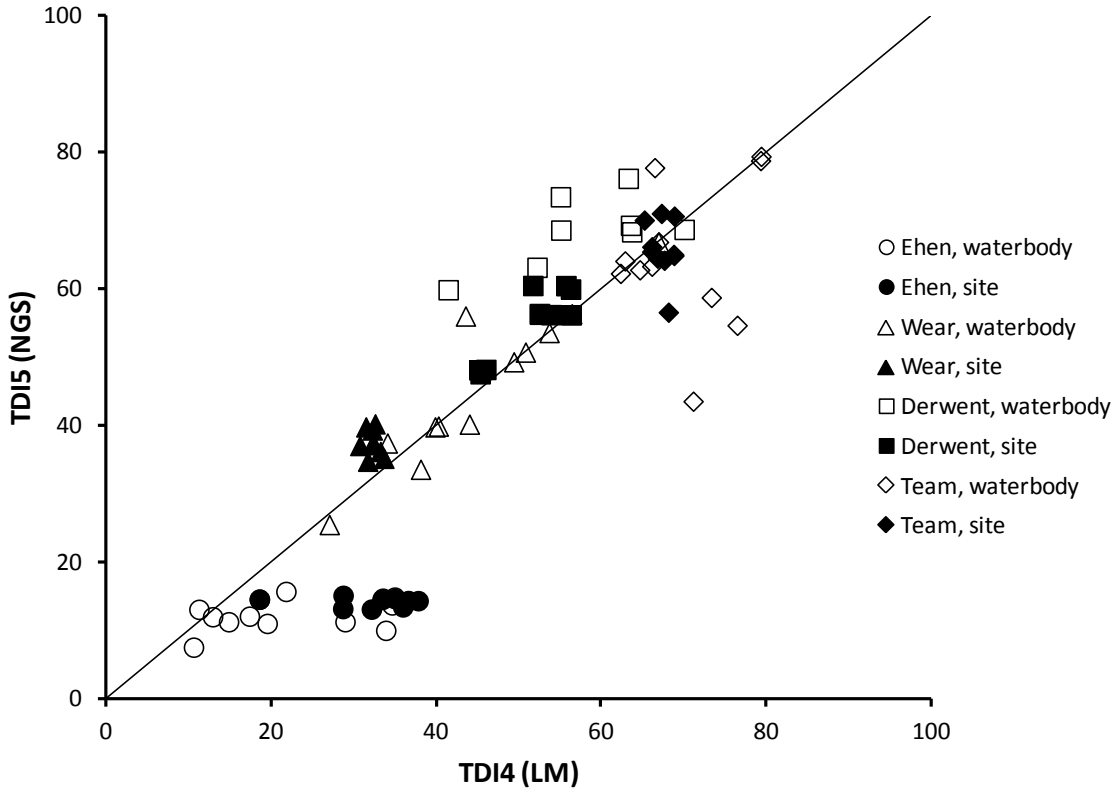


Figure 7.1 Within water body and within site variation in LM and NGS analyses of diatom samples from 4 contrasting river sites in England

Notes: Results are expressed as TDI4 (LM) and TDI5 (NGS) and the diagonal line indicates slope = 1 (LM = NGS).
Closed symbols = within site variation on a single day.
Open symbols = within water body variation over the course of a year.
See text for more details.

7.3.2 Within site and analytical variation (Experiment 1)

Although replicate analyses of 3 samples from one site per water body collected on the same day tended to have less variation than samples collected over time or between sites in the same water body (see Sections 7.3.3 and 7.3.4), there was considerable variation among LM analyses from the River Ehen (high status site) and among NGS analyses from the River Team (poor/bad status site) (Figure 7.1). The lack of variation in NGS in the former may represent the distinctive flora in the River Ehen, which is challenging to LM analysts and not all of whose representatives are represented in the barcode database at present. Within site variation in the River Team was on a similar scale to that observed for the Rivers Derwent and Wear; however, considerable within water body variation was observed for the NGS results, along with some marked differences between LM and NGS.

The most abundant diatom observed with LM was *Luticola goeppertiana*, a species not in the barcode library, while 2 Gomphonema species dominated NGS analyses. One of these (*Gomphonema pseudobohemicum*) was not recorded at all by the LM analyses and is, in any case, a species of oligo- to mesotrophic, circumneutral to slightly acid streams (Hofmann et al. 2011); the other Gomphonema species was present but in lower numbers, This suggests that part of the variation described in Section 6 may represent shortcomings in the breadth of species (and genotypes) in the barcode library at present: if a species is not represented with a reference DNA barcode, it will

be assigned to the species that has the closest barcode match. In addition, the River Derwent (moderate status) has consistently higher results for NGS than for LM, presumably due to similar factors.

In most cases, analytical variation was of a similar magnitude for both LM and NGS, although tending to be slightly lower for LM than for NGS for the River Derwent and lower for NGS than for LM in the Rivers Ehen and Wear (Table 7.3). Variance was much higher for both methods in the River Team compared with other rivers, though it was still lower for NGS than LM. Between-sample variation showed variation between samples in all cases except for NGS in the River Team.

Table 7.3 Variation within (analysis of 3 separate slides) and between replicate samples from the same site (each approximately 10m apart) at 4 water bodies of contrasting ecological quality in northern England

Location	LM		NGS	
	Variance within samples (*n* = 3)	F	Variance within samples (*n* = 3)	F
Ehen, Oxbow				
A	0.053		0.029	
B	0.743		0.066	
C	0.152	42.34 ***	0.010	57.9 ***
Wear, Wolsingham				
A	0.021		0.014	
B	0.305		0.201	
C	0.120	87.35 ***	0.542	58.36 ***
Derwent, Ebchester				
A	0.042		0.121	
B	0.001		0.011	
C	1.384	75.18 ***	0.086	1629 ***
Team, Causey Arch				
A	44.52		12.63	
B	33.62		28.54	
C	23.16	6.05 *	12.54	1.46 N.S.

Notes: Variances were homogeneous for all datasets.
Within sample variation expressed as variance; between-sample variance expressed as F.
* $p < 0.05$; ** $p \geq 0.05, < 0.01$; ***: $p \geq 0.01, < 0.001$; N.S. = not significant

This experiment was performed with a single LM analyst and a single NGS sequencer. A direct comparison of between-operator variation for LM and NGS is not possible, but an insight into this is given in Figure 7.2, which shows between-operator variation for 1 sample from each of the 4 locations, alongside the median of between-operator and instrument variation for samples from each of the 4 locations (see Section 5.1.2 for more details). Variation among LM operators was highest at the River Ehen – an

oligotrophic, soft water stream in north-west England with a challenging assemblage of diatoms. In all cases, however, between-operator variation in LM analyses was greater than between instrument variation in NGS, demonstrating that the NGS approach produces more consistent results.

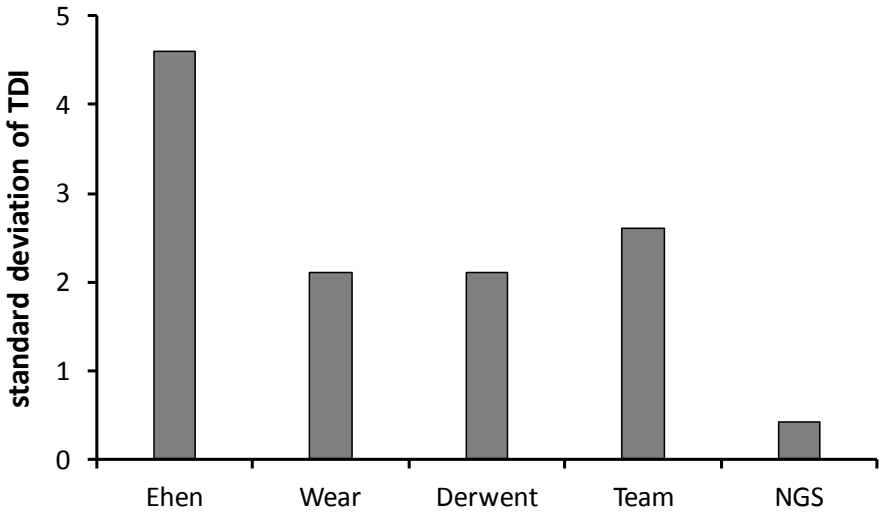


Figure 7.2 Variation (as standard deviation of TDI) between analytical results (LM) from experienced analysts for one sample from each water body reported in Table 7.2 alongside results from tests of analytical specificity for NGS

Notes: Details of the tests of analytical specificity are given in Section 5.1.2.

7.3.3 Within water body variation (spatial and temporal)

Both temporal (Figure 7.3a) and spatial (Figure 7.3b) variation within water bodies (expressed as standard deviation) were of a similar magnitude for LM and NGS (Spearman's rank correlation coefficient $r = 0.72$, $p < 0.01$; and $r = 0.790$, $p < 0.01$ respectively). However, variation in NGS tended to be lower (that is, most points below line indicating slope = 1) in more cases for each of site, temporal and water body variation.

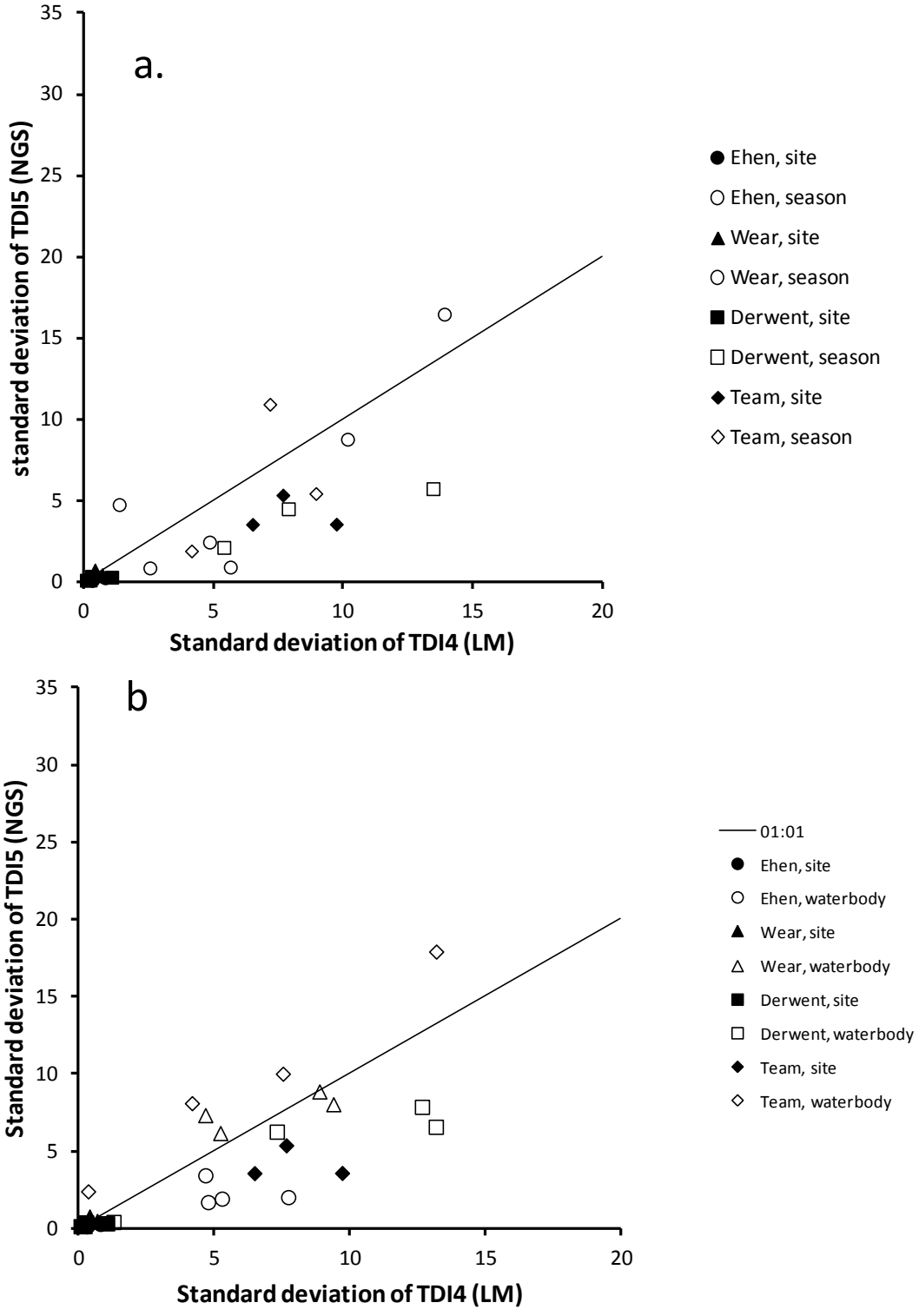


Figure 7.3 Within site and within waterbody variation in LM and NGS analyses of diatom samples from 4 contrasting river sites in England expressed as standard deviation: (a) water body variation expressed as spatial variation within the water body (*n* = 3) on 4 separate occasions; and (b) water body variation expressed as temporal variation (*n* = 4) at each of 3 locations per water body

Notes: Diagonal line indicates slope = 1 (that is, identical variability using both methods).
Closed symbols = within site variation on a single day
Open symbols = within water body variation over the course of a year

Following this overview, each water body was analysed separately. Preliminary F_{max} tests indicated that the assumption of homogeneous variances was violated once for LM analyses (seasonal variation in the River Derwent) and 4 times for NGS analyses (both seasonal and between-site variation in the Rivers Derwent and Ehen).The non-parametric Kruskal–Wallis and Friedman tests were therefore used in place of conventional analysis of variance.

Temporal variation exceeded spatial variation in almost all cases using LM (Table 7.4), though it was only significant (that is, $p < 0.001$) in the River Derwent. Despite this, seasonality was apparent in all cases for LM (Figure 7.4), though less so for NGS (Figure 7.5). The seasonal patterns also varied between rivers, with the lowest TDI values recorded in the summer in all but the Ehen, where winter samples were lowest. Highest values were recorded in the autumn (Ehen), winter (Wear, Team) or spring (Derwent) for LM (Figure 7.4). In contrast, seasonal patterns were less pronounced for NGS (Figure 7.5; Table 7.4). Spatial variation within water bodies for NGS was significant only in the Rivers Ehen and Team.

Table 7.4 Outcome of one-way Kruskal–Wallis (KW) and two-way Friedman (F) tests on within water body variation in TDI determined by LM and NGS

River	LM			NGS		
	Spatial	Temporal	2-way	Spatial	Temporal	2-way
	KW	KW	F	KW	KW	F
Ehen	3.50 N.S.	6.28 N.S.	7.4 N.S.	8.0 *S.	1.87 N.S.	6.6. N.S.
Wear	1.19 N.S.	7.51 N.S.	7.0 N.S.	1.19 N.S.	6.49 N.S.	5.8 N.S.
Derwent	1.08 N.S.	8.44 *	4.5 N.S.	7.42 *	1.36 N.S.	6.0 *
Team	1.42 N.S.	5.61 N.S.	4.2 N.S.	7.38 *	0.74 N.S.	1.0 N.S.

Notes: * $p < 0.05$; N.S. = not significant

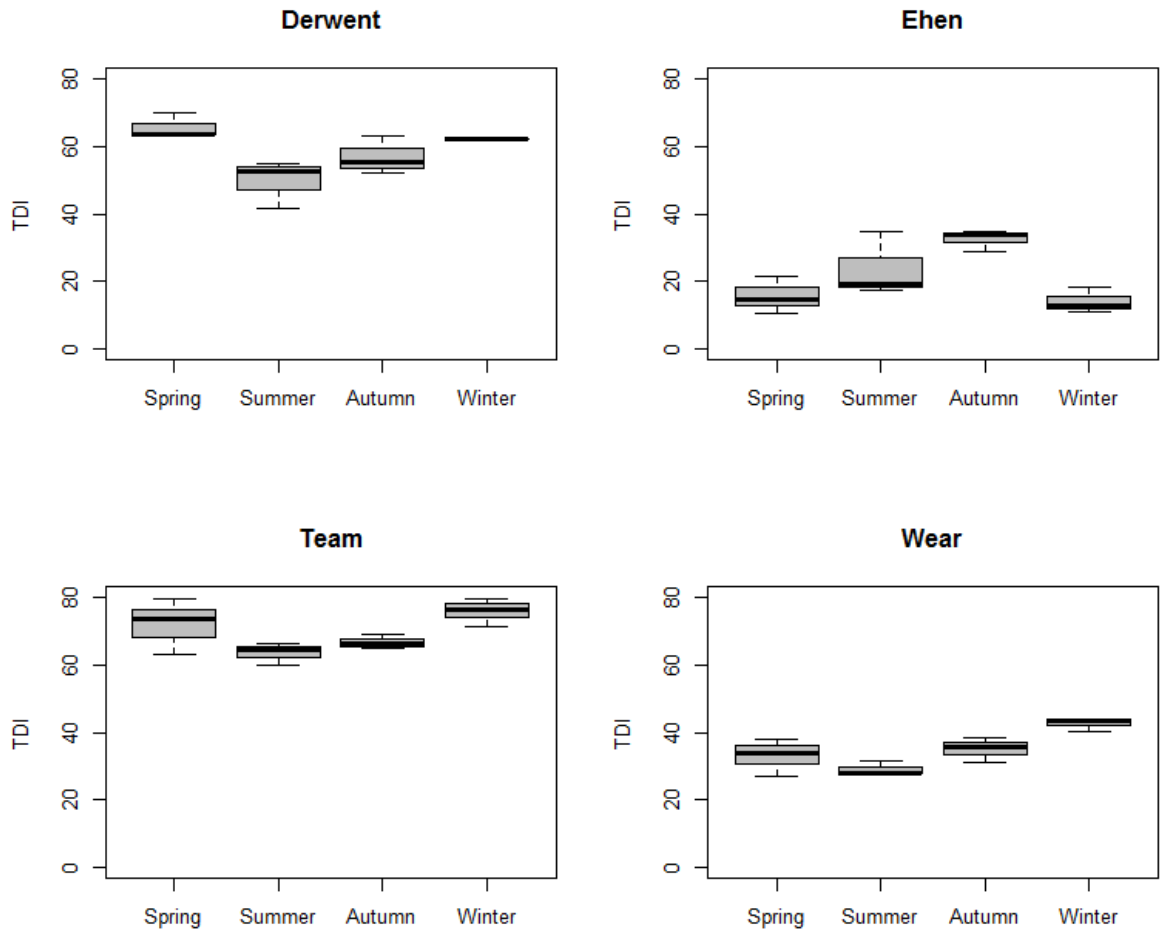


Figure 7.4 Seasonal variation in TDI4 (LM analyses) in the Rivers Ehen, Wear, Derwent and Team

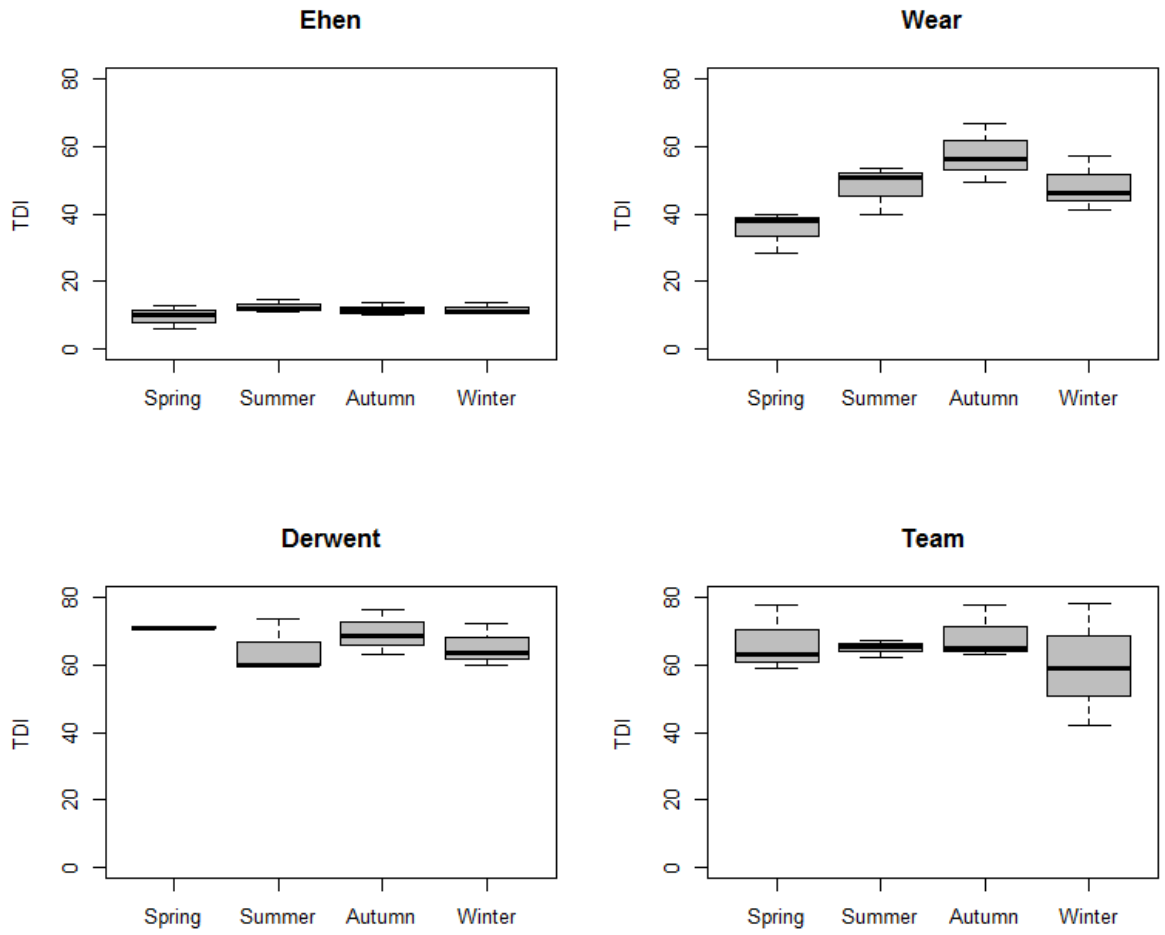


Figure 7.5 Seasonal variation in TDI5 (NGS analyses) in the Rivers Ehen, Wear, Derwent and Team

7.3.4 Overview of sources of uncertainty in LM and NGS assessments of ecological status using diatoms

The results from the previous 2 sections can now be collated and combined with data from Section 5.1.2 to allow comparisons of the scale of the different sources of uncertainty associated with LM and NGS (Figures 7.6 to 7.9). One additional source of variation is included on these plots, that is, that between diatoms and other algae. This is based on variation between the Norwegian non-diatom index, PIT (Schneider and Lindstrøm 2011) and the TDI based on samples collected from the same site on the same day (Schneider et al. 2013); 95% confidence limits of predictions were estimated by eye and then halved to give an approximate value that was used on all 4 plots to indicate the scale of an uncertainty component that would otherwise be invisible.

These plots allow comparisons between different sources of variation within a single water body. Caution is needed for comparisons between water bodies as assumptions regarding homogeneity of variance are not always satisfied (see Section 7.3.3) and standard deviations will be influenced by the site mean. Some generalisations are, nonetheless possible (assuming standard deviation in TDI of <2 = low, 2–6 = intermediate and >6 = high):

Analytical variation (that is, replicate analyses of the same sample by a single analyst and by several analysts) generally has low levels of variation for both NGS and LM

(Table 7.3). Reproducibility (that is, replicate analyses by several individuals/laboratories) is higher for NGS than for LM. Repeatability (that is, replicate analyses by the same individual/laboratory) is similar or slightly lower for NGS compared with LM.

Variability at higher spatial and temporal scales is generally greater than for analytical variation for both LM and NGS (Figures7.6 to 7.9). However, it varies considerably from river to river. There was no consistent trend of LM being either lower or higher than NGS. It is possible that apparently low levels of variation at these scales in the River Ehen, in particular (Figure 7.6), may be an artefact of the limited coverage of the flora found at this site in the barcode library. However, barcodes for missing species can be added to the barcode library in the future as they become available, allowing the precision of the NGS method to improve over time.

Analytical variation by both approaches is generally lower or of a similar magnitude to the variation between ecological status estimates based on diatoms and non-diatoms. Water body spatial and temporal variation of diatoms, whether by NGS or LM, in contrast, is similar or higher.

Overall, NGS provides greater analytical precision than the current LM approach. However, the benefits of the greater analytical precision obtained from NGS are dampened, to some extent, by other sources of error (for example, between season, within site and within water body). This means that it is unlikely to lead to greater confidence of class for water body level status classifications. However, it does have the potential to improve consistency of analysis through automation (see Section 5.1.2), particularly at sites where there is a challenging assemblage of diatoms, as this appears to be an area where variability is introduced in the LM approach.

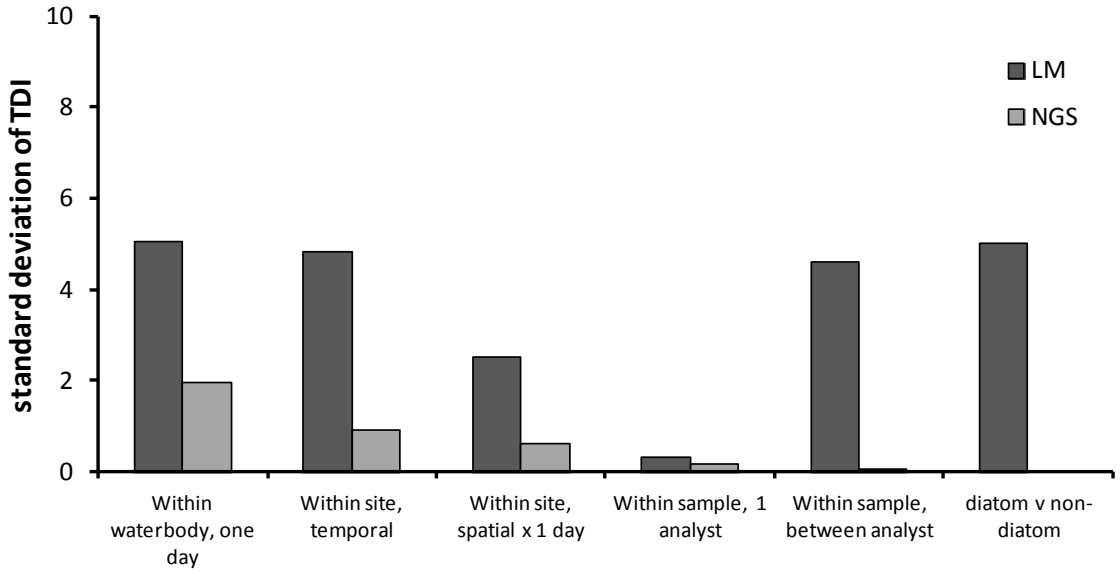


Figure 7.6 Sources of uncertainty in TDI calculated using LM and NGS data from samples collected from the River Ehen (high status)

Notes: Within sample, between-analyst variation for NGS is 0.068.

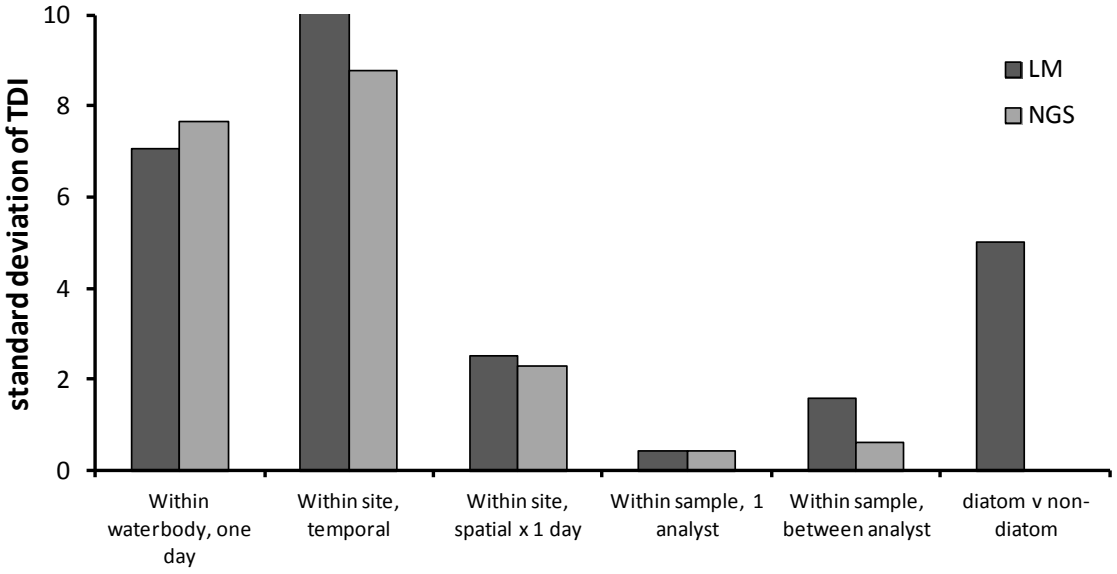


Figure 7.7 Sources of uncertainty in TDI calculated using LM and NGS data from samples collected from the River Wear (good status)

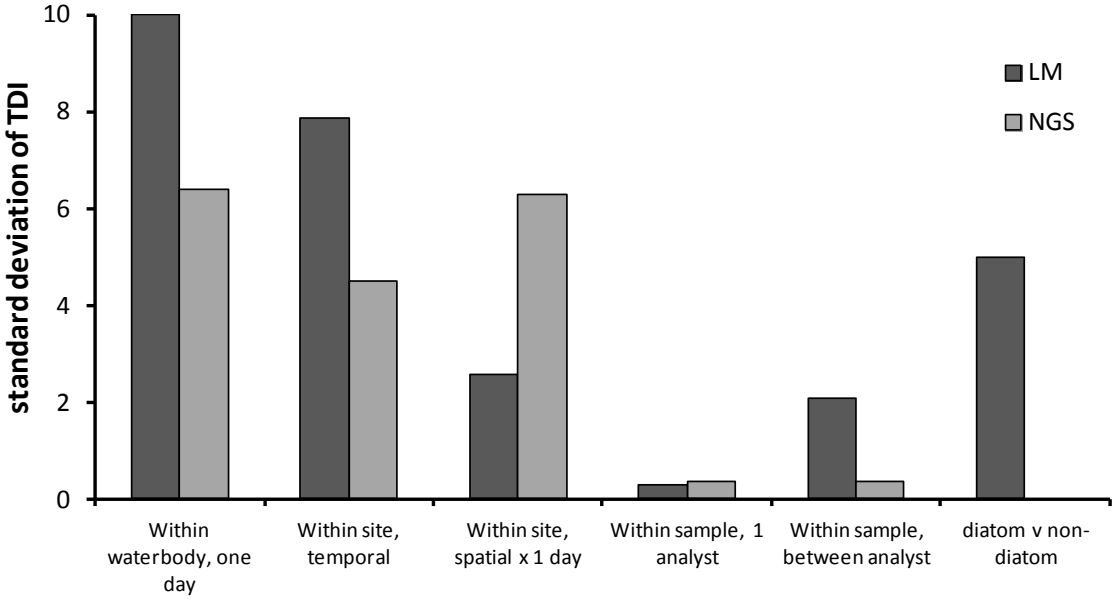


Figure 7.8 Sources of uncertainty in TDI calculated using LM and NGS data from samples collected from the River Derwent (moderate status)

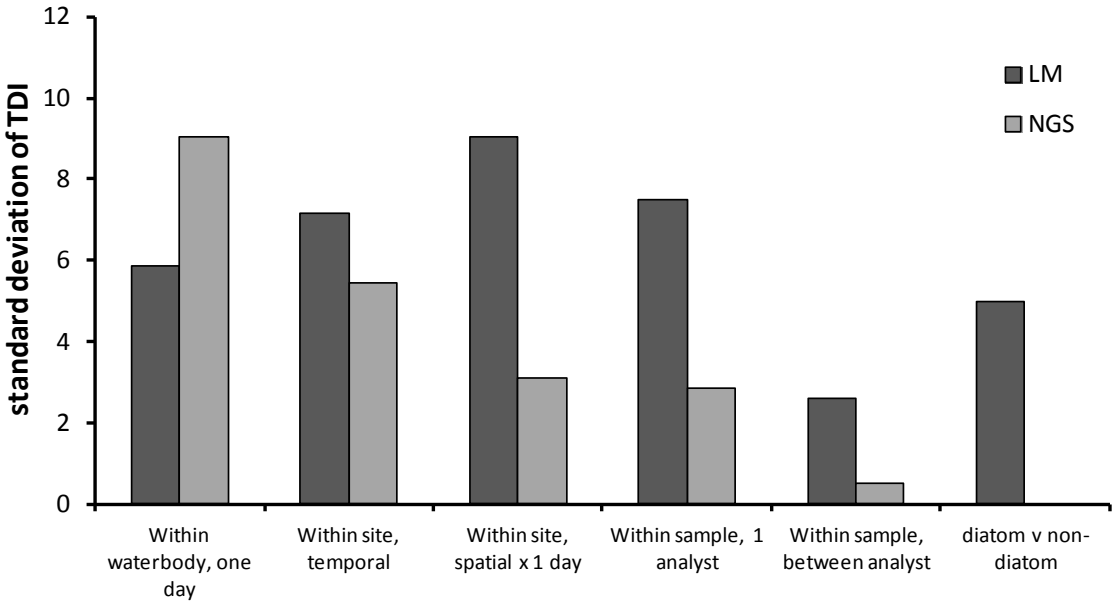


Figure 7.9 Sources of uncertainty in TDI calculated using LM and NGS data from samples collected from the River Team (poor/bad status)

8 Case study: application of the method to an operational investigation

8.1 Introduction

Having developed an NGS compatible metric and examined the performance of this at different spatial and temporal scales, the final test was to apply the method to a 'live' operational investigation of a series of small water bodies and compare the outcomes with those from the current technique to understand how the method might fit into an ecological assessment toolkit.

A study was therefore developed in conjunction with local Environment Agency staff which focused on subcatchments of the River Browney (a tributary of the River Wear) to enable the impact of 3 sewage treatment works (STWs) to be assessed; see schematic map of the area (Figure 8.1). Although the validity of sampling upstream and downstream of point source discharges has been questioned, local Environment Agency staff believe that this is the best way to demonstrate to utility companies that particular STWs are directly responsible for changes in ecology. The study also enabled the impact of the largest of the 3 STWs to be differentiated from the impact of storm sewers serving the village of Lanchester in County Durham. This, in effect, constitutes the 'before' component of a before–after–control–impact study design, widely used for assessing environmental impacts (Underwood 1991, Downes et al. 2002).

Smallhope and Stockerley Burns constitute a single water body in the Wear catchment for Water Framework Directive reporting purposes (GB103024077330) with an overall classification of bad status, driven by invertebrates, with fish and phosphorus at poor status. All other supporting elements that have been measured are at high status.

Smallhope and Stockerley Burns receive inputs from Knitsley (5,172 population equivalent) and Crook Hall (4,809 population equivalent) STWs respectively, both of which receive effluent from houses and businesses on the western outskirts of the town of Consett. The 2 streams join about 2km upstream of Lanchester, and Smallhope Burn receives some storm drainage and urban run-off before the effluent from Lanchester STW (5,447 population equivalent) just above the confluence with the River Browney.

Upstream of the confluence with Smallhope Burn, the River Browney (GB103024077320) is classified as poor status due to the condition of the fish; invertebrates and all supporting elements are at high status. Downstream of the confluence with Smallhope Burn (GB103024077551), the river is moderate status, again due to the condition of the fish; however, phosphorus drops to poor status. The phosphorus failures, combined with the lack of data for phytobenthos, provided the rationale for this particular study.

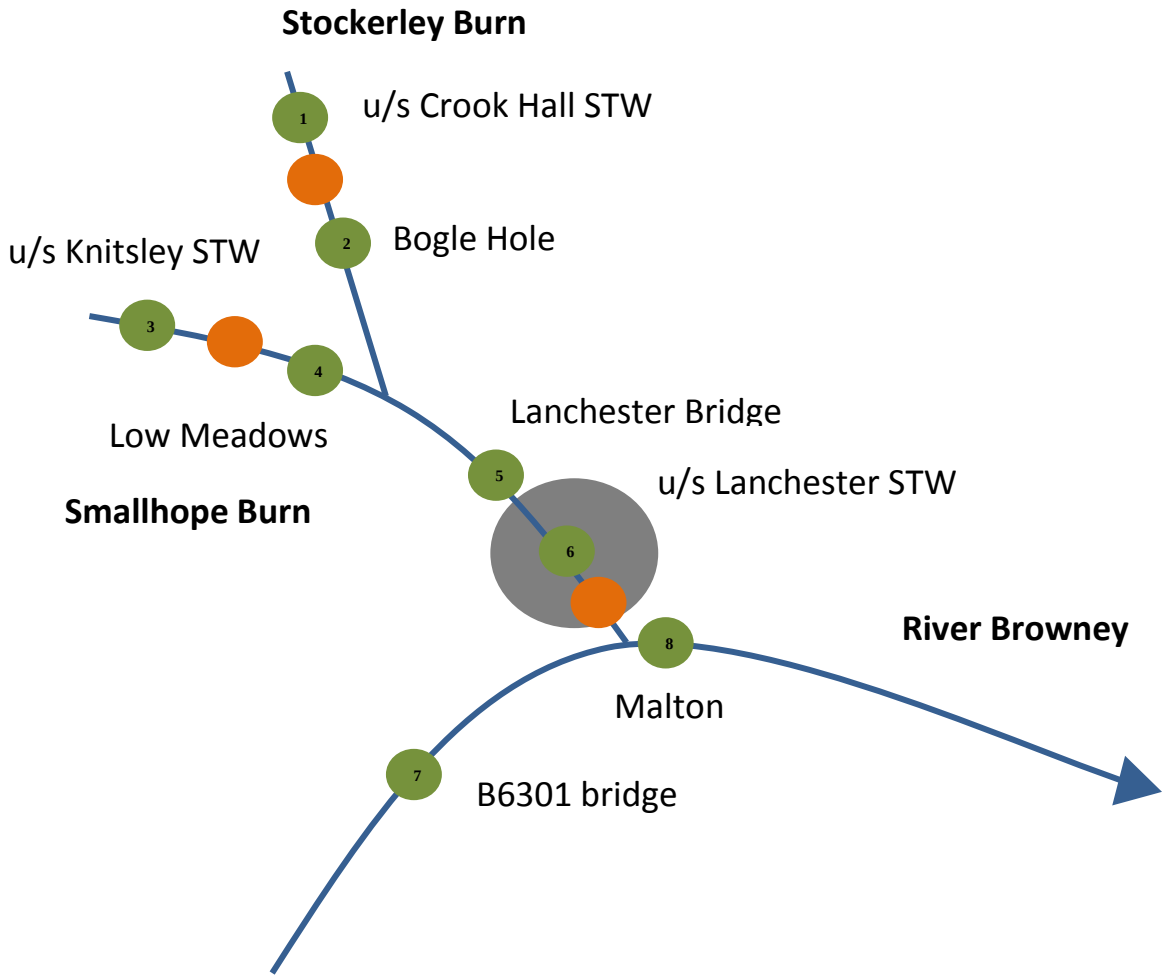


Figure 8.1 Schematic map of the upper River Browney and tributaries showing the location of STWs (orange circles), sampling sites (green circles) and the town of Lanchester (grey circle)

8.2 Methods

8.2.1 Study design

The locations and characteristics of the sites are listed in Table 8.1. U/s Crook Hall and Bogle Bridge bracket Crook Hall STW, while Knitsley Bridge and Low Meadows bracket Knitsley STW. Lanchester Bridge and u/s Lanchester STW examine the impact of the built-up area around Lanchester. Effluent from Lanchester STW enters Smallhope Burn close to the confluence of the River Browney and the sites at the B6301 bridge and Malton allow the effect of this to be assessed (Figure 8.1).

Samples were collected in summer 2014, autumn 2014 and winter 2015; samples were also collected in spring 2014 but these could not be analysed by NGS. Water chemistry for the period under consideration was obtained from the Environment Agency.

Table 8.1 Locations and characteristics of sites visited during investigation of the River Browney subcatchments

Site	Water body/site	NGR	Altitude (m)	Alkalinity (mg $CaCO_3$ per litre)
	Stockerley Burn			
1	u/s Crook Hall STW	NZ 122 508	234	224.2
2	Bogle Bridge	NZ 132 502	175	91.4
	Smallhope Burn			
3	Knitsley Bridge	NZ 121 483	150	82.4
4	Low Meadows	NZ 151 482	120	73.2
5	Lanchester Bridge	NZ 165 479	115	73.2
6	u/s Lanchester STW	NZ 174 467	110	113.6
	River Browney			
7	B6301 bridge	NZ 166 463	105	77.4
8	Malton	NZ 178 464	100	96.2

8.2.2 Statistical analyses

The approach to statistical analyses was similar to that outlined in Section 6.2.2. But because there were a large number of spatial and temporal samples from a limited area with a relatively short gradient of ecological diversity (see below), only limited use was made of multivariate analyses as these might accentuate the importance of relatively small differences, leading to a risk of over-interpretation of the data.

Phosphorus concentrations likely to support different ecological status classes at each site were calculated following UKTAG (2013). The median value of predictions was plotted to make Figure 8.2 easier to read; the full range of predictions for the 8 sites are as follows:

- good status: 0.040–0.054 mg L^{-1}
- moderate status: 0.115–0.143 mg L^{-1}
- poor status: 0.845–0.927 mg L^{-1}

There are no UK standards for nitrate concentrations likely to support good ecological status. However, the Republic of Ireland threshold for good status of 1.8 mg nitrate-N per litre provides an approximate indication of the state of the river with respect to this nutrient.

8.3 Results

8.3.1 Water chemistry

The effect of the STWs at Crook Hall (between sites 1 and 2) and Knitsley (between sites 3 and 4) is clearly shown in the increase in phosphorus concentrations between these sites (Figure 8.2), as is the effect of the confluence of Smallhope Burn (including Lanchester STW as well as the upstream works) on the River Browney (between sites 7 and 8). The upstream locations at Stockerley Beck (site 1), Smallhope Burn (site 3) and the River Browney (site 7) have phosphorus concentrations likely to support good ecological status; however, there is significant enrichment at all the downstream sites. In Smallhope Burn and the River Browney, concentrations are unlikely to support ecology above moderate status while Stockerley Beck has very high concentrations, unlikely to support ecology above poor status. Stockerley Beck (bearing effluent from Crook Hall), however, appears to have little additional impact on Smallhope Burn downstream of the confluence (between sites 5 and 6). Similar patterns are shown by nitrate-N (Figure 8.3), though increases downstream of STWs are not so pronounced and, in addition, both Smallhope Burn (site 3) and the River Browney (site 7) show signs of enrichment upstream of any major point source inputs.

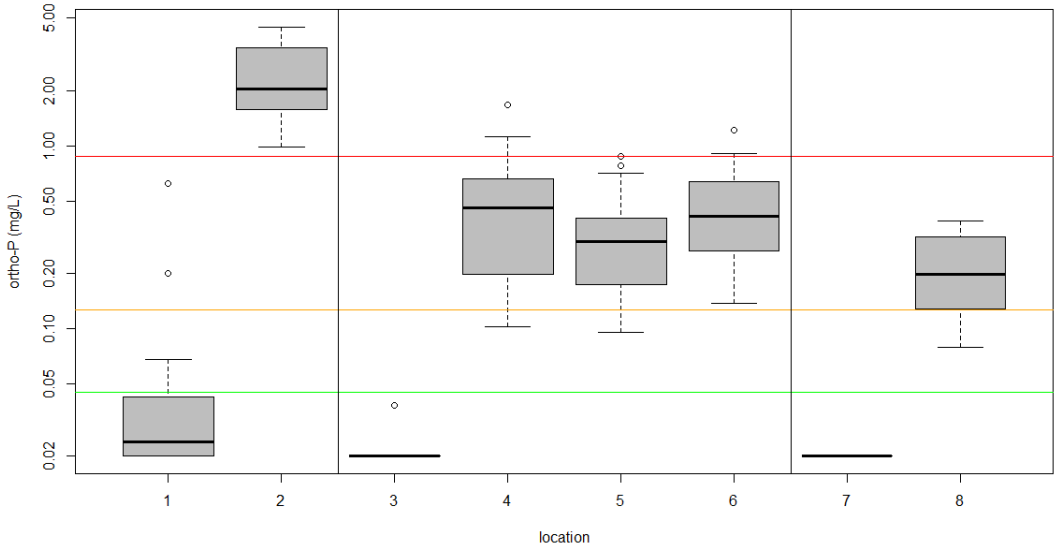


Figure 8.2 Variation in reactive phosphorus in Stockerley and Smallhope Burns and the upper River Browney

Notes: Boxplots summarise data collected between 2012 and 2014.
Horizontal lines show the median site-specific predictions for boundaries between good and moderate status (green), moderate and poor status (orange), and poor and bad status (red).
Sites 1 and 2: Stockerley Beck, u/s Crook Hall STW and Bogle Hole;
Sites 3–6: u/s Knitseley STW, Low Meadows, Lanchester Bridge and u/s Lanchester STW on Smallhope Burn
Sites 7 and 8: B6301 bridge and Malton on River Browney

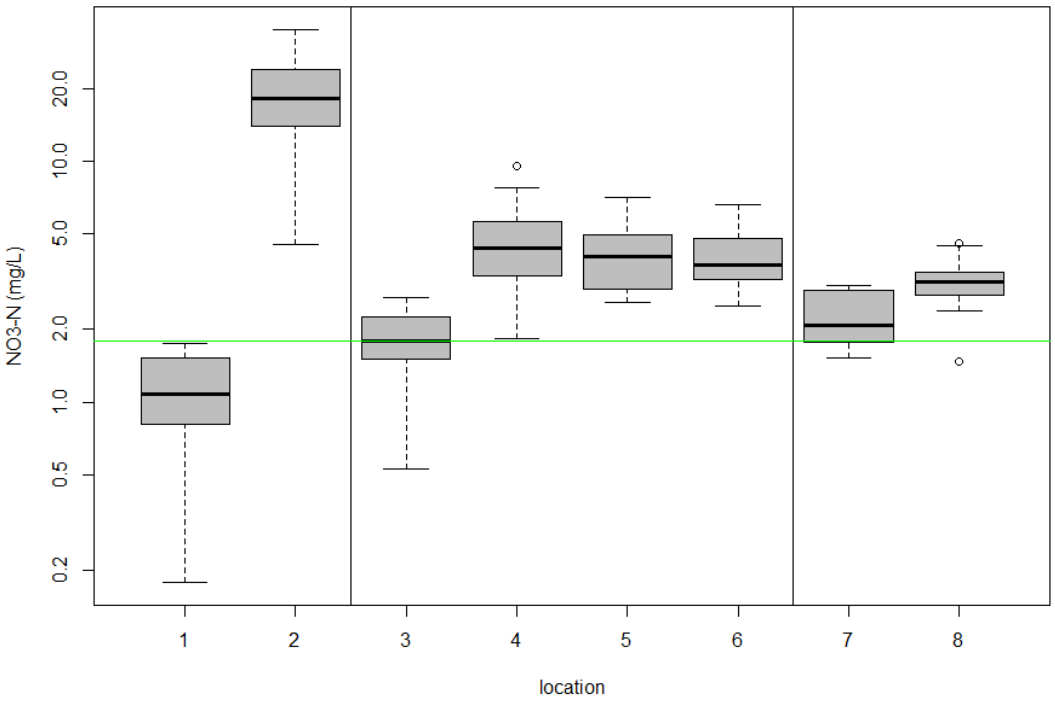


Figure 8.3 Variation in nitrate-N in Stockerley and Smallhope Burns and the upper River Browney

Notes: Boxplots summarise data collected between 2012 and 2014.
Horizontal line: Republic of Ireland standard for nitrate-N concentrations likely to support good ecological status.
Sites 1 and 2: Stockerley Beck, u/s Crook Hall STW and Bogle Hole
Sites 3–6: u/s Knitsley STW, Low Meadows, Lanchester Bridge and u/s Lanchester STW on Smallhope Burn
Sites 7 and 8: B6301 bridge and Malton on River Browney

8.3.2 Diatom analysis: LM and NGS

The expectation, based on the results presented in Section 6, is that there should be a close relationship between TDI4 (LM) and TDI5 (NGS). Although this is the case for most samples (Figure 8.4), there are 4 samples where TDI4 is much higher than TDI5, all of which have low numbers of sequence reads (ranging from 100 to 180) compared with the other NGS samples. Samples with sequence reads less than 3,000 would normally fail quality control and be repeated. Here they have been excluded from further analyses. When the low read samples are removed, the correlation between the 2 approaches becomes highly significant ($r = 0.753$, $p < 0.001$).

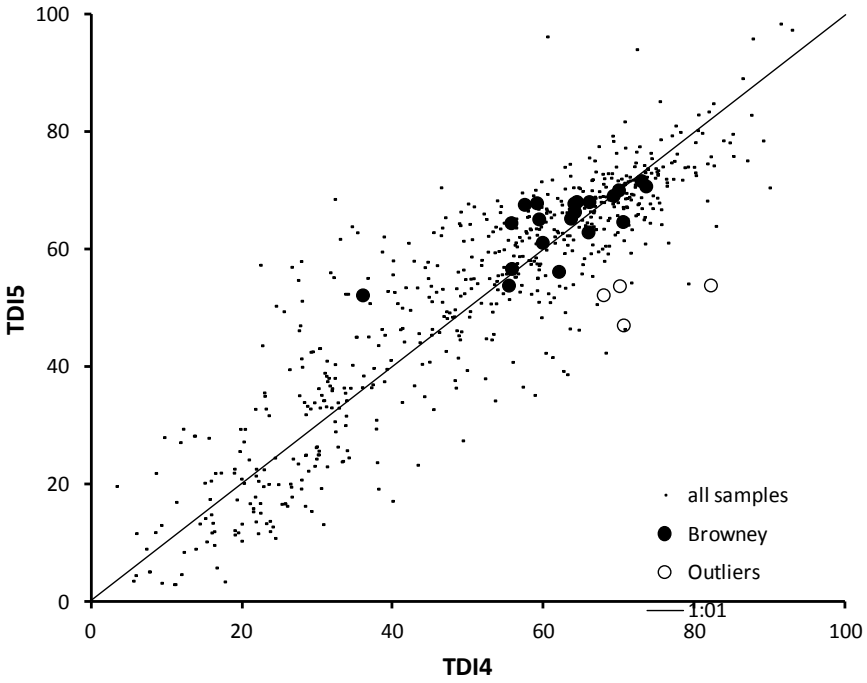


Figure 8.4 Relationship between TDI4 (LM) and TDI5 (NGS) with sites from River Browney subcatchments overlain

Notes: Open circles show samples from the Browney catchment that are outliers.
Diagonal line indicates slope = 1

Figure 8.5a and Figure 8.5b show the difference calculated for the 8 sites for LM and NGS respectively.

Based on the phosphorus data (Figure 8.2), an increase in TDI would be expected between sites 1 and 2, 3 and 4 and 7 and 8. Diatoms analysed by both LM and NGS do not appear to pick up the effect of phosphorous downstream of Knitsley STW on Smallhope Burn (between sites 3 and 4) nor below sites 7 and 8 where Smallhope Burn, bearing the effluent from Lanchester STWs joins the River Browney, although TDI values are very similar between the 2 methods.

The effect of Crook Hall STW on Stockerley Beck (between sites 1 and 2) is picked up by a slight increase in TDI using LM, but the change is not mirrored by the NGS data.

The difference between LM and NGS is apparent at site 1, where a higher than expected TDI5 value is observed. Figure 8.6 shows the difference in composition between the LM and NGS outputs for samples from site 1 for those taxa with RA >5%. The reasons for the higher than expected TDI5 at this site are likely due to the influence of lower numbers of *Achnanthidium minutissimum* recorded using NGS (common in streams with low to moderate concentrations of nutrients) compared with LM and higher proportions of taxa such as *Navicula lanceolata* whose ecological spectra extend into more enriched conditions. On one occasion, *Melosira varians* was recorded by NGS but not by LM. The up weighting of *A. minutissimum* and down weighting of *M. varians* and *N. lanceolata* in TDI5 (Table 6.1) should have accounted for some of the effect of these taxa but, clearly, the impact is still apparent in this instance.

The mismatch between diatoms and water chemistry is most acute at site 7 (B6301 bridge on the River Browney). This site is surrounded by farmland and nitrate-N concentrations appear to be slightly elevated (Figure 8.3). One possible explanation is that the monthly analyses of unfiltered reactive phosphorus is underestimating the true

bioavailable phosphorus load, perhaps reflecting pulses associated with rainfall. However, this is beyond the scope of the current study, which focuses on differences between LM and NGS analyses.

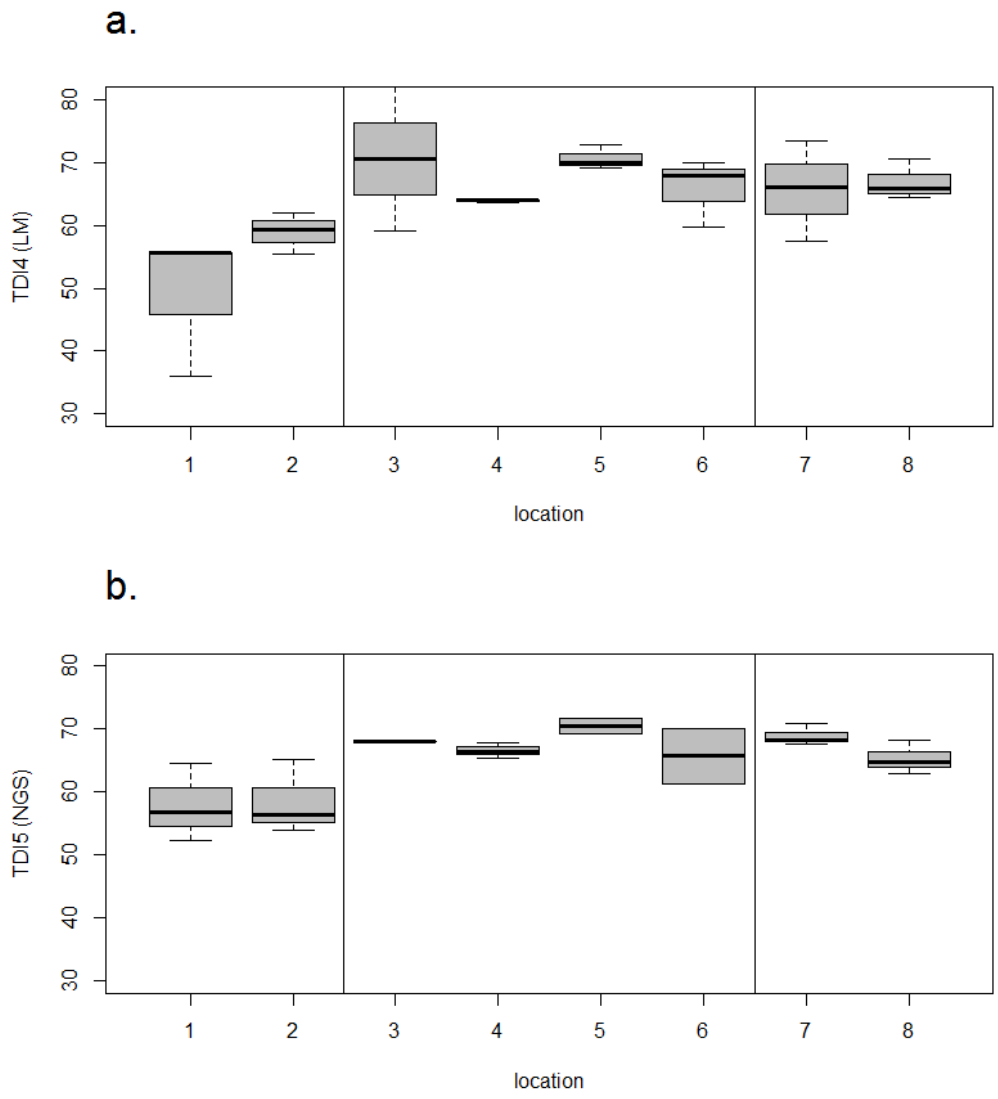


Figure 8.5 Variation inTDI4 (a) and TDI5 (b) in the Stockerley and Smallhope Burns and the upper River Browney

Notes: Boxplots summarise 3 seasonal samples collected between summer 2014 and winter 2015.
Samples with low numbers of sequences have been removed from the TDI5 plot.
All boxes are based on *n* = 3 samples, with the exception of TDI5 data for site 3 (2 outliers purged; *n* = 1) and sites 5 and 6 (*n* = 2).

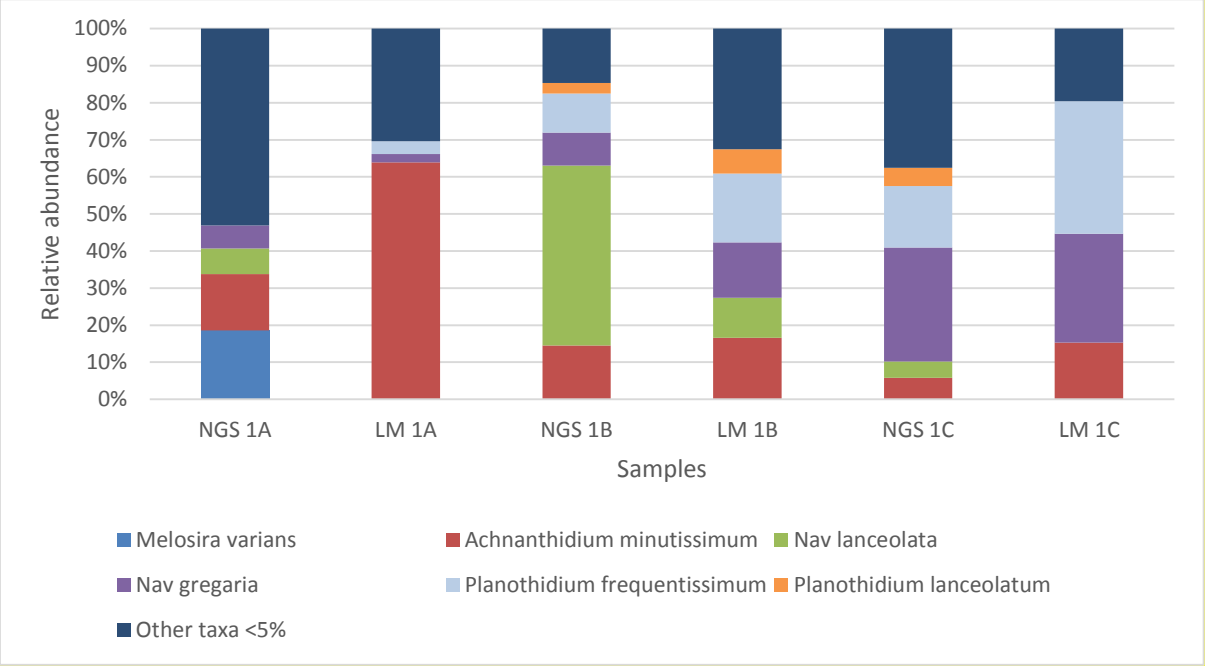


Figure 8.6 Variation in composition of taxa at site 1 between LM and NGS samples

Notes: Bar chart shows the distribution of taxa from 3 seasonal samples collected between summer 2014 and winter 2015 (A = summer, 2014 B = autumn, 2014 and C = winter, 2015).

8.4 Discussion

The objective of this case study was to apply the NGS method to a real life investigation and compare the outcome of NGS with LM to understand its response to pressures – in this case phosphorus and the impact of STWs. The upper River Browney and its subcatchments are of ongoing interest to local Environment Agency staff and therefore represented an appropriate 'real-time' test of the NGS method in a catchment whose general features are well known to the study team.

In terms of the biology, the correlation between the TDI values for the 2 approaches was highly significant ($r = 0.753$, $p < 0.001$), demonstrating the close agreement between the 2 methods. However, there were mismatches between the chemistry and the diatom results obtained using both LM and NGS. A possible explanation is that upstream locations are set amid productive farmland and the low levels of phosphorus recorded by routine chemical sampling may underestimate the short-term pulses of nutrients associated with high flow events which the diatoms are responding to (see, for example, Snell et al. 2014). Unfortunately, this has represented a 'step into the unknown' for the NGS method insofar as the details of inputs and influences on the biota still need to be unravelled. Further work is therefore needed at a small scale to understand the relationship between NGS and LM data and the response to phosphorus. This needs to be carried out on a catchment where the biology and water chemistry are fully understood.

9 Discussion

9.1 Introduction

This project is the first large-scale proof of concept to establish the suitability of combining rbcL DNA barcoding with NGS (metabarcoding) for species identification and RA estimates of diatoms in rivers. Significant correlation between the current LM TDI4 and the recalibrated NGS based TDI5 has been demonstrated, despite an incomplete rbcL DNA barcode reference database. There have been limited demonstrations in the past (Kermarrec et al. 2014, Visco et al. 2015, Zimmerman et al. 2014); of these Visco et al. (2015) achieved quantification, targeting the 18S fragment on a smaller dataset and with lower agreement with LM than was achieved in this project. Other studies using NGS profiling of eukaryotes have successfully demonstrated links between environmental habitat heterogeneity and molecular sequencing patterns (Lallias et al. 2015). These studies reveal the growing body of evidence to support the use of NGS, not only for profiling diatom assemblages, but also for other biological taxa including vertebrates (Hänfling et al. 2016), indicating the potential for this technology to generate data that can be used in ecological assessments.

Though now well-established as part of the ecological assessment toolkit in Europe and beyond (Kelly 2013, Poikane et al. 2016), diatom analysis currently requires highly trained individuals to spend considerable lengths of time with microscopes. There are a number of uncertainties associated with LM assessments (Prygiel et al. 2002, Kelly et al. 2009), a significant part of which is associated with the analytical process itself (Kahlert et al. 2012, Kahlert et al. 2016). There is therefore a strong case for exploring alternative technology, with potential for greater specificity and which may be more suited to large-scale assessment. This study has demonstrated that NGS is one alternative that shows great promise.

It is also important to remember that current methods based on LM are also, to some extent, artificial. The use of cleaned diatom slides offers benefits in greater taxonomic sensitivity, but at the expense of losing information about non-diatom algae (an important component of many biofilms) as well as extracellular structures such as stalks and tubes, and about which individuals of which species were alive at the time of sampling. Moreover, methods for LM data analysis focus on enumeration of individuals, regardless of cell size. There can be, for example, a 100× difference in the biovolume of a single cell of *Achnanthidium minutissimum*, compared with one of *Ulnaria ulna*, yet both have equal influence on a TDI calculation.

Some authors have advocated abandoning traditional taxonomic approaches (for example, Baird and Hajibabaei 2012, Woodward et al. 2013). Although agreeing that there is great potential with NGS approaches to explore aspects of biodiversity and ecosystem function that are difficult to measure using traditional taxonomic approaches, establishing that NGS can provide comparable information to existing methods is an important first step. As current ecological classifications are based in part on assessments made of diatom assemblages, it is important that new methods are compared with methods currently accepted by the regulatory bodies. Having an established baseline also ensures that OTUs are grounded in reality. One aspect of trying to understand the relationship between LM and NGS involved the painstaking task of comparing OTUs with binomials (see Section A1.2.2 in Appendix 1) and visually interrogating phylogenetic trees. This revealed nomenclature issues between databases and cryptic diversity within complexes and identified algal contaminants that would have been missed had the more radical approaches proposed by Baird and Hajibabaei (2012) been adopted.

Having established that there is a significant correlation between the NGS approach and the existing diatom assessment method (chapter 6), albeit with some caveats (Sections 7 and 8), it is now possible to begin to consider how to provide added value contained within the NGS data, exploiting the intrinsic information on diversity using OTU information in combination with species assignments. So long as these metrics can be linked to legislative drivers such as the Water Framework Directive, then an NGS metric should be effective.

NGS is a rapidly emerging field, and unlike other molecular analysis techniques such as PCR, the development of platforms to generate more data at ever decreasing cost continues. To put this into context, prices of instruments and sequencing runs have both dropped by 10 times over the last 5 years, yet the amount of data generated per run has increased 30-fold. This offers the future potential that a method based on NGS will continue to decrease in price, whereas the price of analysis of methods based on microscopy has not changed for many years.

The potential for the use of NGS in ecological assessment extends beyond diatoms. This project has established several general principles of relevance to projects examining the potential of NGS in other spheres of ecological assessment. These include:

- the value of looking critically at barcode length
- the importance of a comprehensive barcode database
- how to handle taxa not included in the barcode database
- issues associated with quantification
- understanding the relationship between NGS and 'traditional' approaches

It is particularly important to approach the latter point with an open mind. While it is not in doubt that differences exist between LM and NGS approaches for analysing diatoms, it is important to bear in mind that the 'traditional' LM approach is, itself, an imperfect reflection of reality (albeit one with which practitioners are familiar). The 2 approaches offer alternative views of the stream ecosystem that need to be reconciled; it is rarely as simple as deciding that one method is 'right' or that it is 'better' than the alternative.

9.2 Development of rbcL barcode and bioinformatics

The identification and development of robust taxonomic markers is not trivial. Accurate species identification is a fundamental criterion (Hebert et al. 2003) for the application of a taxonomic marker for molecular detection. Furthermore, features should include the universality across the taxa of interest, the reflection of evolution of the studied species and ideally low variation in copy number across the taxa of interest (Chase et al. 2007).

With the advent of NGS, an additional characteristic needs to apply for a marker to be suitable for high-throughput sequencing, that is, its length should be short to fit currently available sequencing platforms (Kress and Erickson 2008). In this project, after a suitable barcode (rbcL) was chosen, a major challenge was to identify an informative region of the rbcL gene satisfying the criterion of length without losing its taxonomic resolution. To the project team's knowledge, this is the first report of the use of this region of rbcL allowing a robust metabarcoding strategy due to its compatibility with Illumina technology, the current market leader in this area.

More specifically, during this project a short barcode region was developed which simultaneously enabled a DNA fragment from a large number of diatom taxa to be

amplified while retaining a sufficient number of informative nucleotide positions to allow discrimination. The aim was to take advantage of the wide availability of short-read sequencers such as the MiSeq (Illumina), enabling the production of data at a cost that allows the technique to become useable for routine monitoring by regulatory agencies.

The use of a short barcode is a pragmatic one and does not provide the taxonomic resolution of the full length rbcL barcode. However, it offers both good resolution and cost-effectiveness. In having this balance, there is a risk that in a few cases ecologically differentiated taxa may not be separated by the short barcode sequence, although no such cases have yet been detected. Moreover, the barcode reference database contains full length barcode sequence data and the analysis pipelines will enable analysis of these data should longer read length sequencing become a viable proposition in the future; see, for example, the MinION sequencing device developed by Oxford Nanopore Technologies (https://nanoporetech.com/products/minion).

Further work could be carried out to refine the bioinformatic pipeline to improve the accuracy of taxonomic assignments. Chimeric sequences, which can occur during PCR amplification and result in sequences that may be partly one species and partly another, are a known PCR artefact in amplicon metagenomic studies and were not screened out in this project. Although software is available to detect chimeric sequences (Edgar 2010, a drawback of taxonomy-free methods of detection can be a high false positive rate (Haas et al. 2011) and the subsequent removal of informative sequences. More recently, chimera detection for metagenomic studies has moved towards reference-based detection methods which require the use of curated sequence databases known to be free of chimeric DNA barcodes (Nilsson et al. 2015).

It is possible, but rather long-winded, to screen the DNA barcode database to assess whether the DNA barcodes are themselves chimeric sequences. This screening process would be an essential step before the introduction of a chimera checking step into the current pipeline and could be carried out as future work. However, given that ultimately the taxonomic abundance data >2% produced during the pipeline is converted to a TDI value, it is unlikely that a small number of chimeric sequences would have an impact on the overall TDI of a sample.

Following the comparison of LM and NGS datasets in Section 6.3.2, further work has been carried out to refine the bioinformatics pipeline to:

- increase the taxa assignment threshold from 90% to 95%
- constrain the analysis to only assign a taxa identity to sequences that are present in the barcode reference library and thus bypass searches in GenBank

Given that the initial pipeline assigned a sequence to a taxon when sequence identity was above 90%, the effect of increasing this threshold to 95% was assessed. With a low 90% threshold, very few OTUs are left unknown or searched against GenBank, meaning that identifications that do not have a good sequence similarity match are potentially being made erroneously. Additional work in this area has shown that, should the threshold be increased to 95%, taxonomy would still be assigned to approximately 75% of each sample (Figure 5.5). In this scenario, the remaining 25% of sequences would be left as ‘unknown’, rather than being assigned an identity, which should produce a more accurate NGS TDI5 than is the case for the results presented in this report. While identifications are required to mirror the current LM method, NGS barcodes can provide a higher level of resolution which may be useful in identifying new taxa that have not yet been described, as well as cryptic and semi-cryptic variation within established taxa which may have ecological value. Any potential new taxa emerging from the ‘unknown’ sequences could be included in the diatom database –

without any taxonomic identification – allowing them to be tracked and identified in other water bodies.

The analysis has also been simplified. Figure 6.10 demonstrated that the diatom species present in the barcode reference database provided a good predictor of TDI. This allowed the analysis to be constrained so that it bypasses GenBank and identifies only sequences within the sample that can be linked to sequences present in the barcode reference database. In addition, GenBank comes with various sources of error and hence sequences submitted and assigned a taxa identify may not always be correct. In addition, the constrained pathway is also computationally faster than the original approach.

The data files produced during the NGS based approach are not prohibitively large (~5Gb per Illumina run of 200 samples) and can be compressed for long-term storage by the Environment Agency. This opens up the potential for a wide range of retrospective studies in the future with the sequence data produced during routine monitoring and with the dataset already archived from this study.

9.3 What was learnt from development of the barcode database?

Correct assignment of NGS data to the appropriate Linnaean binomial is of prime importance to the development of a viable NGS based ecological assessment procedure. The situation for diatoms is complicated by the number of new developments in underlying taxonomy, many of which are, themselves, driven by the insights that molecular biology has provided. In some cases, these insights clarify differences between species that present challenges to traditional analyses (Rovira et al. 2015) which, in turn, allow ecological differences to be unravelled (Kelly et al. 2015). In other cases, such studies throw doubt on species defined on morphological criteria alone (Kermarrec et al. 2013, Rovira et al. 2015, Duleba et al. 2016).

The barcode database at the heart of this project contains sequences from 176 species (at the time of writing the number is increasing through the incorporation of additional barcodes becoming available through trusted online databases). A substantial amount of effort went into the development of this database, which still represents less than 10% of the total number of UK diatom species recorded from British and Irish freshwaters. However, this list does include representatives of most of the commonly encountered taxa and is sufficient to account for most of the variation in TDI analyses (Figure 6.10).

There is, nonetheless, no cause for complacency. Inferences based on a nationwide dataset can look less impressive when differences within small geographical areas are examined, and where the absence of a key taxon may influence the sensitivity of the index. Although the number of quantitatively important taxa that are not represented in the database is small (Sections 6.3.1 and 6.3.2), the situation is complicated because several species are known or suspected to be complexes. Furthermore, phylogenetic analyses of diatoms (for example, Rovira et al. 2015) suggested that the rbcL gene evolves more rapidly in some lineages than in others (for example, more rapidly in *Nitzschia* group II than in group I in the study by Rovira and colleagues), potentially biasing NGS data when the same stringency threshold is applied throughout. The same phenomenon has also been observed with other genes or combinations of genes (for example, see the behaviour of Rhabdonema, Striatella, Florella and Astrosyne in the three-gene tree of Lobban and Ashworth 2014).

Ideally, species should be represented in a barcode database by a series of strains exhibiting the full range of genetic variation. Otherwise, potential differences between

LM and NGS outcomes will be accentuated, although the NGS analysis performed identified OTUs as clusters of sequences rather than sequences that are identical to sequences present in the taxon database. Different rates of molecular evolution also illustrate that no single stringency threshold will perform equally in all groups of diatom, in terms of separating closely related species.

One way to increase coverage of the barcode database would be to continue the approach adopted here, sequencing more strains and linking them to the appropriate Linnaean binomial. This may be 'best practice' (Zimmermann et al. 2014), but it is also expensive and depends on being able to select and grow unialgal strains of a wide range of target species. Two alternatives are to infer barcodes directly from comparisons between LM and NGS data, as demonstrated, for example, in this project for *Achnanthes oblongella* (Section A1.2.4) or to adjust the bioinformatics pathways to enable unassigned OTUs to be curated at an appropriate taxonomic level and linked to an appropriate binomial at a later date.

All of these approaches assume a continuing relevance for Linnaean binomials. In practice, these provide a series of a priori categories to which entities identified by either LM or NGS are assigned. Each of these categories can then be linked to autecological information, from which the final status assessment is derived. The assumption is that the information associated with each binomial adds substantial value to the assessment outcome. In theory, a system based purely on OTUs (that is, bypassing Linnaean binomials completely) could work as efficiently, once it had been calibrated against the principal environmental gradients.

As a result of the work in the present study and elsewhere, some practical issues that need to be taken into account in the development of any diatom barcode database have been identified. These are as follows.

- The commonly used freshwater media (for example, Guillard and Lorenzen's WC medium) are themselves selective, giving rather poor results with species from acid oligotrophic waters.
- Even when a range of media are employed, some species may still remain refractory in culture. In these, amplification from single cells may provide reference sequences and allow culturing to be bypassed, but it may be difficult or impossible to provide adequate voucher specimens to document the morphology of the organism that has been barcoded.
- Efficient isolation of a variety of targeted diatom species requires a very unusual combination of dexterity and detailed knowledge of diatom morphology and cytology, as well as an understanding of their ecological preferences.
- Given such a highly skilled culturist, the time and effort spent in isolating and culturing is small relative to that needed for harvesting and the preparation and documentation (including photography) of voucher specimens.

9.4 Relationship of NGS with LM approach

This project has gone further than any other projects in demonstrating that a full NG based analogue of existing ecological assessment methods is possible. In particular, the project has demonstrated that it is possible to achieve semi-quantitative outcomes from NGS. The initial choice of the rbcL gene proved fortuitous in this respect, as there is a predictable relationship between the number of individuals and the number of

reads. Interpretation of this relationship is, however, complicated for the following reasons.

- The number of rbcL reads per cell appears to be influenced by the number of chloroplasts. Although there have been no studies specifically focused on diatom chloroplasts, it is likely that copy number per chloroplast and per cell will vary between species, and between different cells (in different environmental conditions or developmental stages) of the same species (Rauwolf et al. 2010). However, each species will probably vary only within certain limits and these limits will differ from those in other species, Figure 6.3 does suggest that the relationship between the RA of sequences is at least partly a consequence of the number of chloroplasts. In many taxa there is one or two chloroplasts per cell; in a few species, however, there are many and these taxa (in particular, *Melosira varians*) tend to dominate the NGS output. In a small number of genera, the number of chloroplasts is not known.

- Traditional LM does not record the number of cells, but rather the number of valves (= half a cell wall, or 'frustules'). In very small diatoms, it can be difficult to determine whether a single valve or complete frustules are present. (NB In some countries, single valves and intact frustules are not differentiated during analyses.)

- The relationship between LM and NGS for any particular taxon has to be determined in a mixture of (typically) 20 or more species; the proportion of species A in NGS and LM, for example, will also be influenced by fluctuations in the proportion of species B, C, D and so on.

Nonetheless, a good correlation was seen between LM and NGS data (Figure 6.7). The only other study that has achieved quantification (Visco et al. 2015, using 18S) showed a relationship with a similar statistical strength which also deviated from 1:1. Samples with taxa with multiple chloroplasts proved to be particularly troublesome in this study, as a few taxa (*Melosira varians, Cyclotella meneghiniana* and *Diatoma vulgare*) could dominate the rbcL output while being present in relatively low numbers in LM data. Furthermore, a few weakly silicified taxa (for example, *Fistulifera saprophila*) were more common in the NGS output than in LM, possibly due to dissolution in the aggressive oxidising mixtures used to prepare samples for LM (Zgrundo et al. 2013). It should not be a surprise, therefore, that simply applying a metric designed for LM data to NGS data did not result in a strong 1:1 fit (Figure 6.7a). Even after new coefficients were derived to calibrate a NGS specific diatom metric, a few taxa required additional weightings to optimise the fit between LM and NGS specific variants of the TDI.

Having a basic metric that captures the dominant nutrient/organic gradient, it is then relatively straightforward to calibrate this against 'expected' values, following the same procedures used to develop the current method (Kelly et al. 2008, Environment Agency 2013). The outcome shows good, though not perfect agreement, suggesting that continuity with existing classifications should be achieved.

The broad spatial relationship established in Section 6 is examined in more detail in Sections 7 and 8, which focus on spatial and temporal variation at different scales, and within the context of investigations as part of Programmes of Measures. Section 7 suggests that there will be 'gains' in terms of greater analytical precision from the NGS method. However, spatial and temporal variation within a water body was, in most cases, greater than the analytical variation for both LM and NGS. These sources of uncertainty are important for determining Confidence of Class and Risk of Misclassification (Clarke 2013, Kelly et al. 2009). The relative scale of this variation in LM and NGS varied from stream to stream, with NGS showing consistently lower

variability only in the River Ehen. Similarly, higher variability of NGS compared with LM was observed in the River Team, and further work to understand the performance of the NGS method in highly polluted rivers is currently underway. Overall, however, there seems to be little or no likelihood of a major gain in overall precision in status assessments as a result of a shift to NGS.

Similar comments apply to the study of the upper River Browney and subcatchments. This is a catchment of ongoing interest to local Environment Agency staff. As such, it represented a 'step into the unknown' for the method. Although the variability between LM and NGS fell within the expected range (Figure 8.6), the presence of outliers – due to failure in the NGS analysis for some samples – amid otherwise well-correlated data suggests further work is needed to understand the relationship between NGS and LM data at a smaller scale.

9.5 Conclusions

Overall, the outcomes from this study are positive: a procedure has been developed that is compatible with the latest high-throughput NGS technologies and successfully correlates with the current LM method. Protocols for collecting, preserving and storing samples for NGS analyses have been modified from existing methods. Procedures for extracting, amplifying and analysing DNA sequences in these samples have been developed and tested, and automated bioinformatics procedures have been devised to produce data that are compatible with outputs from current LM analyses. This, in turn, has allowed the similarities and differences between the 2 approaches to be evaluated and, from this point, a new metric – a variant of the current TDI (TDI4) – optimised for NGS (TDI5) to be developed.

This is remarkable given that it has been achieved using a barcode database that includes less than 10% of the diatom species that have been described from the UK. As more laboratories contribute barcodes to online databases, the method will continue to improve. However, it is unlikely that full comprehensive coverage of all diatom species will be achieved at a sufficiently high quality in the near future due to issues with the isolation and culturing of some diatom species. This is an area ripe for international collaboration. However, there is potential for exploring parallel approaches to document taxa without the need for culturing and sequencing from pure cultures, particularly as understanding of the species concept in diatoms continues to evolve (Mann 1999, Mann 2010).

When the variation within water bodies is studied in greater detail, however, the picture is not always so clear. In most cases, the levels of variation encountered were similar to those experienced in LM based studies. Where no consistent trend emerged (Figures 7.6 to 7.8), this can probably be explained by a combination of in-stream processes working at a variety of spatial and temporal scales, and issues with the post-NGS data handling such as handling of OTUs that cannot be assigned reliably and the weighting applied to multiple chloroplast taxa that are still being explored.

The collection of large sequencing datasets using NGS enables the possibility of future investigative analyses for the Environment Agency with regard to relatively simple multi-site and multi-year investigations using comparative metagenomics approaches developed by microbial ecologists. The future value of NGS sequencing datasets, such as those collected after implementation of this method, should not be underestimated. Such datasets provide a much larger opportunity for cost-effective large-scale 'big data' research and monitoring improvements that would not be possible with the current slide-based LM methods.

Finally, the project provides a template for how similar projects involving other organism groups and water body types could be organised. The aspiration of producing

NGS ‘mirrors’ of existing techniques is considered a sensible starting point, as it forces a close examination of the relationship between NGS and ‘traditional’ data. Once this has been achieved, however, the door can open to second generation methods that move beyond simplistic metrics and unlock the huge potential of NGS to evaluate ecosystem function in ways that can enhance assessments and, thereby, regulation and management (Sagarin et al. 2009).

9.6 Recommendations for further work

The following areas have been identified for further investigation or refinement prior to the method being implemented for classification of river water bodies.

Improve the utility of NGS outputs

Expand the barcode database

Although the overall performance of the method is good using the current barcode database and most of the variation within the diatom assemblages is being captured, further strengthening of the database will increase the resilience of the method, particularly in situations where samples are dominated by rare or unusual taxa. ‘Low frequency, high impact’ taxa whose absence may have a disproportionate effect on metric calculations should be targeted and barcodes obtained. In addition, the coverage – and understanding – of taxa suspected to be genetically diverse should be increased.

The current barcode library is largely the result of one year’s full-time effort by a postdoctoral researcher to culture and sequence diatoms. Additional sequences have been added from GenBank and other sources. This has ensured coverage of diatoms that are common and which grow easily in culture. As the barcode library grows in size, so the effort needed to plug gaps also increases, as particular taxa need to be targeted and cultured. There is also a risk that target taxa may not grow well in culture and cannot therefore be sequenced. There will be a continued need to add barcodes from online sources.

The addition of new sequences from cultures isolated from UK locations, where these are known to occur, should also continue. This will not be possible for every missing taxon, but it should be possible to:

1. Identify sites where a species is known to occur from existing records
2. Visit the site at a time when the species is known to be abundant
3. Culture biofilm samples to isolate the taxon in question
4. Sanger sequence the taxon to generate the barcode

Improvements to post-NGS data handling

Currently there is an understanding of how NGS and LM data differ – and recognition that the 2 sorts of data should not be regarded as equivalent in all respects. However, there is not a good understanding of why these difference exist.

The contribution made by rbcL reads from diatoms with multiple chloroplasts is thought to be a major factor. Additional data mining should be explored to test this hypothesis and to consider ways in which the accuracy of TDI outputs might be improved so that there can be greater confidence in the data.

Improvements to species identification by investigating the requirement for OTU clustering in the future

The creation of OTUs at 97% similarity from the raw NGS sequences can group very similar species together into one OTU. In the short to medium term, as improvements to computing power are realised, it may be possible to move from a computational power-saving OTU based analysis pipeline to one where each individual NGS sequence is analysed instead. This can be investigated by comparing datasets with and without OTU clustering, alongside a comparison between the current pipeline's BLAST identification of sequences versus machine learning classification systems.

Improve the coverage of poor and bad status classes to give a better overview of the method performance (see Table 6.2).

Care was taken to select sites that covered the full range of conditions encountered in England as part of the calibration dataset to develop the method. Despite this, the calibration dataset was biased towards high, good and moderate status when the final classifications were calculated. Therefore additional poor/bad status sites should be included in the calibration dataset to complete the ecological quality gradient.

Extend geographical coverage of the method to other parts of the UK

The work reported here is based largely on samples collected by the Environment Agency in England. Having established the performance of the method in England, further testing is required to ensure that the method is also applicable in Scotland, Wales and Northern Ireland.

Test the method on an independent dataset

In this project, TDI5 was developed and tested using a single dataset – the Environment Agency's 2014 sampling programme. Although bootstrapping was used to overcome the potential circularity of this process, generation of a new matched LM and NGS dataset would permit an independent test of the performance of the NGS method.

Test the method in real or simulated 'operational investigations' where there is good a priori evidence of a change in diatom assemblage composition within a short distance

The study reported in Section 8 attempted a real-time operational investigation using NGS where the relationship between chemistry and biology within the subcatchment studied was not clear ahead of the work. Effects were expected due to the location of point source inputs and the existing results from water quality and invertebrate analyses. However, there was a poor relationship between chemistry and (LM) diatoms, perhaps reflecting intermittent diffuse inputs missed by routine (monthly) chemistry. Low numbers of reads for some of the NGS outputs, which should have been detected and the extractions repeated, also reduced the number of data points available for analysis. A new study should be conducted based on sites around the UK where a relationship between chemistry and LM diatoms has already been established so that the performance of the NGS method can be evaluated without the complication of simultaneously trying to understand the relationship between pressures and biology in the catchments in question.

Evaluate the potential cross-contamination introduced by the current sampling method

NGS based analysis may be more sensitive than LM to low-level cross-contamination resulting from current sampling procedures. To ensure cross-contamination is minimised, an investigation into the efficacy of a cleaning step in the sampling process and the use of deionised or tap water for rinsing is required.

Test the transferability of the river method to lakes

The current method has been calibrated for rivers. The barcode database is likely to be important in determining the transferability of the method to lakes. Most of the common diatom species are found in both rivers and lakes, but there are a few that are more prolific in lakes. The current database has inadequate representation of, for example, Cymbella (and relatives), Denticula and Epithemia. There is also likely to be a stronger planktonic diatom signal from lake data, and it may be necessary to incorporate planktonic taxa in the barcode database in order to filter them out during bioinformatics analysis.

Although development of a lake method would require samples from lakes across the alkalinity and pressure gradients, a preliminary investigation using samples from England alone may give some insights into the scale of modification required to develop an operational lake assessment tool.

Consider the effect of method change on long-term dataset

There is often a need to maintain long-term datasets to track temporal change, which is particularly important for environment agencies in justifying and reporting on the efficacy of nutrient control measures in catchments. These datasets have been built on the results of the LM method, and there is a need to examine how NGS-computed TDI values relate to temporal trends of LM-derived TDIs, and consider reasons for any inconsistencies.

9.6.1 Preparation for implementation

Once the method has reached a stage where operational implementation by the relevant UK agencies is considered feasible and desirable, there will be a number of implementation issues to be considered. Details may be specific to each agency, depending on current systems in use, but will include:

- finalising the DNA barcode database*
- ensuring taxa have appropriate codes to allow input of NGS data to the agencies' data archive systems
- updating of classification software (DARLEQ) and associated guidance
- adopting NGS based assessment as a recognised UK method for Water Framework Directive classification and intercalibration of the method as required by the Water Framework Directive
- knowledge transfer and staff training in implementation of the new method

* Although taxa will continue to be added to the database over time, there is a need to determine a point at which a stable version is adopted for the purposes of an operational classification tool. This does not mean the taxa list is permanently fixed, but

future revisions would need to be considered in the context of the impact on classification results.

References

ANDERSON, M.J., 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26 (1), 32-46.

BAIRD, D.J. AND HAJIBABAEI, M., 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, 21 (8), 2039-2044.

BARNES, M.A., TURNER, C.R., JERDE, C.L., RENSHAW, M.A., CHADDERTON, W.L. AND LODGE, D.M., 2014. Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science and Technology*, 48 (3), 1819-1827.

BENNION, H., KELLY, M.G., JUGGINS, S., YALLOP, M.L., BURGESS, A., JAMIESON, B.J. AND KROKOWSKI, J., 2014. Assessment of ecological status in UK lakes using benthic diatoms. *Freshwater Science*, 33 (2), 639-654.

BIRKS, H.J.B., LINE, J.M., JUGGINS, S., STEVENSON, A.C. AND TER BRAAK, C.J.F., 1990. Diatoms and pH reconstruction. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 327 (1240), 263-278.

CAISOVÁ, L., MARIN, B. AND MELKONIAN, M., 2011. A close-up view on ITS2 evolution and speciation – a case study in the Ulvophyceae (Chlorophyta, Viridiplantae). *BMC Evolutionary Biology*, 11, 262.

CAPORASO, J.G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F.D., COSTELLO, E.K., FIERER, N., GONZALEZ PENA, A., 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. AND KNIGHT, R., 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7, 335-336.

CEN, 2014a. *EN 13946: 2014. Water quality – Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers*. Geneva: Comité European de Normalisation.

CEN, 2014b. *EN 14407:2014. Water quality – Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters.* Geneva: Comité European de Normalisation.

CHASE, M., COWAN, R., HOLLINGSWORTH, P., VAN DEN BERG, C., MADRIÑÁN, S., PETERSEN, G., SEBERG, O., JØRGSENSEN, T., CAMERON, K., CARINE, M., PEDERSEN, N., HEDDERSON, T., CONRAD, F., SALAZAR, G., RICHARDSON, J., HOLLINGSWORTH, M., BARRACLOUGH, T., KELLY, L. AND WILKINSON, M., 2007. A proposal for a standardised protocol to barcode all land plants. *Taxon*, 56 (2), 295-299.

CLARKE, R.T., 2013. Estimating confidence of European WFD ecological status class and WISER Bioassessment Uncertainty Guidance Software (WISERBUGS). *Hydrobiologia*, 704 (1), 39-56.

COLEMAN, A.W., 2009. Is there a molecular key to the level of 'biological species' in eukaryotes? A DNA guide. *Molecular Phylogenetics and Evolution*, 50 (1), 197-203.

DOWNES, B.J., BARMUTA, L.A., FAIRWEATHER, P.G., FAITH, D.P., KEOUGH, M.J., LAKE, P.S., MAPSTONE, B.D. AND QUINN, G.P., 2002. *Monitoring Ecological Impacts: Concepts and Practice In Flowing Waters.* Cambridge: Cambridge University Press.

DULEBA, M., KISS, K.T., FÖLDI, A., KOVÁCS, J., BOROJEVIC, K.K., MOLNÁR, L.F., PLENKOVIC-MORAJ, A., POHNER, Z., SOLAK, C.N., TÓTH, B. AND ÁCS, É., 2015., Morphological and genetic variability of assemblages of *Cyclotella ocellata* Pantocsek/*C. comensis* Grunow complex (Bacillariophyta, Thalassiosirales). *Diatom Research*, 30 (4), 283-306.

EDGAR, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26 (19), 2460-2461.

ELAND, L.E, DAVENPORT, R. AND MOTA, C.R., 2012. Evaulation of DNA extraction methods for freshwater eukaryotic microalgae. *Water Research*, 46, 5355-5364.

ENVIRONMENT AGENCY, 2011. *A review of molecular techniques for ecological monitoring*. Report SC090010. Bristol: Environment Agency.

ENVIRONMENT AGENCY, 2013.*The integration of macrophyte and phytobenthos surveys as a single biological quality element for the Water Framework Directive.* Report SC070034/T4. Bristol: Environment Agency.

EUROPEAN COMMISSION, 2008. Commission Decision of 30 October 2008 establishing, pursuant to Directive 2000/60/EC of the European Parliament and of the Council, the values of the Member State monitoring system classifications as a result of the intercalibration exercise. *Official Journal of the European Union*, L 332, 10.12.2008, 20-44.

EUROPEAN COMMISSION, 2013. Commission Decision of 20 September 2013 establishing, pursuant to Directive 2000/60/EC of the European Parliament and of the Council, the values of the Member State monitoring system classifications as a result of the intercalibration exercise and repealing Decision 2008/915/EC. *Official Journal of the European Union*, L 266, 8.10.2013, 1-47.

EVANS, K.M., WORTLEY, A.H. AND MANN, D.G., 2007. An assessment of potential diatom 'barcode' genes, cox1, rbcL, 18S and ITS rDNA. and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist*, 158 (3), 349-364.

FAWLEY, M.W. AND FAWLEY, K.P., 2004. A simple and rapid technique for the isolation of DNA from microalgae. *Journal of Phycology*, 40 (1), 223-224.

FOURTANIER, E. AND KOCIOLEK, J.P., 1999. Catalogue of the diatom genera. *Diatom Research*, 14 (1), 1-190.

GUILLARD, R.R.L. AND LORENZEN, C.J., 1972. Yellow-green algae with chlorophyllide c. *Journal of Phycology*, 8 (1), 10-14.

HAAS, B.J., GEVERS, D., EARL, A.M., FELDGARDEN, M., WARD, D.V., GIANNOUKOS, G., CIULLA, D., TABBAA, D., HIGHLANDER, S.K., SODERGREN, E., METHE, B., DESANTIS, T.Z., THE HUMAN MICROBIOME CONSORTIUM, PETROSINO, J.F., KNIGHT, R. AND BIRREN, B.W., 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, 21 (3), 494-504.

HÄNFLING, B., LAWSON HANDLEY, L., READ, D. S., HAHN, C., LI, J., NICHOLS, P., BLACKMAN, R.C., OLIVER, A. AND WINFIELD, I.J., 2016. Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, 25 (13), 3010-3119.

HALL, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, 95-98.

HAMSHER, S.E., EVANS, K.M., MANN, D.G., POULÍČKOVÁ, A. AND SAUNDERS, G.W., 2011. Barcoding diatoms: exploring alternatives to COI-5P. *Protist*, 162 (3), 405-422.

HARTLEY, B., 1996. *An Atlas of British Diatoms* (ed. P.A. Sims; illustrated by H.G. Barber and J.R. Carter). Bristol: Biopress.

HEBERT, P.D.N., CYWINSKA, A. AND BALL, S.L., 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270 (1512), 313–321.

HOFMANN, G., WERUM, M. AND LANGE-BERTALOT, H., 2011. *Diatomeen im Süßwasser-Benthos von Mitteleuropa* [Diatoms in the freshwater benthos of Central Europe]. Rugell, Liechtenstein: ARG Gantner Verlag KG.

JONES, H.M., SIMPSON, G.E., STICKLE, A.J. AND MANN, D.G., 2005. Life history and systematics of *Petroneis* (Bacillariophyta) with special reference to British waters. *European Journal of Phycology*, 40 (1), 61-87.

JOSHI, N.A. AND FASS, J.N., 2011. *Sickle – a sliding-window, adaptive, quality-based trimming tool for FastQ files, Version 1.33* [software]. Available from: https://github.com/najoshi/sickle [Accessed 26 July 2017].

JUGGINS, S., 2015. *rioja: analysis of quaternary science data. R package version 0.9-6*. Avaialbe from: http://cran.r-project.org/package=rioja [Accessed 26 July 2017].

KAHLERT, M., ALBERT, R.-L., ANTTILA, E.-L., BENGTSSON, R., BIGLER, C., ESKOLA, T., GÄLMAN, V., GOTTSCHALK, S., HERLITZ, E., JARLMAN, A., KASPEROVICIENE, J.,. KOKOCIŃSKI, M., LUUP, H., MIETTINEN, J., PAUNKSNYTE, I., PIIRSOO, K., QUINTANA, I., RAUNIO, J., SANDELL, B., SIMOLA, H., SUNDBERG, I., VILBASTE, S. AND WECKSTRÖM, J., 2009. Harmonization is more important than experience – results of the first Nordic-Baltic diatom intercalibration exercise 2007 (stream monitoring). *Journal of Applied Phycology*, 21 (4), 471-482.

KAHLERT, M., KELLY, M.G., ALBERT, R.-L., ALMEIDA, S., BEŠTA, T., BLANCO, S., DENYS, L., ECTOR, L., FRÁNKOVÁ, M., HLÚBIKOVÁ, D., IVANOV, P., KENNEDY, B., MARVAN, P., MERTENS, A., MIETTINEN, J., PICIŃSKA-FAŁTYNOWICZ, J., ROSEBERY, J., TORNÉS, E., VAN DAM, H., VILBASTE, S. AND VOGEL, A., 2012. Identification verus counting protocols as sources of uncertainty in diatom-based ecological status assessments. *Hydrobiologia*, 695 (1), 109-124.

KAHLERT, M., ÁCS, E., ALMEIDA, S.F.P., BLANCO, S., DREßLER, M., ECTOR, L., KARJALAINEN, S.M., LIESS, A., MERTENS, A., VAN DER WAL, J., VILBASTE, S. AND WERNER, P., 2016. Quality assurance of diatom counts in Europe: towards harmonized datasets. *Hydrobiologia*, 772 (1), 1-14.

KATOH, K. AND STANLEY, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30 (4), 772-780.

KELLY, M.G., 2013. Data rich, information poor? Phytobenthos assessment and the Water Framework Directive. *European Journal of Phycology*, 48 (4), 437-450.

KELLY, M.G. AND WHITTON, B.A., 1995. The Trophic Diatom Index: a new index for monitoring eutrophication in rivers. *Journal of Applied Phycology*, 7 (4), 433-444.

KELLY, M.G., CAZAUBON, A., CORING, E., DELL'UOMO, A., ECTOR, L., GOLDSMITH, B., GUASCH, H., HÜRLIMANN,J., JARLMAN, A., KAWECKA, B., KWANDRANS, J., LAUGASTE, R., LINDSTRØM, E.-A., LEITAO, M., MARVAN, P., PADISÁK, J., PIPP, E., PRYGIEL, J., ROTT, E., SABATER, S., VAN DAM, H. AND

VIZINET, J., 1998. Recommendations for the routine sampling of diatoms for water quality assessments in Europe. *Journal of Applied Phycology*, 10 (2), 215-224.

KELLY, M., JUGGINS, S., GUTHRIE, R., PRITCHARD, S., JAMIESON, J., RIPPEY, B., HIRST, H. AND YALLOP, M., 2008. Assessment of ecological status in U.K. rivers using diatoms. *Freshwater Biology*, 53 (2), 403-422.

KELLY, M., BENNION, H., BURGESS, A., ELLIS, J., JUGGINS, S., GUTHRIE, R., JAMIESON, J., ADRIAENSSENS, V. AND YALLOP, M., 2009. Uncertainty in ecological status assessments of lakes and rivers using diatoms. *Hydrobiologia*, 633 (1), 5-15.

KELLY, M.G., TROBAJO, R., ROVIRA, L. AND MANN, D.G., 2015. Characterizing the niches of two very similar Nitzschia species and implications for ecological assessment. *Diatom Research*, 30 (1), 27-33.

KERMARREC, L., BOUCHEZ, A., RIMET, F. AND HUMBERT, J.-F., 2013. First evidence of the existence of semi-cryptic species and of a phylogeographic structure in the *Gomphonema parvulum* (Kützing) Kützing complex (Bacillariophyta). *Protist*, 164 (5), 686-705.

KERMARREC, L., FRANC, A., RIMET, F., CHAUMEIL, P., FRIGERIO, J.-M., HUMBERT, J.-F. AND BOUCHEZ, A., 2014. A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, 33 (1), 349-363.

KRAMMER, K. AND LANGE-BERTALOT, H., 1986. *Die Süßwasserflora von Mitteleuropa. 2 Bacillariophyceae. Teil 1: Naviculaceae* [The Freshwater Flora of Central Europe. 2 Bacillariophyceae. Part 1: Naviculaceae]. Stuttgart: Gustav Fischer Verlag.

KRAMMER, K. AND LANGE-BERTALOT, H., 1997. *Die Süßwasserflora von Mitteleuropa. 2 Bacillariophyceae. Teil 2: Bacillariaceae, Epithemiaceae, Surirellaceae* [The Freshwater Flora of Central Europe. 2 Bacillariophyceae. Part 2: Bacillariaceae, Epithemiaceae, Surirellaceae], 2nd edition, with a new appendix. Stuttgart: Gustav Fischer Verlag.

KRAMMER, K. AND LANGE-BERTALOT, H., 2000. *Die Süßwasserflora von Mitteleuropa. 2 Bacillariophyceae. Teil 3: Centrales, Fragilariaceae, Eunotiaceae* [The Freshwater Flora of Central Europe. 2 Bacillariophyceae. Part 3: Centrales, Fragilariaceae, Eunotiaceae], 2nd edition. Stuttgart: Gustav Fischer Verlag.

KRAMMER, K. AND LANGE-BERTALOT, H., 2004. *Die Süßwasserflora von Mitteleuropa. 2 Bacillariophyceae. Teil 4: Achnanthaceae. Kritische Ergänzungen zu Achnanthes s.l., Navicula s. str., Gomphonema* [The Freshwater Flora of Central Europe. 2 Bacillariophyceae. Part 4: *Achnanthaceae.* Critical Additions to Achnanthes s.l., Navicula s. Str., Gomphonema], Heidelberg: Spektrum Akademischer/Gustav Fischer.

KRESS, W.J. AND ERICKSON, D.L., 2008. DNA barcodes: genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (8), 2761-2762.

LALLIAS, D., HIDDINK, D.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. AND CREER, S., 2015.Environmental metabarcoding reveals heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine ecosystems. *The ISME Journal*, 9, 1208-1221.

LIANG, Z. AND KEELEY, A., 2013. Filtration recovery of extracellular DNA from environmental water samples. *Environmental Science and Technology*, 47 (16), 9324-9331.

LIN, L.I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45 (1), 255-268.

LOBBAN, C.S. AND ASHWORTH, M.P., 2014. *Hanicella moenia*, gen. et sp. nov., a ribbon-forming diatom (Bacillariophyta) with complex girdle bands, compared to *Microtabella interrupta* and *Rhabdonema* cf. *adriaticum*: implications for Striatellales, Rhabdonematales, and Grammatophoraceae, fam. nov. *Journal of Phycology,* 50 (5), 860-884.

MANN, D.G., 1999. The species concept in diatoms. *Phycologia*, 38 (6), 437-495.

MANN, D.G., 2010. Discovering diatom species: is a long history of disagreements about species-level taxonomy now at an end? *Plant Ecology and Evolution*, 143 (3), 251-264.

MANN, D.G., THOMAS, S.J. AND EVANS, K.M., 2008. Revision of the diatom genus Sellaphora: a first account of the larger species in the British Isles. *Fottea*, 8 (1): 15-78.

MANN, D.G., SATO, S., TROBAJO, R., VANORMELINGEN, P. AND SOUFFREAU, C., 2010. DNA barcoding for species identification and discovery in diatoms. *Cryptogamie, Algologie*, 31 (4): 557-577.

MARTIN, M., 2001. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17 (1), 10-12.

MCCUNE, B. AND GRACE, J.B., 2002. *Analysis of Ecological Communities*. Gleneden Beach, OR: MjM Software Design.

MONIZ, M.B. AND KACZMARSKA, I., 2009. Barcoding diatoms: is there a good marker? *Molecular Ecology Resources*, 9 (Suppl. 1), 65-74.

MONIZ, M.B. AND KACZMARSKA, I., 2010. Barcoding of diatoms: nuclear encoded ITS revisited. *Protist*, 161 (1), 7-34.

NILSSON, R.H., TEDERSOO, L., RYBERG, M., KRISTIANSSON, E., HARTMANN, M., UNTERSEHER, M., PORTER, T.M., BENGTSSON-PALME, J., WALKER, D.M., DESOUSA, F., GAMPER, H.A., LARSSON, E., LARSSON, K-H., KOLJALG, U., EDGAR, R.C. AND ABARENKOV, K., 2015. A comprehensive, automatically updated fungal ITS sequence dataset for reference-based chimera control in environmental sequence efforts. *Microbes and Environments*, 30 (2), 145-150.

OKSANEN, J., KINDT, R., LEGENDRE, P. AND O'HARA, R.B., 2007. *vegan: Community Ecology Package, version 1.8-5*, released 11 January 2007. Available from: https://cran.r-project.org/src/contrib/Archive/vegan/ [Accessed 28 July 2017].

PARDO, I., GÓMEZ-RODRÍGUEZ, C., WASSON, J.-G., OWEN, R., VAN DE BUND, W., KELLY, M., BENNETT, C., BIRK, S., BUFFAGNI, A., ERBA, S., MENGIN, N., MURRAY-BLIGH, J., OFENBÖECK, G., 2012. The European reference condition concept: a scientific and technical approach to identify minimally-impacted river ecosystems. *Science of the Total Environment*, 420, 33-42.

PERES-NETO, P. AND JACKSON, D., 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, 129 (2), 169-178.

POIKANE, S., KELLY, M.G. AND CANTONATI, M., 2016. Benthic algal assessment of ecological status in European lakes and rivers: challenges and opportunities. *Science of the Total Environment*, 568, 602-613.

POTAPOVA, M., 2012. New species and combinations in monoraphid diatoms (family Achnanthidiacaeae) from North America. *Diatom Research,* 27 (1), 29-42.

PRYGIEL, J., CARPENTIER, P., ALMEIDA, S., COSTE, M., DRUART, J.-C., ECTOR, L., GUILLARD, D., HONORÉ, M.-A., ISERENTANT, R., LEDEGANCK, P., LALANNE-CASSOU, C., LESNIAK, C., MERCIER, I., MONCAUT, P., NAZART, M., NOUCHET, N., PERES, F., PEETERS, V., RIMET, F., RUMEAU, A., SABATER, S., STRAUB, F., TORRISI, M., TUDESQUE, L., VAN DER VIJVER, B., VIDAL, H., VIZINET, J. AND ZYDEK, N., 2002. Determination of the biological diatom index (IBD NF T 90-354): results of an intercomparison exercise. *Journal of Applied Phycology*,14 (1), 27-39.

RAUWOLF, U., GOLCZK, H., GREINER, S. AND HERMANN, R.G., 2010. Variable amounts of DNA related to the size of chloroplasts III: biochemical determinations of DNA amounts per organelle. *Molecular Genetics and Genomics*, 283 (1), 35-47.

R DEVELOPMENT CORE TEAM, 2017. *R: A Language and Environment For Statistical Computing*. *Reference Index.* Version 3.4.1 (2017-06-30). Vienna: R Foundation for Statistical Computing. Available from: https://cran.r-project.org/manuals.html [Accessed 28 July 2017].

ROVIRA, L., TROBAJO, R., SATO, S., IBÁÑEZ, C. AND MANN, D.G., 2015. Genetic and physiological diversity in the diatom *Nitzschia inconspicua*. *Journal of Eukaryotic Microbiology*, 62 (6), 815-832.

ROUND, F.E., CRAWFORD, R.M. AND MANN, D.G., 1990. *The Diatoms: Biology and Morphology of the Genera*. Cambridge: Cambridge University Press.

SAGARIN, R., CARLSSON, J., DUVAL, M., FRESHWATER, W., GODFREY, M.H., LITAKER, W., MUNÕZ, R., NOBLE, R., SCHULTZ, T. AND WYNNE, B, 2009. Bringing molecular tools into enviornmental resource management: untangling the molecules to policy pathway. *PLOS Biology*, 7 (3), 426-430.

SCHNEIDER, S.C. AND LINDSTRØM, E.-A., 2011. The periphyton index of trophic status PIT: a new eutrophication metric based on non-diatomaceous benthic algae in Nordic rivers. *Hydrobiologia*, 665 (1), 143-155.

SCHNEIDER, S.C., KAHLERT, M. AND KELLY, M.G., 2013. Interactions between pH and nutrients on benthic algae in streams and consequences for ecological status assessment and species richness patterns. *Science of the Total Environment* 444, 73-84.

SCHULTZ, M.E., 1971. Salinity-related polymorphism in the brackish-water diatom *Cyclotella cryptica*. *Canadian Journal of Botany*, 49 (8), 1285-1289.

SNELL, M.A., BARKER, P.A., SURRIDGE, B.W.J., LARGE, A.R.G., JONCZK, J., BENSKIN, C.M., REANEY, S., PERKS, M.T., OWEN, G.J., CLEASBY, C., DEASY, C., BURKE, S. AND HAYGARTH, P.M., 2014. High frequency variability of environmental drivers determining benthic community dynamics in headwater streams. *Environmental Science: Processes & Impacts,* 16 (7), 1629-1636.

STEVENSON, M., 2010. *epiR: Functions for analysing epidemiological data*, Version 0.9-27, released 20 September 2010. Aavailable from: https://cran.r-project.org/src/contrib/Archive/epiR/ [Accessed 28 July 2017].

TER BRAAK, C.J.F. AND BARENDREGT, L.G., 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. *Mathematical Biosciences*, 78 (1), 57-72.

TER BRAAK, C.J.F. AND LOOMAN, C.W.N., 1986. Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio*, 65 (1), 3-11.

TROBAJO, R., CLAVERO, E., CHEPURNOV, V.A., SABBE, K., MANN, D.G., ISHIHARA, S. AND COX, E.J., 2009. Morphological, genetic and mating diversity within the widespread bioindicator *Nitzschia palea* (Bacillariophyceae). *Phycologia*, 48 (6), 443-459.

UKTAG, 2013. *Updated recommendations on phosphorus standards for rivers. River Basin Management (2015-2021).* Final report. Water Framework Directive UK Technical Advisory Group.

UNDERWOOD, A.J., 1991. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. *Australian Journal of Marine & Freshwater Research*, 42, 569-587.

UNTERGASSER, A., CUTCUTACHE, I., KORESSAAR, T., YE, J., FAIRCLOTH, B.C,, REMM, M. AND ROZEN, S.G., 2013. Primer3 – new capabilities and interfaces. *Nucleic Acids Research*, 40 (15), e115.

VISCO, J.A., APOTHÉLOZ-PERRET-GENTIL, L., CORDONIER, A., ESLING, P., PILLETT, L. AND PAWLOWSKI, J., 2015. Environmental monitoring: inferring diatom index from next-generation sequencing data. *Environmental Science and Technology*, 49 (13), 7597-7605.

VON STOSCH, H.A. AND FECHER, K., 1979. 'Internal thecae' of *Eunotia soleirolii* (Bacillariophyceae): development, structure and function as resting spores. *Journal of Phycology*, 15 (3), 233-243.

WHITTON, B.A., JOHN, D.M., JOHNSON, L.R., BOULTON, P.N.G., KELLY, M.G. AND HAWORTH, E.Y., 1998. *A Coded List of Freshwater Algae of the British Isles*, LOIS Publication Number 222. Wallingford: Institute of Hydrology.

WOODWARD, G., GRAY, C. AND BAIRD, D.J., 2013. Biomonitoring for the 21st century: new perspectives in an age of globalisation and emerging environmental threats. *Limnetica*, 32 (2), 159-174.

ZHANG, J., KOBERT, K., FLOURI, T. AND STAMATAKIS,A., 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30 (5), 614-620.

ZHANG, Z., SCHWARTZ, S., WAGNER, L. AND MILLER, W., 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, 7 (1-2), 203-214.

ZGRUNDO, A., LEMKE, P., PNIEWSKI, F., COX, E.J. AND LATALA, A., 2013. Morphological and molecular phylogenetic studies on *Fistulifera saprophila*. *Diatom Research*, 28 (4), 431-443.

ZIMMERMANN, J., JAHN, R. AND GEMEINHOLZER, B., 2011. Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocol. *Organisms Diversity & Evolution*, 11, 173-192.

ZIMMERMAN, J., ABARCA, N., ENK, N., SKIBBE, O., KUSBER, W.-H. AND JAHN, R, 2014. Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. *PLOS One*, 9 (9), e108793.

List of abbreviations

BLAST	Basic Local Assignment Search Tool
bp	base pair
COI	cytochrome c oxidase subunit 1
DNA	deoxyribonucleic acid
DTAB	dodecyltrimethylammonium bromide
eDNA	environmental DNA
EDTA	ethylenediaminetetraacetic acid
eTDI	expected Trophic Diatom Index
EQR	Ecological Quality Ratio
IMS	industrial methylated spirits
ITS	internal transcribed spacer
LM	light microscopy
MID	multiple identifier
NCBI	National Center for Biotechnology Information [USA]
NGR	National Grid Reference
NGS	next generation sequencing
NMDS	non-metric multidimensional scaling
OTU	Operational Taxonomic Unit
PCR	polymerase chain reaction
PROMpT	Primary Rapid Overview of Metagenomic Taxonomy
QIIME	Quantitative Insights Into Microbial Ecology
RA	relative abundance
rbcL	ribulose bisphosphate carboxylase large chain gene
SEM	scanning electron microscopy
SSU	small ribosomal subunit
STW	sewage treatment works
TDI	Trophic Diatom Index
UV	ultraviolet
WFD	Water Framework Directive

Glossary

Bioinformatics	Field of biology that uses computer science, statistics, mathematics and engineering to study and process biological data.
Bioinformatics pipeline	Steps involved in extracting, processing and analysing raw data generated, for example, by next generation sequencing.
BLAST®	Basic Local Assignment Search Tool – bioinformatics tool that finds regions of local similarity between DNA or protein sequences (http://blast.ncbi.nlm.nih.gov/Blast.cgi).
DNA barcoding	Identification of a species or taxon based on PCR amplification and sequencing of a standard region of DNA (often the mitochondrial cytochrome oxidase 1 gene).
GenBank®	Annotated collection of publicly available DNA sequences housed at the National Center for Biotechnology Information (USA) (www.ncbi.nlm.nih.gov/genbank/).
Illumina sequencing	Next generation sequencing on a platform developed by the company Illumina, such as MiSeq™ (www.illumina.com/systems/miseq.html) used in the current study.
Metabarcoding	A rapid method of biodiversity assessment that combines 2 technologies: • DNA based taxon identification (DNA barcoding) • high-throughput DNA sequencing (NGS) It uses universal PCR primers to mass-amplify DNA barcodes from mass collections of organisms or from environmental DNA.
Next generation DNA sequencing (NGS)	Also known as high-throughput sequencing, 'next generation sequencing' is the catchall term used to describe a number of different modern sequencing technologies, including Illumina (Solexa). These recent technologies allow the sequencing of DNA that is much quicker and cheaper than the previously used Sanger sequencing, and as such have revolutionised the study of genomics and molecular biology.
Operational taxonomic unit (OTU)	Clusters of similar rbcL barcode variants. It is a means of categorising taxa based on their sequence similarity. Each cluster represents a taxonomic unit for example, species or genus.
Polymerase chain reaction (PCR)	A method of amplifying the number of copies of a target region of DNA using oligonucleotide primers which permits downstream analysis such as DNA sequencing.
Primer	A short single-stranded stretch of DNA that is complementary to the DNA sequence of a target region. A pair of primers, flanking the target region, is required for PCR amplification. The primers bind to the target DNA during PCR and prime

	the addition of nucleotides, generating millions of copies of the target sequence.
PROMpT	A bioinformatics pipeline system for rapid metagenomic analysis of NGS amplicon sequencing data with a simple web interface, allowing non-informatic users access to the benefits from NGS sequencing (https://github.com/passdan/prompt). While built for NGS, there is also the capacity to load light microscopy data into the pipeline to allow easy comparisons between methods. It is designed to be implemented in analysis of your chosen taxonomic clade, requiring only reference sequences formatted into a BLAST (Basic Local Alignment Search Tool - http://blast.ncbi.nlm.nih.gov/Blast.cgi) database and a taxonomic hierarchy that can both be defined by the user. It also allows for correction factors to be applied to the reference sequences to allow for polyploidy or the effect of size differences in the community.

Appendix 1: Proof of concept – testing the feasibility of developing diatom ecological assessment metrics from NGS data

A1.1 Introduction and method development

The overall objective of this work is to develop a cost-effective operational molecular diatom tool to determine water quality for the Water Framework Directive using diatom DNA barcodes combined with next generation sequencing (NGS). In addition, it is hoped it will liberate molecular techniques from the research environment, and demonstrate their power and utility within a regulatory framework. This will hopefully facilitate their uptake into other areas of the Environment Agency's monitoring programme such as macroinvertebrate and fish monitoring.

The proof of concept phase was carried out from September 2011 to March 2014 to develop and test an alternative means of evaluating ecological status using benthic diatoms and NGS rather than light microscopy (LM) as the basis for sample analysis. A brief overview of the work is provided here. The chloroplast-based rbcL gene was selected on the basis of prior studies (see Section A.1.1.3), as the most suitable barcode for routine environmental assessment using diatoms.

A1.1.1 Sample handling and transfer

For this proof of concept phase of the project, diatom samples were collected from rivers in England by Environment Agency Area staff as part of routine surveys in autumn 2011 by brushing the top surface of 5 cobbles with a clean toothbrush to remove the biofilm (following standard Environment Agency protocols). Samples were returned to the laboratory where 15ml of biofilm/water suspension was removed and centrifuged to generate a pellet of diatoms and frozen at -20°C. The remaining sample was preserved in Lugol's iodine for morphological analysis. The preserved sample and frozen pellet were then transferred at 4°C – using the Environment Agency's infrastructure – to a laboratory in Exeter where they were again stored at -20°C. When sufficient numbers had been collected, the samples were dispatched under dry ice to Cardiff University. The Lugol's preserved samples were transferred to Bowburn Consultancy (Durham) for morphological analysis and the associated pelleted diatom samples were stored at -70°C prior to DNA extraction. Approximately 100 samples were collected; a subset was used to test whether the NGS approach could provide meaningful diatom species metrics compared with LM.

A1.1.2 DNA extraction

Initial DNA extractions were performed using a commercial procedure, Qiagen DNeasy® Plant Mini Kit (69104). But although DNA was extracted and barcodes amplified, it was evident both from spectrophotometric analysis of the DNA and the

dilution required to perform some amplifications that the template DNA was of varying quality.

To ensure the templates generated were of consistent high quality, the extraction procedure was re-optimised. Three extraction procedures were compared: the DNeasy Plant Mini Kit (Qiagen Ltd); the Instagene DNA matrix (Bio-Rad); and a procedure developed using a hybrid involving glass bead lysis into a DTAB extraction (Fawley and Fawley 2004). The hybrid protocol yielded a simple and rapid technique for extraction of DNA from diatoms followed by a DNeasy column purification. The latter technique yielded DNA of consistent quality and qualities sufficient for the purposes of this study. UV/visible region spectra of typical extractions are shown in Figure A1.1.

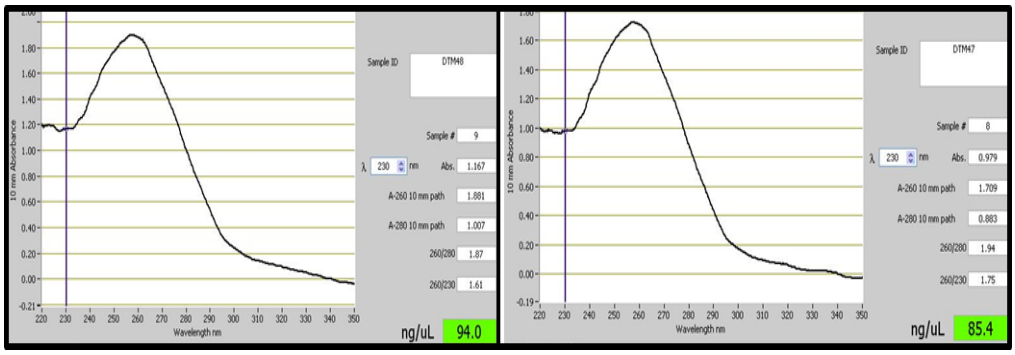


Figure A1.1 Representative spectrophometric analysis of DNA extracted from diatom samples

DNA extractions were also trialled on diatom samples that had been preserved in Lugol's iodine, the standard preservative for samples for LM. However, these yielded exceptionally low quantities of DNA, which were unsuitable for further analysis.

A1.1.3 Amplification of rbcL-3′ barcode from environmental samples

To test whether it was possible to amplify rbcL barcodes from diatom samples collected by Environment Agency Area staff, the 3′ prime (3P or 3′) end of the Rubisco rbcL gene was targeted using forward primer Cfd_F (CCRTTYATGCGTTGGAGAGA) and reverse primer DPrbcL7 (AARCAACCTTGTGTAAGTCT) (Hamsher et al. 2011) to amplify ~850 bp amplicons.

Amplifications were performed on genomic DNA in 25µl reaction volumes using the following conditions: one cycle at 94°C for 3 minutes, followed by 30 cycles at 94°C (30 seconds), 55°C (30 seconds) and 72°C (1 minute). A final step at 72°C (10 minutes) was included. PCR products were electrophoresed through a 1.5% agarose gel at 4–5V per cm, and visualised using SYBR Safe stain (Invitrogen) under UV light, with a 100 bp Plus Gene Ruler ladder (Fermentas). Although differences in the intensity of the band were observed between samples, ~90% of the amplification succeeded with no further optimisations.

A1.1.4 Validation of diatom rbcL-3′ barcode generation

The success of amplifying rbcL from an environmental sample was assessed by cloning and sequencing representatives from diatom samples. For this 1µl of the gel purified rbcL-3′ amplicon was cloned using the Topo cloning kit (Invitrogen) according to the manufacturer's instructions. Briefly, 5µl cloning reactions were set up containing 1µl PCR product, 1µl salt solution, 1µl water and 1µl of the Topo cloning vector. The

reaction was incubated at room temperature for 10 minutes and then 2µl was used to transform the One Shot® Mach1™-T1R Competent Cells provided in the kit using the procedure defined for the chemically competent cells. Ten microlitres and 40µl of the transformation were spread onto LBkan plates and colonies grown at 30°C for 24 hours and 37°C for 18 hours respectively. Colonies from the LBkan plates were transferred to 5ml of LBkan and grown overnight at 37°C. Plasmid preparations were then performed on 4.5ml of the LBkan using a Promega SV mini prep kit following the manufacturer's instructions with 2 minor modifications:

- DNA was dried by centrifuging for 2 minutes in a fresh tube prior to elution
- elution was performed in 75µl of sterile distilled water

DNA was quantified using a Nanodrop and digested with *Eco*R1, which cuts either side of the Topo vector. Successful recombinants were selected and sequenced using Sanger sequencing (MSBU Cardiff University). The diatom DNA barcode products were confirmed as being rbcL-3′ using the BLASTX algorithm against the Non-redundant GenBank Database (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The identified closest matching species were ascertained and recorded (data not shown).

A1.1.5 Development and testing of rbcL diatom barcode

To determine which portion of the Rubisco rbcL gene to target, 50 diatom rbcL gene sequences were retrieved from the European Bioinformatics Institute (EMBL) database using the Sellaphora sequences detailed in Hamsher et al. (2011) as a retrieval tag. Five of these sequences were discarded due to poor quality (runs of *n* bases in the sequence) or short lengths. An alignment of the remaining 45 sequences was constructed using ClustalW, and the degree of consensus bases (>55% consensus from the 45 sequences) was determined at the 5′and 3′ ends of the gene. From this output, it was estimated that the 5′ end of the gene showed 65% consensus over 734 bases, with 58% over 725 bases at the 3′ end. It was therefore judged that either end of the gene would be equally useful in the context of this work.

NGS technology developments

At the start of the proof of concept work, the range of second generation NGS technologies universally relied on a clonal amplification step prior to sequencing, either emPCR (Roche GS FLX or SOLID technologies) or surface-based amplification (Illumina and Ion torrent). This amplification step provided a limit to the size of the amplicon that could then be sequenced.

For GS FLX and GS FLX+, various 'long emPCR' protocols have been proposed, although those routinely using the platforms advise amplicons of <600 bp and the protocols available prior to 2012 recommended amplicons ideally between 300 and 400 bp. This therefore provided this project with a significant hurdle since the current rbcL barcode that has been validated previously and used for phylogenetics is ~850 bp, yielding a ~950 bp amplicon when the required NGS sequencing primer, multiple identifier (MID) tags and calibration sequences have been added to each end.

A significant technical development was announced in November 2012 by Roche, the manufacturers of the GS FLX platform. This involved enhancements in software and reagent developments, which allowed the GS FLX+ platform to sequence amplicons of up to 1,200 bp, providing a sequence range distribution with a modal distribution centred on 950 bp. To take advantage of these developments, platforms must have both software upgrades and exploit 'new' suites of reagents and protocols. The

immediate advantage to this project was that it would allow the use of the established rbcL-3′ barcode and exploit all of the information content of this fragment.

Since expert confidence in the provision of sequence data from the longer amplicon was low[2] and additional developments might present more cost-effective long-term solutions, it was decided that the project should attempt to assimilate the opportunities presented by the rapidly changing NGS landscape.

A summary of the current competing technologies is provided in Table A1.1, which illustrates that exploitation of a smaller amplicon would allow utilisation of technologies that would significantly reduce the costs associated with NGS analysis to a tenth or a fifth of those for the GS FLX+ platform. It was therefore decided to evaluate shorter rbcL barcode amplicons, with a particular focus on the informatics content of the outputs.

Informatic design of custom NGS compatible primer sets

To derive whether shorter amplicons compatible with alternative NGS could be developed, it was necessary to establish their validity. Previous research had identified the longer rbcL-3′ fragment as a potential barcode for environmental analyses as it fulfilled the following criteria.

- It provides appropriate taxonomic resolution.
- Validated and optimised primers exist for its cross-species amplification.
- The primers have been tested for environmental diatom analysis.

It was therefore essential to establish that alternative primers which would amplify shorter fragments could fulfil these criteria. To develop additional NGS compatible primer sets, >1,100 rbcL sequences were downloaded from GenBank® (www.ncbi.nlm.nih.gov/genbank/). These were filtered to select those that contained the majority of the rbcL-3′ region (used for phylogeny reconstructions) and to remove species/taxa redundancy. This yielded 349 sequences. These were aligned using the software tool Muscle and the variation across the sequence determined.

A number of ‘regions’ displayed conservation and these were examined by eye to identify possible priming sites. The parameters used included:

- ≤8 fold redundancy would yield a match to all species represented
- terminal 3′ bases were invariant
- T_m matched those primers already used to generate the 850 bp amplicon
- did not represent repetitive sequence
- primers displayed no significant hairpins, self-priming or primer-primer interaction

[2] Personal communication from Edinburgh Genetics, Centre for Genomic Research Liverpool, and the Food Standards Agency’s genomic unit at York

Table A1.1 Comparison of NGS platforms

Platform	Company	Read length	Accuracy	Number of reads (millions)	Multiplex [5]	Depth per sample (K)	Cost per run [1]	Cost per sample [1]	Instrument run time (hours)
GS FLX	Roche	750 bp	98.9%	0.5	25	20,000	£5,000	£200	24
Ion Torrent PGM	Life Technologies	400 bp	98.3%	5	200	25,000	£500	£2.50	12
MiSeq	Illumina	2 × 300 bp	99.2%	25	200	125,000	£1,000	£5	56
HiSeq2500	Illumina	2 × 250 [4] bp	99.7%	250	384 [3]	650,000	£3,500	£9	60
MinION	Oxford Nanopore	>5kb bp	95[2]%	0.1 [2]	5	20,000	£350	£70	48

Notes:
[1] The costs per run and cost per sample are for the direct sequencing costs only and do not include sample preparation costs, which are comparable between platforms.
[2] MinION accuracy and read numbers are estimates based on data released by Oxford Nanopore for R9 flow cells.
[3] Only 384 barcodes are currently commercially available for HiSeq but more could be custom synthesised.
[4] Standard mode HiSeq gives 2 × 150 bp HiSeq2500 run in fast mode can produce 2 × 250 bp. This is likely to be too short for the rbcL mini-barcode.
[5] Multiplexing to achieve the depth of coverage was used during the project, not necessarily the maximum that could be achieved.

Two regions were identified (Figure A1.2), fortuitously dividing the rbcL gene into 3 equal sections of approximately 300 bp. Degenerate primers were designed (Figure A1.2 and Table A1.2) from these regions and used for further analysis.

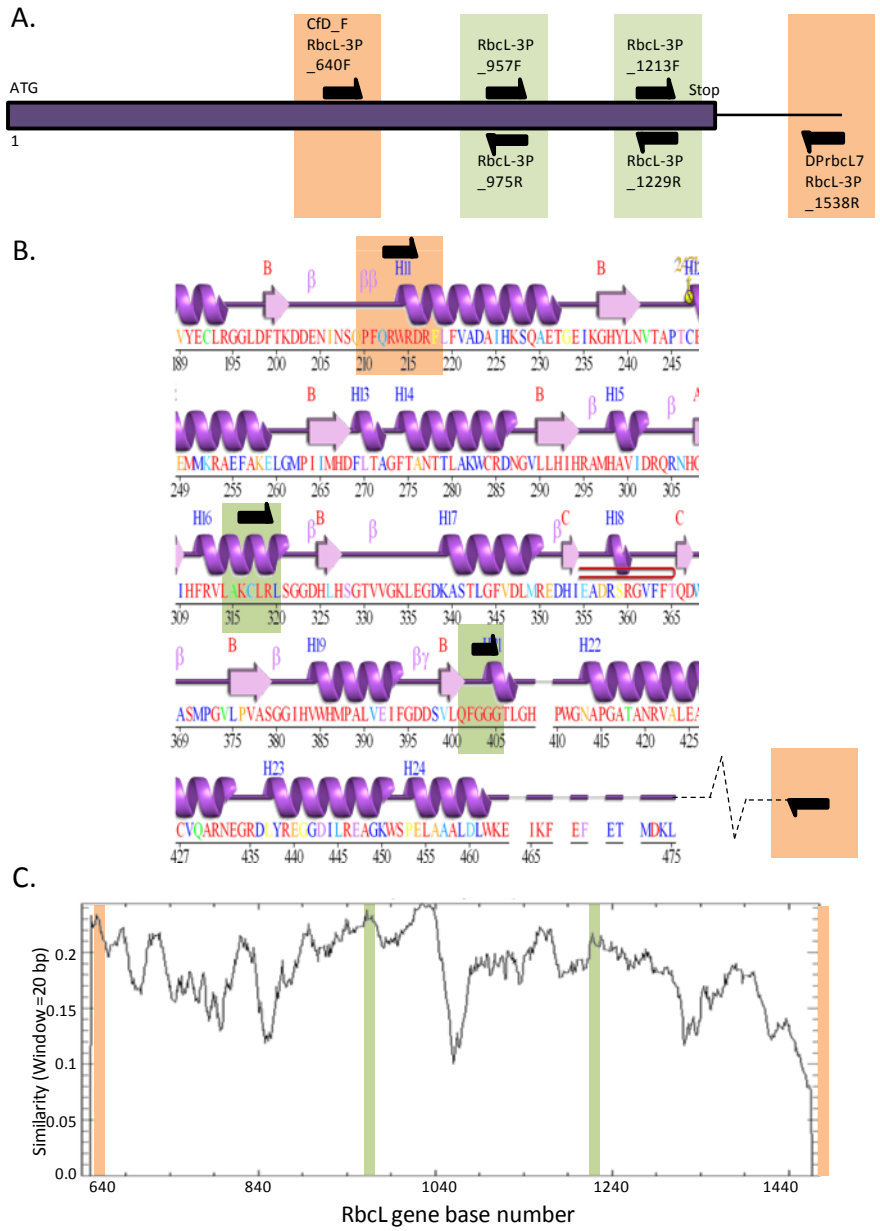


Figure A1.2 Design of rbcL NGS compatible primers

Notes: Approximate location of new and established primers are overlaid onto illustrations of the gene encoding the large Rubisco subunit (Panel A) as well as the secondary protein fold (Panel B) derived for *Synechococcus elongates.*
Nucleotide and protein numbering is initiated from the first base of the methionine or the methionine itself, respectively.
Primers where the background shading is given in orange represent the established primer set, while new primers are denoted using a green background shading.
The similarity observed when the non-redundant 349 rbcL sequences that cover the 3′ region of the rbcL gene used for barcoding is shown in Panel C. As with Panel A, the nucleotide number numbering is initiated from the gene's start codon. Unfortunately very few of the sequences report the 'full sequence' including the rbcL-3′ primer sequence and therefore the conservation plot does not incorporate this region.

Table A1.2 Degenerate primers designed for NGS rbcL amplicon generation

CfD_F || rbcL-3P_640F (640): CCRTTYATGCGTTGGAGAGA
DPrbcL7|| rbcL-3P_1538R (1538): AARCAACCTTGTGTAAGTCT [5′-AGACTTACACAAGGTTGYTT-3′]
rbcL-3P_957F T_m 54°C: 5′ R TGG ATG CGT ATG KSW GG 3′
rbcL-3P_975R T_m 55°C 5′- ACC WSM CAT ACG CAT CCA -3′ [5′- TGG ATG CGT ATG KSW GGT -3′]
rbcL-3P_1213F: 5′- TTY GGT GGT GGT ACW ATI GG -3′
rbcL-3P_1229R: 5′- ATW GTA CCA CCA CCC AAC TGT A -3′ [5′- TAC AIT TIG GTG GTG GTA CWA T -3′]

Notes: All primer are given 5′–3′ on the positive strand unless otherwise indicated

To determine whether reducing the size of the fragment amplified for barcode purposes would have an impact on the taxonomic resolution and the ability to exploit those sequences submitted to the data repositories, bespoke software was developed to identify and extract specific regions of the rbcL genes. Simulated amplicons were either bracketed by specific primers or selected as a specific size starting from a primer site. Hamsher et al. (2011) used a fragment of 748bp 3′ of the rbcL-3P_640F (CfD_F) primer to validate the rbcL-3′ primers (CfD_F and DPrbcL7 now assigned the systematic names rbcL-3P_640F and rbcL-3P_1538R respectively). A number of sequences available in GenBank and their respective information content were therefore analysed for a suite of regions of the rbcL-3' region both in its entirety and with simulated primer sub-sequences (Figure A1.3, Table A1.3 and Table A1.4). The metric employed to explore information content was the number of operational taxonomic units (OTUs), defined at a series of thresholds representing the percentage identity of the sequences within an OTU. This metric was selected in preference to classical metrics because, although relaxing the sequence identity match when assigning species will always provide additional assignment, it increases the potential error and removes potentially valuable information. This is especially relevant for NGS sequencing approaches where technical error is higher than with classical Sanger approaches.

Initial analysis was performed using all rbcL sequences submitted to the databases (Table A1.3). This showed that only 6 database sequences representing full plasmid genomes contain the complete rbcL-3′ region. The sequence employed by Hamsher et al. (2011) (primer region/amplicon A1 in Table A1.3) is represented by 383 entries, while the sequence spanned by rbcL-3P_640F (CfD_F) and rbcL-3P_975R (primer region/amplicon D in Table A1.3) is represented within 727 entries.

It is misleading to compare the OTU representation since each group of sequences contains different qualities of species and taxa redundancy. Therefore, the analysis was repeated employing the 349 sequences used for the design of the new NGS primers (see above). This analysis clearly shows that amplicon D, representing the fragment between rbcL-3P_640F (CfD_F) and the rbcL-3P_975R, is contained in 301 entries and represents the largest number of OTUs at 268 at 0.97% identity (Table A1.4). These analyses suggest that this first amplicon may be optimal for NGS based analysis of environmental samples.

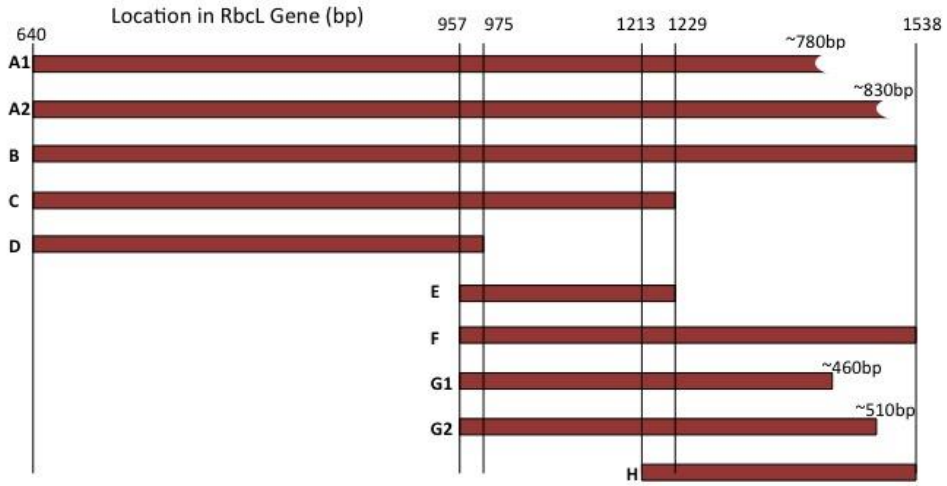


Figure A1.3 Regions of rbcL-3′ gene exploited for bioinformatic analysis

Notes: Numbering used to define location in rbcL gene defined from first base of the start codon.

Table A1.3 OTU analysis of full GenBank representation of rbcL-3′ regions

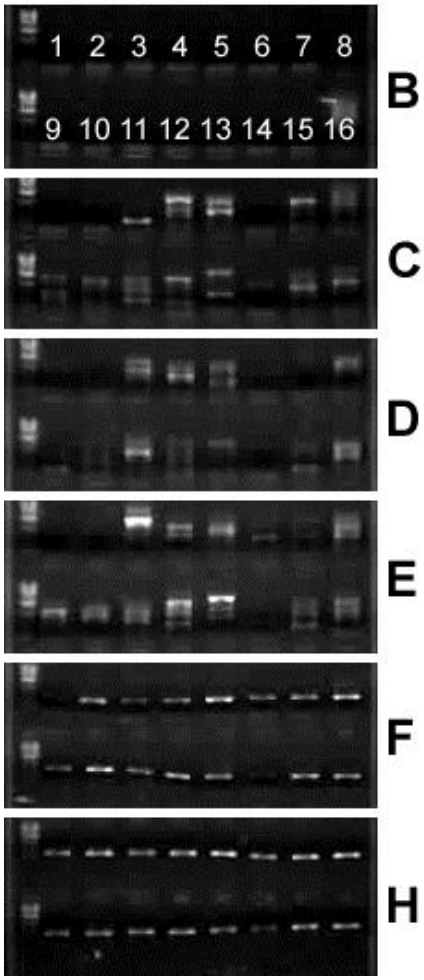
Primer region	No. Applicable reads from all GenBank records	No. OTUs @0.97	No. OTUs @0.95	No. OTUs @0.90
A1	383	202	139	48
A2	62	24	17	6
B	6	5	5	3
C	570	262	190	63
D	727	345	268	92
E	694	271	188	44
F	10	7	6	5
G1	520	214	152	48
G2	95	24	15	9
H	16	10	10	6

Table A1.4 OTU analysis of 349 GenBank entries for selected rbcL-3′ regions

Primer region	No. Applicable reads from reduced fasta file	No. OTUs @0.97	No. OTUs @0.95	No. OTUs @0.90
A1	243	125	174	43
C	301	165	222	57
D	301	268	232	92
E	314	142	194	36
G1	270	128	182	35

Experimental validation of NGS specific primer designs

It was essential to confirm that the proposed alternative primers would amplify a wide spectrum of diatom species if they were to be compatible with diatom assemblage analysis. To establish compatibility and to optimise the specific PCR conditions, amplifications were performed using DNA templates representing 16 diatom species isolated as part of the rbcL reference database development (Section 3). The analysis revealed that amplicons H and F provided single amplicon bands with all 16 species tested.



1	0614	Tabellaria flocculosa
2	0641	Melosira varians
3	0643	Fragilaria gracilis
4	0647	Pseudostaurosira brevistriata
5	0648	Cyclotella meneghiniana
6	0654	Gomphonema acuminatum
7	0659	Diatoma vulgare
8	0660	Nitzschia palea
9	0666	Achnanthidium minutissim var. cf jackii
10	0667	Encyonema silesiacum
11	0680	Navicula cf radiosa
12	0710	Sellaphora joubaudii
13	0715	Mayamaea atomus var. permitis
14	0729	Eunotia cf pectinalis
15	0762	Neidium cf. ampliatum
16	0876	Stephanodiscus hantzschii

Figure A1.4 Cross-species validation of primer sets (left) with representative phylogenetically diverse clones (right)

Notes: Amplicons are labelled as given in Figure A1.3.

Samples 1–16 are selected to cover a wide range of diatoms including radial centrics (blue), polar/thalassiosiroid centrics (green), araphid pennates (pink) and raphid pennates (yellow).

A1.1.6 Amplification of NGS compatible MID tagged barcodes representing diatom assemblage

Amplification trial of alternative NGS primer sets

The initial approach was to perform a head-to-head analysis of the 3 possible amplicons – amplicon B (representing the original validated rbcL barcode; Hamsher et al. 2011), amplicon F and amplicon H – using the GS FLX+ long sequence protocol and evaluating the species representation achieved by each amplicon. Should the shorter amplicons provide similar species representation, subsequent analyses could exploit alternative and more cost-effective technologies. This being the case, a 96-well plate design was adopted to combine the appropriate diatom-specific primers (Table A1.2) with an appropriate amplicon specific MID tag. The design of the plate was engineered to allow for high-throughput analysis of 11 environmental diatom samples with the 3 alternate sets of primers and subsequent analysis of the individual samples. Primers were supplied at 100 µM and diluted to working concentrations.

Unfortunately, subsequent amplification tests for amplicons F and H failed to generate complementary amplicons from each sample. Due to time constraints, this approach was abandoned in favour of just sequencing the original longer amplicon (amplicon B). Further work on developing a short rbcL barcode is described in Section 4.

Amplification of rbcL-3′ for diatom community analysis by NGS

The only European NGS supplier that would guarantee delivery of a long sequence was MWG-Biotech. Therefore, primers were designed that combined sample-specific MID tags together with the rbcL-3′ primers CfD_F || rbcL-3P_640F (640): CCRTTYATGCGTTGGAGAGA and DPrbcL7|| rbcL-3P_1538R (1538).

The MWG-Biotech protocol recommended the production of MID tagged amplicons which would be ligated to sequence adapters then size selected, purified, qualified and combined into pools prior to NGS analysis. This protocol would yield extended sequences from both the forward and reverse direction of the amplicons since the ligation of the sequencing adapter was non-specific. However, this approach significantly reduced the costs of the amplification primers and reduced the possibility of primer-based artefacts.

To remove PCR bias that may arise from initial primer hybridisation, it is standard practice to perform triplicate amplifications prior to pooling each sample for NGS analysis. Furthermore, to reduce any proofreading errors generated by *Taq* polymerase, the NGS amplification exploited a proofreading, hot start polymerase. A total of 17 successful amplifications from environmental diatom samples were provided to MWG-Biotech for sequencing. It should be noted that, within this batch, a small number were at the limits of the permissible concentrations, a fact that was identified during MWG-Biotech's quality control analysis.

A1.2 Analysis of NGS data

A1.2.1 Development of PROMpT: bioinformatic pipeline software

NGS data were processed using the PROMpT pipeline software (https://passdan.github.io/prompt), developed in parallel to this study. It was augmented with customisation for rbcL diatom analysis utilising the diatom rbcL reference sequences generated (Section 3). This allowed for integration of both the

NGS amplicon data and the classical LM analysis, allowing direct comparison as visualised in Figure A1.5.

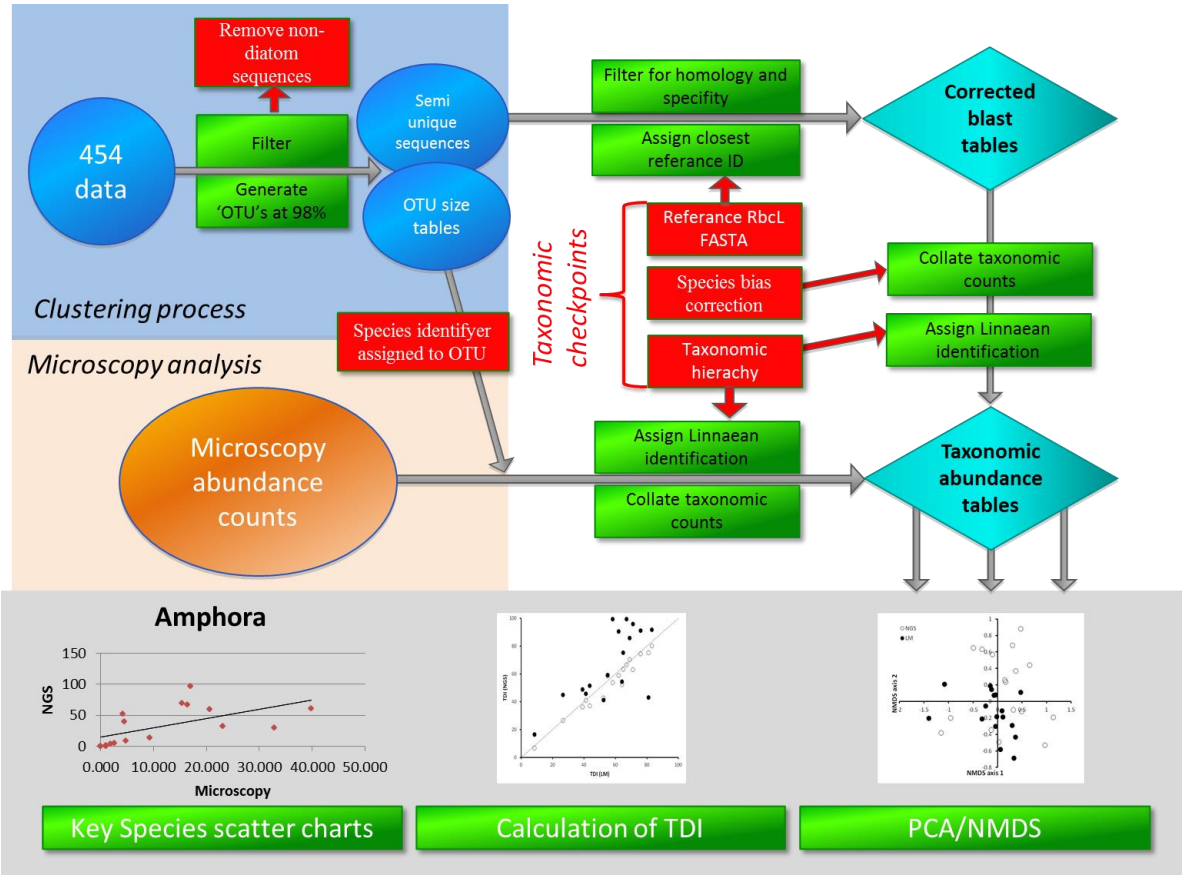


Figure A1.5 Overview of analytical workflow of PROMpT

Notes: PCA = principal component analysis

A1.2.2 Analysis of NGS data

From the 17 samples submitted for NGS analyses under the long read GS FLX+ protocol, 247,657 sequences were obtained which passed the initial machine quality control that removed sequences with no data or mixed sequences data. The quality of the raw data was analysed, yielding a sequence distribution with:

- a maximum sequence length of 929bp (Figure A1.6B)
- a GC content of 39% (Figure A1.6D)
- an average quality score of Q = 30

Q is equivalent to Illumina 1.9 quality score; this approximates to an error rate of Q10 = 1/10, Q20=1/100, Q30= 1/1000 and Q40 =1/10,000.

The quality score is not constant through the length of the sequence and substantially degrades through the length of the sequence (Figure A1.6A). At around 550 bp, the interquartile range representing 95% (represented by the yellow boxes in Figure A1.6A) of the sequences starts to fall below Q20. This led to all further analysis using only sequences that were 550 bp, removing shorter sequences as not having sufficient sequence and longer sequences due to error rates.

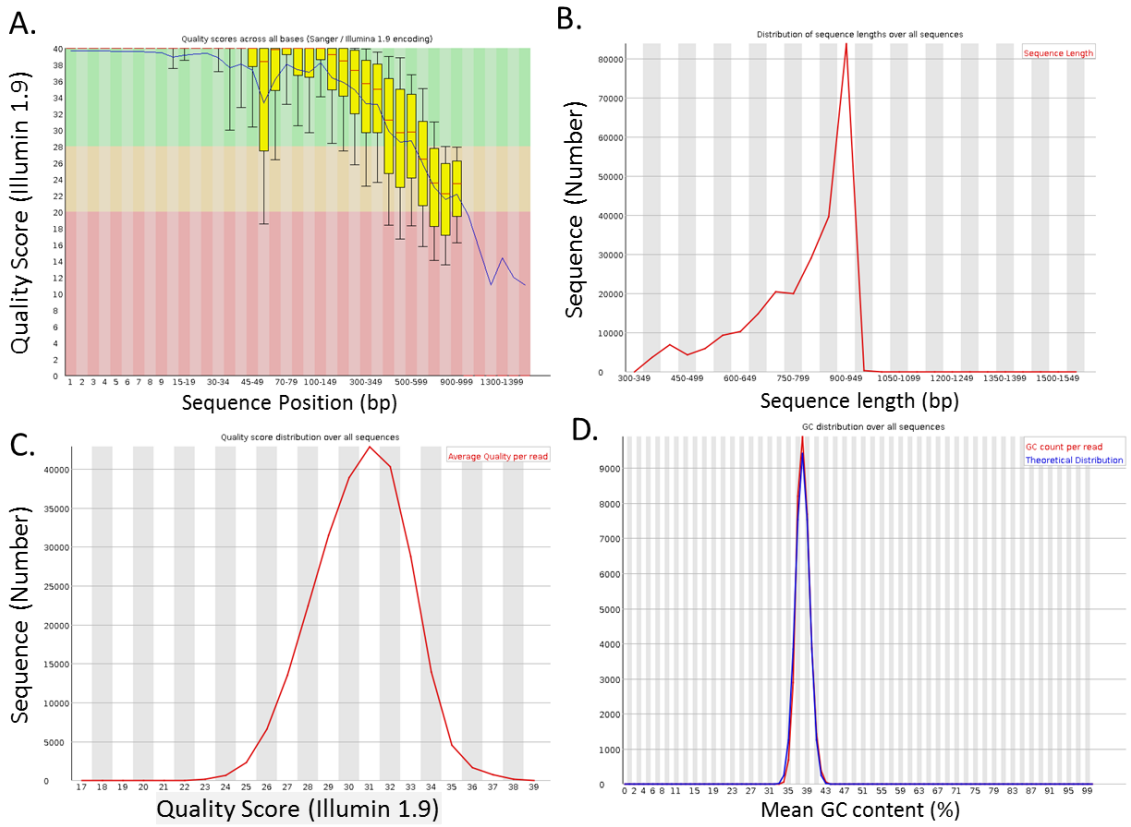


Figure A1.6 Quality analysis of raw GS FLX+ data

The sequences were then divided between those where the sequence was derived from the forward (CfD_F || rbcL-3P_640F (640)) primer sites and those derived from the reverse primer site (DPrbcL7|| rbcL-3P_1538R). The distribution of the sequences is given in Table A1.5.

Sequence representation within the samples was not consistent due to the low concentration of amplified product used for the NGS analysis. In microbial community analyses, samples with <3000K counts would normally be excluded. However, this is done on the basis of representation of the community, and as such the lower complexity of diatom assemblages when compared with their microbial counterparts may allow for lower numbers to be used.

Preliminary analysis and phylogenetic verification

Initially the forward sequences were trimmed and analysed using the PROMpT data analysis workflow described above. An initial analysis was used to derive overall diversity indices and OTUs (99%) for all samples (Table A1.5).

Table A1.5 Results of initial analysis to obtain information about diversity and number of OTUs

Sample ID	Number of sequences		Chao index	Shannon index	OTUs (99%)
	Forward	Reverse			
DTM100	786	705	305.0	3.8	191
DTM113	8,519	8,015	283.0	1.7	140
DTM15	10,388	9,484	240.1	3.5	172
DTM34	5,261	6,584	238.2	3.8	169
DTM42	12,938	12,678	202.6	3.3	126
DTM44	7,122	7,323	38.0	2.3	33
DTM45	2,627	2,698	201.5	2.8	99
DTM47	5,627	5,506	248.1	3.3	152
DTM55	14,664	13,070	57.1	1.1	42
DTM56	8,066	6,681	141.0	2.9	99
DTM69	6,018	3,482	317.6	3.4	199
DTM72	4,647	3,616	48.1	0.6	33
DTM73	1,004	470	65.0	1.8	25
DTM74	357	525	154.2	2.5	96
DTM96	1,829	1,785	71.3	1.2	46
DTM98	2,794	3,133	57.2	2.6	39
DTM99	952	913	175.0	1.3	37

Notes: Further details of the diversity indices are given in Caporaso et al. (2010).

The data were resampled to provide a theoretical calculation of the number of sequences required to optimise these metrics (Figure A1.7). This suggested that only ~500 sequences are required to report on the full OTU composition of the sample, but that about 10-fold additional sequence data are needed to capture the total richness of these samples.

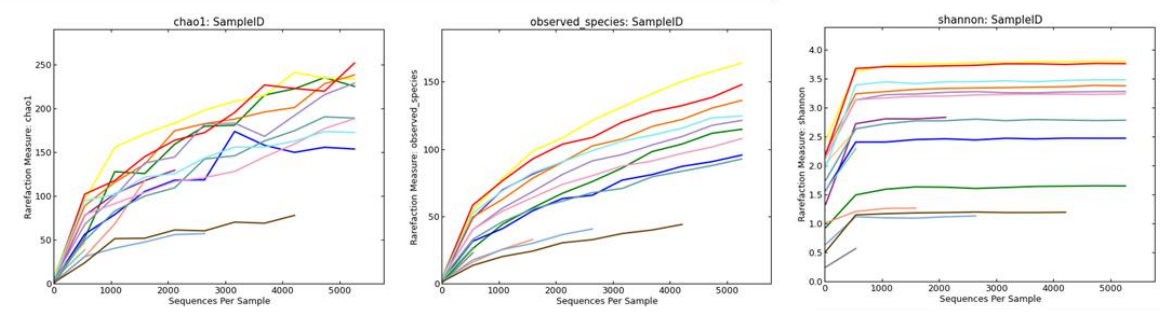


Figure A1.7 Diversity metric analysis of diatom community data

Quality correction of taxonomic annotations from reference data

Initial PROMpT analysis was performed at a 97% and 96% level of taxonomic stringency (the percentage match that was accepted to assign identity to an OTU) using the rbcL reference database (Section 3). Initial analysis showed poor correlation with LM data and calculated TDIs (see Box 1 in Section 1.1).

To investigate the cause of these discrepancies, the top 1% of the OTUs from each sample were aligned against full length reference sequences (~500) and a 'guide' maximum likelihood phylogenetic tree was constructed without bootstrapping. These trees allowed visual interrogation between the OTUs and reference sequences. The decision to perform this analysis with guide trees (not bootstrapped) was based on the practical computation time (days) that would have been required to perform the bootstrapping on trees with 500–800 constituents. The trees generated were navigated manually and the OTUs analysed. This analysis was performed at the following levels.

- If reference sequences were within 3% of the OTU, the identity of the accessions was check against an up-to-date copy of the barcode reference database.
- If no reference sequence within 3% existed within the guide tree, the taxa dictionary was interrogated for species where the frequency of observation matched with the OTU sequence frequency. If a relevant accession was identified, the partial sequence database was integrated into the analytical pipeline used for analysis.
- If no significant match was observed for the sequence, GenBank was interrogated for matches using the BLAST algorithm. Matching sequences with significant provenance were included in the analytical pipeline.
- If no matches were observed, the OTUs were further analysed across samples to see if sequences could be used to infer a species (see specific examples below).

This manual analysis was very useful for identifying major issues with the preliminary analysis. These included the following.

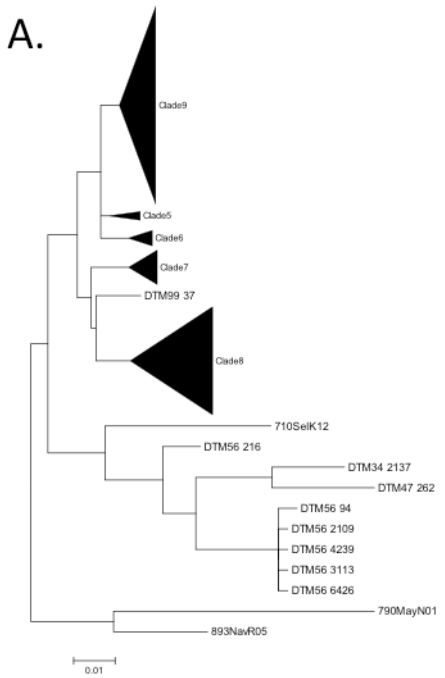
- Nomenclatural issues between reference databases were detected and harmonised, contributing to a significant improvement in species assignment.
- The inclusion of partial sequences (often the forward element of the rbcL-3′ fragment) significantly assisted species identification.
- Inclusion of specific GenBank sequences where appropriate provenance existed improved taxa assignment.
- Some species represent complexes that contained significant cryptic or semi-cryptic diversity that is difficult to detect by LM. By including appropriate OTUs representative of variants within these complexes, significant species reassignment was seen (see Section A1.2.3).
- OTUs for some species could be inferred due to their occurrence frequency and relative phylogenetic position (see Section A1.2.4).
- Significant numbers of Xanthophyta sequences were observed within the sequence reads. Removal of these reads redressed some significant discrepancies between molecular and taxonomic data (see Section A1.2.5).

These issues were addressed by making subtle adjustment to the PROMpT's analytical code. The code was altered to recognise 3 classes of sequence within its sequence database. These included:

- no prefix – verified sequences within the reference database
- 'g' prefix – GenBank sequences
- 'i' prefix – species inferred by the analysis
- 'n' prefix – non-algal sequences

A1.2.3 Identification of species complexes

Analysis of the individual samples identified a number of species complexes where the taxonomic differentiation is very subtle. This was evident in one species in particular, *Eolima minima* (synonym: *Navicula minima*). Significant OTUs obtained by NGS were associated with the single *Eolima minima* reference sequence. To determine the full diversity of this species, sequences representing 1% of the constitute sequences of each sample and with a close phylogenetic relationship to the *Eolima minima* reference sequence were mined from all samples. A maximum likelihood guide tree was then generated using these OTUs, the *Eolima minima* reference sequence and 3 other closely related sequences (accessions 710, 790 and 893) (Figure A1.8). Clades were then generated representing ~1% divergence, and the inter- and intra-divergence was calculated (Figure A1.8B and Figure A1.8C). This analysis resulted in the inclusion of the additional barcode sequences for *Eolima minima* to the analytical pipeline and reference database ('I' suffix – inferred, 'g' suffix – GenBank: iDTM42_2671, iDTM42_719, iDTM44_274, iDTM42_1080, iDTM96_900, gAM710427, gEF143279, gJQ610175 and gKF959642).



B.

Clade9	0.9%
Clade5	1.3%
Clade6	1.0%
Clade7	1%
Clade8	2.2%

C.

	8	7	9	5
Clade8				
Clade7	0.034			
Clade9	0.04	0.037		
Clade5	0.042	0.035	0.024	
Clade6	0.043	0.032	0.026	0.021

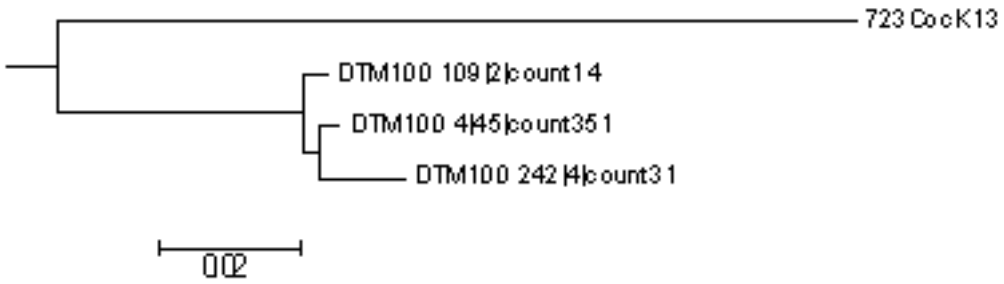
Figure A1.8 Phylogenetic analysis of *Eolima minima complex*: (A) maximum likelihood tree of *Eolima minima* OTUs; (B) estimates of average evolutionary divergence over sequence pairs within groups; and (C) estimates of evolutionary divergence over sequence pairs between groups

A1.2.4 Inferred taxa analysis: *Achnanthes oblongella*

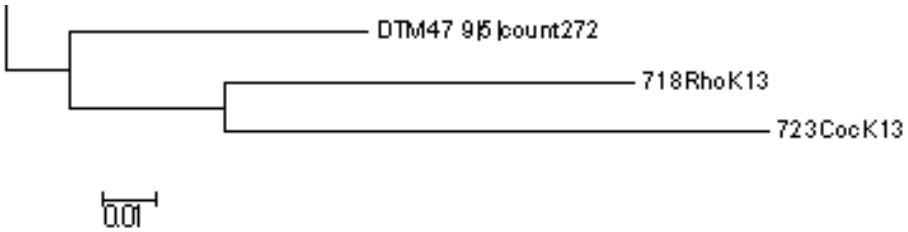
The dominant species occurring within samples were assembled in order to identify species that were missing from the barcode reference database and where there may be good evidence for species inference.

LM analysis had identified *Achnanthes oblongella* in 2 samples at the following frequencies: DTM100 at 84%, DTM47 at 11%. Initially, a highly represented non-assigned clade was identified in DTM100; this consisted of OTU DTM100_109, DTM100_4 and DTM100_242 (Figure A1.9A), which together accounted for 51% of all reads (these percentages have not been adjusted for the Xanthophyta that are also found within the sample). The association with the sequence from accession 723 (*Cocconeis pediculus*) can be ignored due to the significant divergence between these sequences. An appropriate clade was also identified in DTM47 that accounted for 5% of all reads (Figure A1.9B). The sequences were combined into a single tree, which confirmed that the OTUs belonged to a single clade (Figure A1.9C). In response to this analysis, the iDTM100_4 ('i' – inferred) sequence was added to the analytical pipeline and reference database to represent *Achnanthes oblongella*.

(A)



(B)



(C)

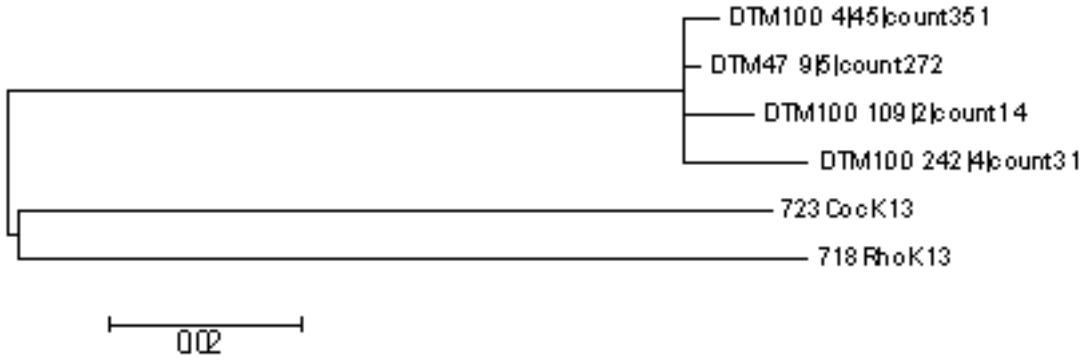


Figure A1.9 Clades of putative *Achnanthes oblongella*: orphan clades were identified individually from DTM100 (A), DTM47 (B) and then the relevant OTUs were combined into a single maximum likelihood guide tree (C)

A1.2.5 Xanthophyta contaminants

The preliminary analysis of DTM98 identified 79% of the sample as Xanthophyta (yellow-green algae), significantly disrupting the proportional representation of the diatom species within the samples. Occurrence of other Xanthophyta was observed within a number of the other samples. Initially, individual rbcL genes from Xanthophyta were added to the analytical pipeline annotated as 'n' or non-diatom, and the workflow was refined to filter out these sequences and provide proportional counts for the diatom constituent alone. In a few samples, all non-stochastic OTUs with sequence representation >5 were analysed, demonstrating that most samples contained some yellow-green algae.

It was considered impractical to mine all the samples for the representative non-diatom sequences and an alternative strategy was adopted whereby GenBank was mined for rbcL genes of Xanthophyta. Testing these sequences for a >90% match against the project's diatom reference database identified 5 sequences whose match to the reference diatom barcodes and phylogenetic context would suggest that these were sequences submitted to GenBank as Xanthophyta but where the current phylogenetic analysis suggested they represented sequences from diatoms. All of these were removed. The remaining 306 Xanthophyta rbcL genes were incorporated into the analysis pipeline with the prefix 'n' to represent non-diatom sequences and added to the reference database to allow them to be pre-filtered prior to proportional calculations.

A1.3 Relating NGS outputs to LM calculated using the TDI

After the inclusion of a range of refinements to the analytical pipeline, the pipeline was rerun on both the forward and reverse sequences from the 17 samples. Analysis of the forward sequences data was used for the comparison with the LM data.

A1.3.1 RA of taxa in analyses by LM and NGS

The hypothesis underlying this work is that the RA of taxa should be similar in an NGS analysis to that obtained by traditional LM analysis. This was tested by examining the RA of genera as estimated by both methods. This in turn assumed that factors that determine the representation of organisms in an NGS analysis are controlled by phylogeny and will not differ markedly between species (though in several of the examples listed below, a single species comprises most of the records for a genus). A number of properties were examined:

- concurrence (whether the same taxon was present in both LM and NGS analyses)
- Spearman's rank correlation between LM and NGS percentages within the dataset
- whether representation was higher in NGS compared with LM, or vice versa
- whether there were any conspicuous outliers

Results are summarised in Table A1.6 with 2 examples, Achnanthidium and Eolimna, also illustrated (Figure A1.10). Both show a general trend of higher representation in LM being matched by higher representation in NGS. In the case of Achnanthidium, however, relative representation in LM is much higher than by NGS (that is, all samples fall below the line indicating slope =1), whereas for Eolimna, representation by NGS

tends to be slightly greater than by LM (samples mostly fall just above line indicating slope = 1). There is also, for Eolimna, one conspicuous outlier where representation by NGS is much higher than would be predicted by LM.

All genera tested showed significant correlations between representation in LM and NGS (Table A1.6) except Nitzschia. NGS gave much greater representation than LM for Cyclotella (centric), Amphora and Eolimna (both raphid), while Melosira (centric), Diatoma and Fragilaria (both araphid), Achnanthidium, Navicula and Rhoicosphenia (all raphid) had greater representation in LM. There were conspicuous outliers for 7 of the 12 genera tested where representation for one or more samples in NGS was substantially higher than predicted from the trend between NGS and LM inferred from other samples. For Fragilaria, Nitzschia and Planothidium, the opposite was also true, with 2 samples showing much greater representation in LM. This may reflect species being detected by LM analysis for which barcodes do not yet exist.

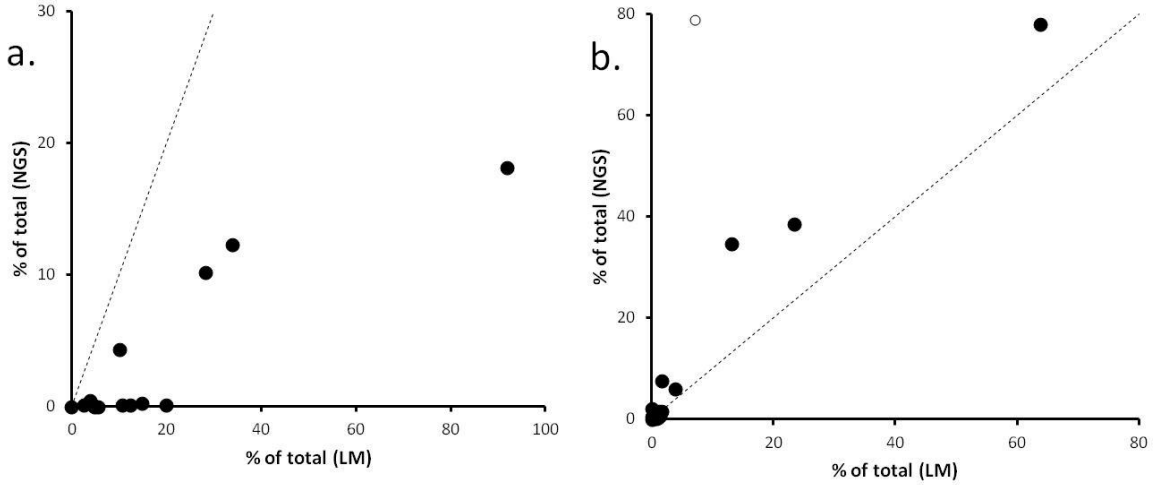


Figure A1.10 Comparison between representation of 2 taxa by traditional LM analysis and NGS: (a) *Achnanthidium*; and (b) *Eolimna*

Notes: Open circle = outlier; dashed line: slope = 1.

Several sources of variability contribute to the differences seen between LM and NGS outputs in this study. Those associated with LM are well understood due to the considerable amount of work over the years. There is an inherent stochastic variability between counts, reflecting the (near-) random distribution of valves on a slide, overlain by between-analyst variation (Prygiel et al. 2000, Kahlert et al. 2009, Kahlert et al. 2012). The latter can be controlled, to some extent, by working within a quality assurance framework (Kelly 2013). Further issues include the underrepresentation of certain taxa due to the dissolution of weakly silicified valves (for example, *Fistulifera*; Zgrundo et al. 2013) and problems caused by contagious distributions of chain-forming genera such as Staurosira and Pseudostaurosira.

An additional set of factors apply when considering NGS. These can be broken down into 2 categories:

- Underlying **real** differences in representation of LM and NGS data (for example, issues with copy number) as well as the possibility of selective amplification of some taxa. The selective amplification may be due to differential efficiency in liberating DNA or subtle differences in primer binding that are exacerbated during the competitive amplification process which occurs during community analysis. This will lead to a systematic deviation from a 1:1 relationship for any particular genus and, in turn, will have knock-on effects on the relationships of other taxa in the sample. A further possibility is that LM analyses do not differentiate between live and

dead cells, with the assumption that living cells will contribute most of the DNA. Although time will elapse before complete DNA degradation, the size on the amplicon means that it is unlikely to survive significantly after the death of the diatom.

- Several of the genera examined also showed occasional outliers, where the representation in one sample greatly exceeded that predicted from the general trend between LM and NGS samples (Figure A1.10). Situations where LM greatly exceeds NGS may indicate 'gaps' in the taxa dictionary that will be filled over time. However, there are also possibilities of occasional overexpression of particular taxa, leading to very high NGS results for a sample.

At this stage no general trend is apparent between the relative representation in LM and NGS based on phylogeny or cell size, though more data and a wider range of analyses (including species- as well as genus-level comparisons) are needed before generalisations can be made.

A1.3.2 Community composition and TDI, as assessed by LM and NGS

The outcome of the process described above is a data matrix in which the composition of the diatom assemblage is expressed in terms of the number of barcodes in an NGS analysis that can be assigned to particular taxa. Of the 17 samples analysed in this study, 8 (47%) had over 90% of the barcodes assigned to binomials in the reference database, and 13 (76%) had over 75% of barcodes assigned. Of the 33 taxa that constituted ≥5% of the total count in at least one LM analysis, 21 (64%) were represented in the reference database, though there are still some issues, particularly where the traditional taxonomy still requires work (for example, *Cocconeis placentula*), where it is suspected that cryptic or semi-cryptic diversity may exist (*Eolimna minima*) or for a few genera where it is known that the reference database is weak, relative to understanding based on morphological taxonomy. This situation should improve as the reference database increases in depth.

Generally, more taxa were identified using LM than NGS (Figure A1.11). This was the case both when the comparison was limited to taxa that could be named using the reference database and when OTUs were used, irrespective of whether a binomial could be applied. More OTUs were recognised by NGS than could be named; the difference ranged from 2 additional OTUs being recorded (DTM96, DTM98 – both with limited diversity due to heavy metals) to 14 (DTM113). DTM113 was also interesting as this was the only sample where a considerably greater number of taxa (as OTUs) were discovered by NGS than by LM.

An NMDS performed using both LM and NGS datasets showed similarities between the positions of samples, as estimated by the composition using the 2 techniques, particularly along the first axis (Figure A1.12; Spearman's rank correlation of axis 1 scores by LM and NGS: 0.57; $p < 0.05$). DTM96 and DTM98 both had very low scores for axis 1 using both methods, possibly reflecting low diversity due to the influence of heavy metal pollution at these sites. Greater differences were observed between LM and NGS approaches for axis 2, with LM analyses generally having lower axis 2 scores (median: -0.18) than NGS analyses (median: 0.24).

If data obtained by the 2 approaches show similar structure in relation to the major environmental gradient (presumed to be water quality), it should follow that ecological indices based on these data should also give similar results. When the TDI is calculated on the RA of taxa to which binomials could be applied via NGS, a significant relationship is obtained (Figure A1.13; Spearman's rank correlation: 0.59; $p < 0.02$).

NGS appears to overestimate the TDI at higher values for reasons that are not entirely clear. The possibility that this is a chance consequence of the subset of samples selected for these preliminary NGS analyses cannot be ruled out. One sample, DTM56, had a much higher TDI value based on LM than that from NGS. This was a diverse sample with a large number of valves belonging to a recently described species, *Platessa bahlsii* (Potapova 2012). If confirmed, this would be the first UK record and, as a consequence, there is no TDI score. However, other taxa in the sample would support the high TDI score assigned.

Table A1.6 Comparison between representation in LM and NGS for common diatom genera

Genus	Maximum RA (LM)	N ≥2% (LM)	Concurrence		Correlation	Slope	Outliers?
			All records	RA >2% only			
Centric diatoms							
Cyclotella	7%	2	65%	100%	0.59 *	NGS >> LM	NGS
Melosira	12%	5	76%	100%	0.86 ***	LM > NGS	×
Araphid diatoms							
Diatoma	5%	2	88%	50%	0.96 ***	LM > NGS	×
Fragilaria	33%	6	65%	67%	0.61 **	LM >> NGS	NGS and LM
Raphid diatoms							
Achnanthidium	92%	14	76%	86%	0.72 **	LM >> NGS	LM
Amphora	94%	12	70%	100%	0.80 ***	NGS > LM	NGS
Eolimna	64%	5	86%	100%	0.82 ***	NGS > LM	NGS
Navicula	18%	14	70%	78%	0.72	LM >> NGS	×
Nitzschia	52%	12	88%	92%	0.44	–	NGS and LM
Planothidium	36%	8	82%	75%	0.60 **	?	NGS and LM
Rhoicosphenia	11%	6	71%	67%	0.70 **	LM > NGS	NGS
Surirella	4%	3	53%	0%	0.50 *	?	×

Notes: Maximum RA (LM) indicates the highest value recorded in the 17 samples in the original analyses using LM to indicate the range over which NGS results should be expected; N ≥ 2% (LM) is also included as this is the effective 'confidence limit' for 'presence' LM analyses based on 300 valves (values lower than this may not be recorded in replicate analyses).
Concurrence (whether the taxon was recorded in both LM and NGS analyses) is presented for all samples and for only those samples where representation in LM exceeds 2%.
Spearman's rank correlation is presented with statistical confidence (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; N.S. = not significant).
'Slope' indicates whether the slope of LM v NGS is greater or less than 1.
'Outlier?' is based on a visual assessment of whether samples deviate from the main trend of the data.

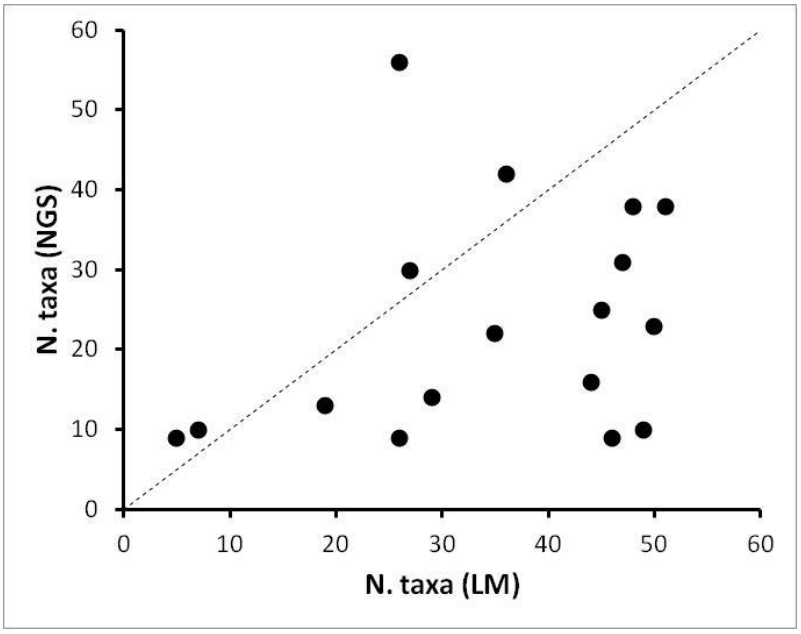


Figure A1.11 Comparison between number of taxa (N. taxa) recorded by LM and NGS

Notes: NGS taxa are based on OTUs; see text for more details.
The diagonal line indicates slope = 1.

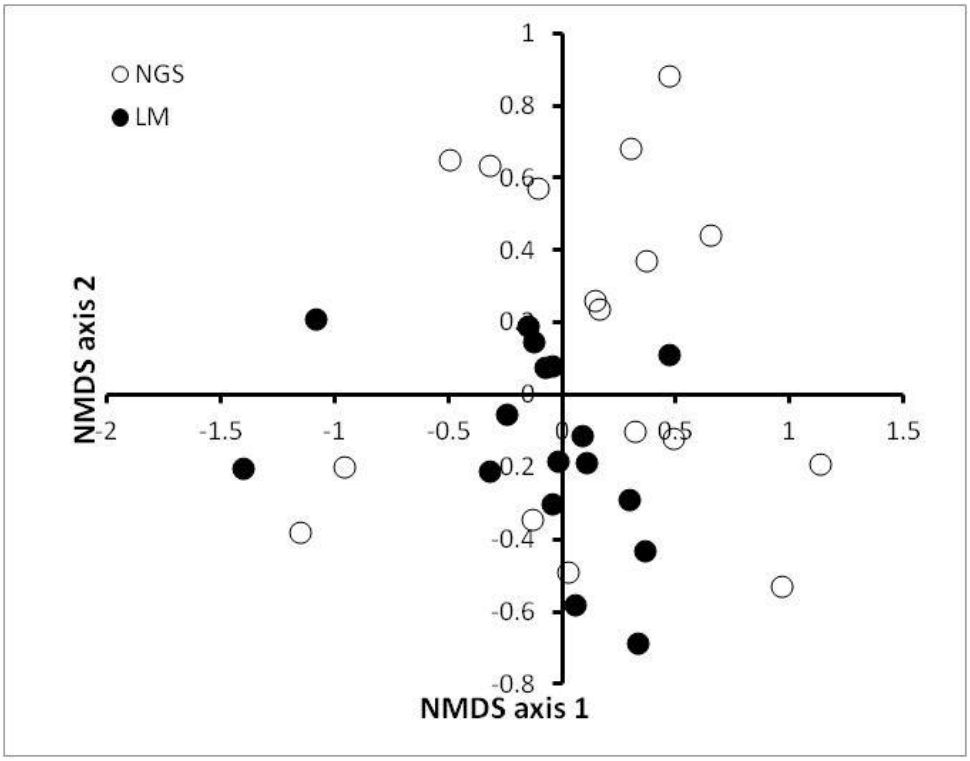


Figure A1.12 First 2 axes of NMDS analysis using combined data from samples analysed by LM and NGS

Variation between TDI values calculated with LM and NGS data may be due to:

- incomplete coverage in the reference database
- a number of factors that influence how the taxa are recorded in either LM or NGS

To differentiate between these 2 causes, TDI values were computed from LM data using only taxa that were also recorded in the NGS analyses. This reduced the total number of taxa from 164 to 64 although, as the TDI is based on a weighted average equation that favours the most abundant taxa, which are well covered by the reference database (see above), this only had a small effect on the TDI calculation based on LM data (Spearman's rank correlation: 0.97; $p < 0.001$). Although further work to improve coverage of the reference database would be useful, this analysis suggests that the most important problem is differences in how taxa are recorded in NGS and LM.

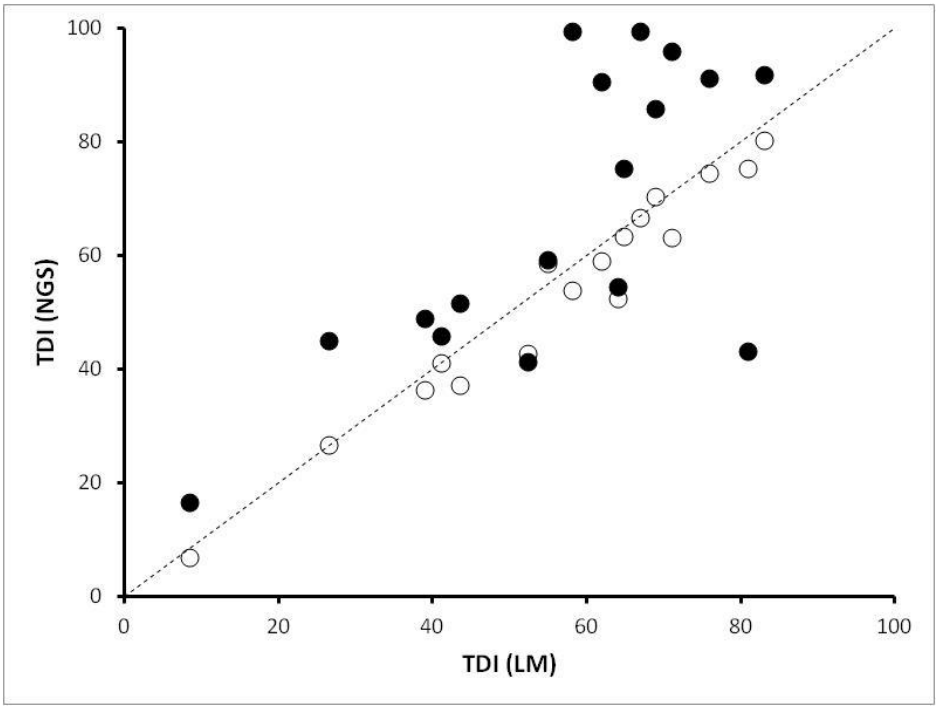


Figure A1.13 Comparison of TDI values computed using traditional LM analyses and NGS

Notes: The x axis shows the TDI based on all taxa identified by LM.
Closed circles show the calculation based on NGS outputs, while open circles show the equivalent value of the TDI based on LM data but using only the taxa available for the NGS calculations.
The diagonal line shows slope = 1.

A1.4 Discussion

Within this proof of concept work, the following main outputs were developed and tested.

- A diatom reference database of rbcL barcodes from known diatom species was developed by isolating and culturing diatom species from water bodies of different ecological quality (see Section 3 for full details).
- A field sampling strategy for collecting and preserving diatom samples was established (Appendix 2).
- A protocol for DNA extraction and amplification from environmental samples was developed. This has since been adapted for automation (Appendix 9).
- Work to develop shorter amplicons compatible with alternative NGS was also undertaken (Section A1.1.5). Use of shorter amplicons would

significantly reduce the cost of NGS, making it a much more attractive proposition for routine analyses. However, a shorter amplicon could not be developed satisfactorily during the lifetime of the proof of concept study but has since been refined (see Section 4).

- A series of bioinformatics procedures were developed to match NGS output with the relevant species in the barcode reference database. This included:
  - steps to screen out non-diatom algae at an early stage
  - routines to manipulate data and produce an output in a form suitable for further analyses (Section A1.2)
- The final stage of the proof of concept project was to relate the NGS outputs to LM results for the same samples (Section A1.3). NGS samples tended to recognise fewer taxa than LM (though this may change as the system develops) and the proportional representation of taxa was often different. However, the 2 datasets showed a similar structure when evaluated using NMDS and TDI values computed from NGS data were significantly correlated (Spearman's rank correlation: 0.59).

The proof of concept project had to overcome several methodological challenges including the generation of a 'gold standard' reference database, which is the backbone for any phylogenetic analysis. The culturing stages, though effective, had a tendency to favour fast-growing cosmopolitan species and, unless substantial effort is devoted to diatom 'horticulture', it is unlikely that this method will provide barcodes for slower-growing species with more specialised requirements, many of which are typically found at low RAs.

An unexpected outcome from this project was the ability to 'discover' new species directly from field samples using NGS, bypassing the need to culture strains (Sections A1.2.3 and A1.2.4). Species discovery or barcode assignment by NGS needs to be used with care and it is suggested the development of a series of carefully considered rules covering issues such as replication, metadata and phylogenetic context to ensure that any inferred barcodes are robust.

- The issues that need to be addressed in the next phase of research are mainly associated with the development of a new suite of primers that would support the amplification of a smaller (300–500 bp) amplicon compatible with the full range of NGS sequencing technologies. This would have 3 advantages:
- significantly improving the cost-effectiveness of the NGS analysis
- increasing the depth of the sampling performed by NGS
- removing the technical error associated with the GS FLX+ platform

Appendix 2: Establishing and deploying a field sampling strategy for diatom community samples compatible with NGS analysis for use by Environment Agency sampling teams

A.2.1 Introduction

The project team engaged with Environment Agency Area sampling staff to establish and deploy a robust procedure for sample collection of diatom community samples compatible with NGS analysis. This process needed to include:

- a Standard Operating Procedure for the sampling teams (Section A.2.3)
- a dispatch protocol to ensure that the samples arrive at the archiving and processing centre

A.2.2 Sample preservation trial

For samples collected for LM analysis, changes to the diatom assemblage after sampling (for example, due to differential growth rates, microbial activity and grazing) is prevented by the addition of either Lugol's iodine or industrial methylated spirits (IMS). However, previous analysis had shown that these methods were incompatible with DNA extraction. An alternative option was to preserve samples by cooling with an ice pack at -4°C. However, sampling teams had no way of maintaining ice packs at low temperatures in the field without adding substantially to the weight of sample batches, making postal delivery impractical. A 'chemical' freezer bag that could be activated in the field and was relatively light was trialled. Even given optimal delivery times, however, the sample would still arrive having experienced substantial time at room temperature.

An additional trial of preservatives was therefore performed. All samples were collected from the River Taff (51.486991, -3.189138) and the following treatment applied.

1. Diatom suspension (15ml) was placed into an empty 15ml Falcon tube, transported directly to the laboratory, where it was centrifuged immediately to pellet the cellular material and frozen at -20°C.
2. Diatom sample (7.5ml) was added to an equal volume of IMS and the sample left at room temperature for 72 hours.
3. Diatom sample (7.5ml) was added to an equal volume of ethanol and the sample left at room temperature for 72 hours.

4. Diatom sample (7.5ml) was added to an equal volume of nucleic acid preservative (3.5 M ammonium sulphate, 17 mM sodium citrate and 13 mM EDTA) and the sample left at room temperature for 72 hours.

After 72 hours at room temperature, samples 2–4 were centrifuged to pellet the cellular material and frozen at -20°C. All samples were then defrosted and DNA extracted using a hybrid glass bead lysis into a DTAB extraction method (Fawley and Fawley 2004). The hybrid protocol yielded a simple and rapid technique for extraction of DNA from diatoms followed by a DNeasy column purification.

The expectation was that 50:50 volume/volume (v/v) addition of nucleic acid preservative and ethanol to the sample would suspend all biological activity, thus preserving community structure. The DNA samples were analysed for DNA recovery and purified using spectral analysis (Figure A2.1). These results confirmed that:

- no DNA could be recovered from IMS preserved material
- both ethanol and the nucleic acid preservative did preserve the integrity

Although the yield of DNA using these methods is half of that achieved for the fresh sample, this represents an equivalent quality when adjusting for the volume of preservative added.

Ethanol and the nucleic acid preservative samples were subsequently successfully tested for the ability to act as a template for rbcL-3′ amplification (Figure A.2.2) and both provide a method for robust sample collection.

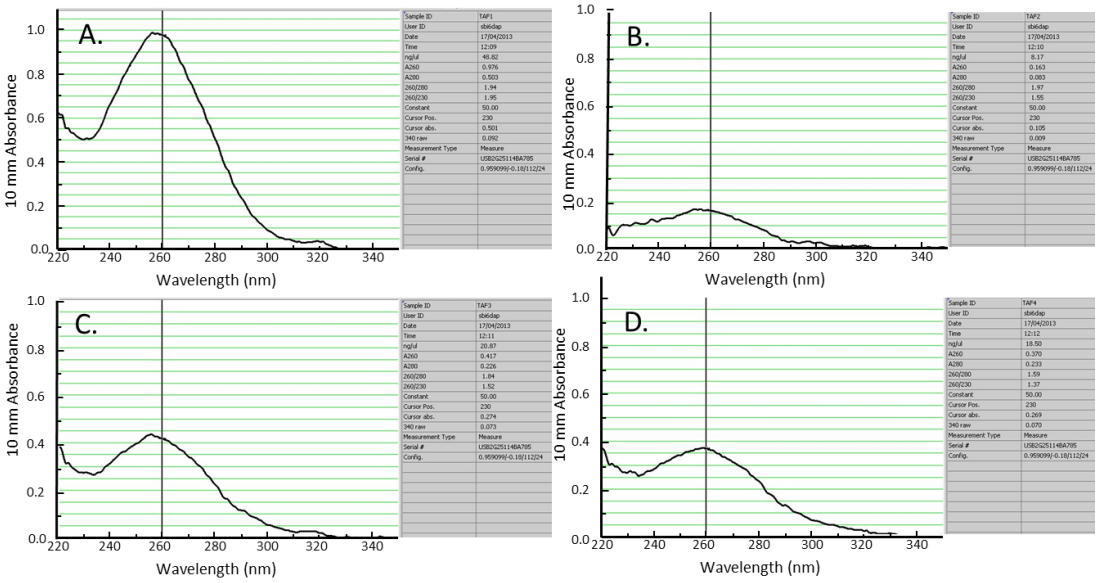


Figure A2.1 Compatibility test for diatom preservation with DNA extraction. DNA was extracted and analysed from diatoms subsampled from an individual community preparation and either immediately centrifuged and preserved at -20ºC (A) or maintained for 72 hours at room temperature with an equal volume of IMS (B), ethanol (C) and nucleic acid preservative (D).

Notes: DNA concentrations were: (A) 48ng μl^{-1} fresh sample; (B) IMS 8.2 ng μl^{-1} (note this is not accurate due to degradation); (C) ethanol 20.9 ng μl^{-1}; and (D) nucleic acid preservative 18.5 ng μl^{-1}.

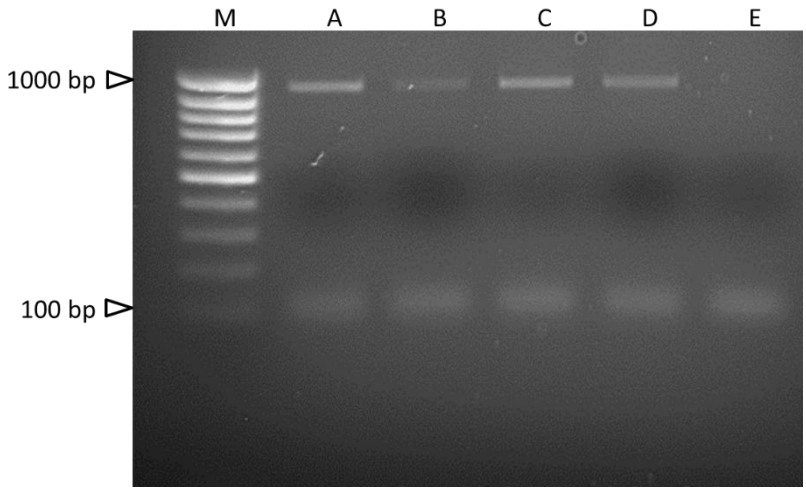


Figure A2.2 rbcL amplification from diatom assemblages after preservation treatments. DNA extracted from environmental samples after differential preservation were amplified using the rbcL-3′ primers previous reported by Hamsher et al. (2011). Lanes show the following samples: (M) 100 bp ladder; (A) fresh sample; (B) 72 hours IMS; (C) 72 hours ethanol; (D) 72 hours nucleic acid preservative; and (E) control PCR with no template DNA.

A.2.3 Standard Operating Procedure: Diatom sample preservation for molecular analysis

Purpose

This document describes the process you must follow when collecting and preserving diatom samples for molecular analysis.

Scope

This method is applicable to all diatom samples collected by Environment Agency staff from rivers and lakes in the UK for DNA analysis.

Justification of method

The chemical added to diatom samples used for DNA analysis is different from the chemical (Lugol's iodine) added to diatom samples collected for standard analysis. It is necessary to use a different preservative as it stabilises and protects the DNA within the cells of the diatoms. It also eliminates the need to immediately process or freeze the samples.

Health and safety

The preservative used in diatom DNA samples is an aqueous, ammonium sulphate based, non-toxic preservative. It is not classified as hazardous, the risk level is low, it can be disposed of down the sink and there are no restrictions on shipment. It can, however, cause skin irritation. Therefore you must avoid contact with skin and wear gloves when handling. If skin contact occurs, wash hands thoroughly with plenty of water. Please refer to the COSHH risk assessment for further information regarding the health and safety risk of this product.

Equipment and supplies for preservation

- 15ml sterile Falcon tubes
- diatom DNA preservative (can be stored at room temperature)
- plastic Pasteur pipettes
- barcode labels
- gloves

Method summary

Sample collection

Diatom DNA samples must be collected using the standard sampling method described in Operational Instruction 27_07, which is in accordance with CEN (2014a) and Kelly et al. (1998). Once the sample has been collected, mix it and decant 5ml of sample to the Falcon tube (15ml centrifuge tube). The remaining sample can then be decanted into the normal sample container. Both samples must then be appropriately labelled with the sample barcode, Biosys site ID and sample date.

Sample preservation

On return to the laboratory, both sample portions need to be appropriately preserved as follows:

Diatom DNA sample (15ml centrifuge tube):

Important! Wear gloves when handling diatom DNA samples to reduce the risk of skin exposure and to avoid contaminating DNA entering the sample.

1. Add 5ml of the diatom DNA preservative to the sample using a pipette. **Important!** There must be equal volumes of sample liquid and diatom preservative.
2. Replace the same cap onto the sample tube and seal it with Parafilm. **Important!** You must make sure the same cap goes back on the same tube to avoid sample contamination.
3. Invert the tube to mix the contents.
4. Store the sample in the freezer.

Important! It is essential that diatom DNA samples are preserved as quickly as possible after collection to reduce DNA degradation. If a sample will not reach the laboratory for more than 24 hours, consider preserving the sample at the depot.

Standard diatom sample

This portion of the sample can be preserved as normal with Lugol's iodine.

What to do with samples

1. Store the preserved diatom DNA samples in the freezer.

2. Once a batch of at least 10 samples has been collected, these should be couriered to the molecular laboratory that will be carrying out the diatom DNA analysis. To save on shipping cost and if capacity is available in your freezer, please store the samples until the end of the sampling campaign and ship in larger batches (for example, one batch at the end of spring sampling and one at the end of autumn sampling).

Appendix 3: Collection locations

The table below lists the locations from which diatom species were collected to provide strains for the rbcL barcode database.

- ID permits cross-reference to individual strain identities.
- Voucher (box slot) refers to the location of the original slide in the herbarium at the Royal Botanic Gardens, Edinburgh
- BC accession number refers to the location of the original slide in Bowburn Consultancy's herbarium and database (where appropriate).

ID	Locality	Date	NGR	Collector	Original ID
P01	Water of Leith at Currie Rugby Club, Balerno, Midlothian	19 May 2012	NT 164667	David Mann	P1
P02	Streams, Pentland Hills, Green Cleuch, above Balerno, Midlothian	19 May 2012	NT 1862	David Mann	P2
P03	Streams, Pentland Hills, Green Cleuch, above Balerno, Midlothian	19 May 2012	NT 1862	David Mann	P3
P04	Streams, Pentland Hills, Green Cleuch, above Balerno, Midlothian	19 May 2012	NT 1862	David Mann	P4
P05	Streams, Pentland Hills, Green Cleuch, above Balerno, Midlothian	19 May 2012	NT 1862	David Mann	P5
P06	Streams, Pentland Hills, Green Cleuch, above Balerno, Midlothian	19 May 2012	NT 1862	David Mann	P6
P07	Streams, Pentland Hills, Green Cleuch, above Balerno, Midlothian	19 May 2012	NT 1862	David Mann	P7
P08	Streams, Pentland Hills, Green Cleuch, above Balerno, Midlothian	19 May 2012	NT 1862	David Mann	P8
P09	Streams, Pentland Hills, Green Cleuch, above Balerno, Midlothian	19 May 2012	NT 1862	David Mann	P9

ID	Locality	Date	NGR	Collector	Original ID
P10	Streams, Pentland Hills, Green Cleuch, above Balerno, Midlothian	19 May 2012	NT 1862	David Mann	P10
C01	Kinleith Burn, Moidart House, Currie, Edinburgh	22 May 2012	NT 187675	David Mann	C1
C02	Kinleith Burn, Moidart House, Edinburgh	22 May 2012	NT 187675	David Mann	C2
C03	Kinleith Burn, Moidart House, Edinburgh	22 May 2012	NT 187675	David Mann	C3
C04	Kinleith Burn, Moidart House, Edinburgh	22 May 2012	NT 187675	David Mann	C4
C05	Kinleith Burn, Moidart House, Edinburgh	22 May 2012	NT 187675	David Mann	C5
B01	Allt a 'Bhalachain, Argyll and Bute	26 May 2012	NN 2705	David Mann	BT1
B02	Allt a 'Bhalachain, Argyll and Bute	26 May 2012	NN 2705	David Mann	BT2
B03	Allt a 'Bhalachain, Argyll and Bute	26 May 2012	NN 2705	David Mann	BT3
B04	Allt a 'Bhalachain, Argyll and Bute	26 May 2012	NN 2705	David Mann	BT4
B05	Allt a 'Bhalachain, Argyll and Bute	26 May 2012	NN 2705	David Mann	BT5
B06	Allt a 'Bhalachain, Argyll and Bute	26 May 2012	NN 2705	David Mann	BT6
B07	Allt a 'Bhalachain, Argyll and Bute	26 May 2012	NN 2705	David Mann	BT7
B08	Allt a 'Bhalachain, Argyll and Bute	26 May 2012	NN 2705	David Mann	BT8
B09	Allt a 'Bhalachain, Argyll and Bute	26 May 2012	NN 2705	David Mann	BT9
B10	Allt a 'Bhalachain, Argyll and Bute	3 June 2012	NN 2705	David Mann	BN1

ID	Locality	Date	NGR	Collector	Original ID
B11	Allt a 'Bhalachain, Argyll and Bute	3 June 2012	NN 2705	David Mann	BN2
C06	River Almond at Cramond, Edinburgh	June 2012	NT 183764	David Mann	CRA1
C07	River Almond at Cramond, Edinburgh	June 2012	NT 183764	David Mann	CRA2
M01	Kinleith Burn, Moidart House, Edinburgh	June 2012	NT 187675	David Mann	M1
M02	Kinleith Burn, Moidart House, Edinburgh	June 2012	NT 187675	David Mann	M2
W01	Water of Leith, Currie, Edinburgh	June 2012	NT 183677	David Mann	WL1
W02	Water of Leith, Currie, Edinburgh	June 2012	NT 183677	David Mann	WL2
T01	River Tay, near Aberfeldy, Perth and Kinross	4 June 2012	–	Cristine Rosique	T01
T02	River Tay, near Aberfeldy, Perth and Kinross, slow flow	14 June 2012	–	Cristine Rosique	T02
T03	River Tay, near Aberfeldy, Perth and Kinross, fast flow	14 June 2012	–	Cristine Rosique	T03
K01	Eudon Beck	20 June 2012	NZ 067300	Martyn Kelly	K01
K02	River Browney, Sunderland Bridge ()112257	20 June 2012	NZ 267383	Martyn Kelly	K02
P11	River Tay, Pitlochry, Perth and Kinross	8 July 2012	–	Shinya Sato	P11
P12	River Tay, Pitlochry, Perth and Kinross	8 July 2012	–	Shinya Sato	P12
P13	River Tay, Pitlochry, Perth and Kinross	8 July 2012	–	Shinya Sato	P13
K03	River Ehen, 'scout camp'	12 August 2012	NY 087153	Martyn Kelly	K03
K04	River Ehen, 'Mill, footbridge'	12 August 2012	NY 081152	Martyn Kelly	K04

ID	Locality	Date	NGR	Collector	Original ID
K05	River Ehen, ‘oxbow’	12 August 2012	NY 072157	Martyn Kelly	K05
K06	Cheriton Stream, Cheriton	19 September 2012	SU 5829 2849	Martyn Kelly	A
K07	River Dever, Branbury (112277)	19 September 2012	SU 4215 42230	Martyn Kelly	B
K08	Pillhill Brook, Upper Clatford (112278)	19 September 2012	SU 35111 44201	Martyn Kelly	C
K09	River Anton, Andover, ‘KFC’	19 September 2012	SU 36446 46388	Martyn Kelly	D
K10	Lambourn, Bagnor (112280)	19 September 2012	SU 4519 6928	Martyn Kelly	E
K11	River Kennet, Stitchcombe Mill	19 September 2012	SU 1676 6870	Martyn Kelly	F
K12	River Wylye, Kingston Deverill	19 September 2012	ST 844372	Martyn Kelly	G
K13	River Wylye, Henford Marsh	19 September 2012	ST 878438	Martyn Kelly	H
B12	Inveruglas Water, by Ben Vane, Argyll and Bute	23 September 2012	NT 2909	David Mann	SL
B13	Inveruglas Water, by Ben Vane, Argyll and Bute	23 September 2012	NT 2909	David Mann	SL inv
B14	Allt Coiregrogain, by Ben Vane, Argyll and Bute	23 September 2012	NT 2909	David Mann	BV str
B15	Allt Coiregrogain, by Ben Vane, Argyll and Bute	23 September 2012	NT 2909	David Mann	BV
N01	Wooler Water near Wooler, Northumbria	28 October 2012	–	David Mann	1
N02	Wooler Water near Wooler, Northumbria	28 October 2012	–	David Mann	2
N03	River near Wooler, Northumbria	28 October 2012	–	David Mann	3
N04	Harthope Burn, Northumbria	28 October 2012	NT 973246	David Mann	4

ID	Locality	Date	NGR	Collector	Original ID
N05	Harthope Burn, Northumbria	28 October 2012	NT 973246	David Mann	5

Appendix 4: Diatom species from which rbcL barcodes obtained

Species	Authority	Number Strains
Achnanthes_pseudoswazi	J.R.Carter 1963	1
Achnanthidium_caledonicum	(Lange-Bertalot) Lange-Bertalot 1999	2
Achnanthidium_lineare	W. Smith; 1855	1
Achnanthidium_minutissimum	(Kützing) Czarnecki 1994	88
Achnanthidium_sp.	Kützing 1844	1
Adlafia_bryophila	(Petersen) Lange-Bertalot In Moser et al. 1997	1
Adlafia_minuscula	(Grunow) Lange-Bertalot in Lange-Bertalot and Genkal 1999	2
Amphora_pediculus	(Kützing) Grunow in Schmid et al. 1875	3
Brachysira_neoexilis	Lange-Bertalot in Lange-Bertalot and Moser 1994	2
Brachysira_vitrea	(Grunow) R.Ross in B.Hartley 1986	1
Cocconeis_pediculus	Ehrenberg 1838	1
Cocconeis_placentula	Ehrenberg 1838	1
Cyclotella_meneghiniana	Kützing 1844	7
Cymbella _sp.	C.Agardh 1830	1
Cymbella_cymbiformis	C. Agardh 1830	1
Diatoma_moniliformis	Kützing 1833	6
Diatoma_tenuis	Agardh 1812	2
Diatoma_vulgaris	Agardh 1812	3
Encyonema_minutum	(Hilse in Rabenhorst) D.G.Mann in Round et al. 1990	4
Encyonema_silesiacum	(Bleisch in Rabenhorst) D.G.Mann in Round et al. 1990	4
Encyonema_sp.	Kützing 1833	6
Encyonopsis_falaisensis	(Grunow) Krammer 1997	2
Encyonopsis_microcephala	(Grunow) Krammer 1997	1
Eunotia_arcus	Ehrenberg 1837	1

Species	Authority	Number Strains
Eunotia_bilunaris	(Ehrenberg) Mills 1934	7
Eunotia_exigua	(Brébisson) Rabenhorst 1864	4
Eunotia_implicata	Norpel, Lange-Bertalot et Alles 1991	1
Eunotia_minor	(Kützing) Grunow in Van Heurck 1881	3
Fistulifera_solaris	S.Mayama, M.Matsumoto, K.Nemoto and T.Tanaka in Matsumoto et al. 2014	1
Fragilaria_capucina	Desmazières 1925	3
Fragilaria_crotonensis	Kitton 1869	1
Fragilaria_gracilis	Østrup 1910	67
Fragilaria_sp.	Lyngbye 1819	7
Fragilaria_mesolepta	Rabenhorst 1861	1
Fragilaria_pararumpens	Lange-Bertalot, G. Hofmann et Werum 2011	19
Fragilaria_perminuta	(Grunow) Lange-Bertalot 2000	2
Fragilaria_radians	(Kützing) Lange-Bertalot in Hofmann et al. 2011	1
Fragilaria_rumpens	(Kützing) Carlson 1913	2
Fragilaria_tenera	(W. Smith) Lange-Bertalot 1980	1
Fragilaria_vaucheriae	(Kützing) Petersen 1938	5
Frustulia_crassinervia	(Brébisson) Lange-Bertalot and Krammer in Lange-Bertalot and Metzeltin 1996	2
Gomphonema_acuminatum	Ehrenberg 1836	3
Gomphonema_sp.	Ehrenberg 1832	2
Gomphonema_clavatum	Ehrenberg 1832	2
Gomphonema_cymbelliclinum	E.Reichardt and Lange-Bertalot 1999	1
Gomphonema_exilissimum	(Grunow) Lange-Bertalot and E. Reichardt 1996	5
Gomphonema_hebridense	Gregory 1854	8
Gomphonema_micropus	Kützing 1844	2
Gomphonema_minutum	(C. Agardh) C. Agardh 1831	2
Gomphonema_parvulum	(Kützing) Kützing 1849	21

Species	Authority	Number Strains
Gomphonema_pseudobohemicum	Lange-Bertalot and E. Reichardt 1993	1
Gomphonema_pumilum	(Grunow) E. Reichardt and Lange-Bertalot 1991	1
Gomphonema_truncatum	Ehrenberg 1832	2
Hannaea_arcus	R.M.Patrick in R.M.Patrick et Reimer 1966	2
Mayamaea_atomus	(Kützing) Lange-Bertalot 1997	2
Melosira_varians	C. Agardh 1827	8
Meridion_circulare	(Greville) C.Agardh 1831	1
Navicula tripunctata	(O.F.Müller) Bory 1822	1
Navicula_angusta	Grunow 1860	1
Navicula_capitata	Ehrenberg 1838	1
Navicula_cryptocephala	Kützing 1844	3
Navicula_cryptotenella	Lange-Bertalot 1985	2
Navicula_gregaria	Donkin 1861	10
Navicula_lanceolata	(Agardh) Ehrenberg 1838	45
Navicula_radiosa	Kützing 1844	7
Navicula_sp.	Bory 1822	2
Navicula_trivialis	Lange-Bertalot 1980	1
Navicula_upsaliensis	(Grunow) Peragallo 1903	1
Navicula_veneta	Kützing 1844	1
Neidium_dubium	(Ehrenberg) Cleve 1894	1
Nitzschia_acicularis	(Kützing) W.Smith 1853	1
Nitzschia_alicae	Hlúbiková and Ector in Hlúbiková et al. 2009	2
Nitzschia_amphibia	Grunow 1862	4
Nitzschia_capitellata	Hustedt in A.Schmidt et al. 1922	1
Nitzschia_dissipata	(Kützing) Grunow 1862	4
Nitzschia_fonticola	Grunow in Van Heurck 1881	4
Nitzschia_frustulum	(Kützing) Grunow in Cleve and Grunow 1880	1
Nitzschia_hantzschiana	Rabenhorst 1860	2
Nitzschia_linearis	(Agardh) W.Smith 1853	7

Species	Authority	Number Strains
Nitzschia_palea	(Kützing) W.Smith 1856	35
Nitzschia_paleacea	Grunow in Van Heurck 1881	2
Nitzschia_perminuta	(Grunow) M. Peragallo 1903	1
Nitzschia_pusilla	(Kützing) Grunow em. Lange-Bertalot 1976	1
Nitzschia_recta	Hantzsch ex. Rabenhorst 1861	2
Nitzschia_romana	Grunow in Van Heurck 1881	1
Nitzschia_sigma	(Kützing) W.Smith 1853	1
Nitzschia_sigmoidea	(Nitzsch) W.Smith 1853	1
Nitzschia_sociabilis	Hustedt 1957	2
Nitzschia_sp.	Hassall 1845	3
Nitzschia_sublinearis	Hustedt 1930	1
Nitzschia_vermicularoides	Lange-Bertalot	1
Parlibellus_protracta	(Grunow) Witkowski, Lange-Bertalot and Metzeltin 2000	1
Peronia_fibula	(Brébisson ex.Kützing) R.Ross 1956	1
Pinnularia_grunowii	Krammer 2000	1
Pinnularia_microstauron	(Ehrenberg) Cleve 1891	3
Pinnularia_neomajor	Krammer 1992	1
Pinnularia_sp.	Ehrenberg 1843	3
Pinnularia_subcapitata	Gregory 1856	4
Planothidium_frequentissimum	(Lange-Bertalot) Round and L.Bukhtiyarova 1996	1
Planothidium_lanceolatum	(Brébisson) Lange-Bertalot 1999	4
Psammothidium_bioretii	(Germain) L.Bukhtiyarova and Round 1996	1
Pseudostaurosira_brevistriata	(Grunow in Van Heurck) D.M.Williams and Round 1987	2
Reimeria_sinuata	(Gregory) Kociolek and Stoermer 1987	3
Rhoicosphenia_abbreviata	(C.Agardh) Lange-Bertalot 1980	1
Sellaphora_joubaudii	(H.Germain) Aboal in Aboal et al. 2003	1
Sellaphora_seminulum	(Grunow) D.G.Mann 1989	1
Stauroneis_phoenicenteron	(Nitzsch) Ehrenberg 1843	1

Species	Authority	Number Strains
Staurosira_cf_subsalina	(Hustedt) Lange-Bertalot 2000	1
Staurosira_elliptica	(Schumann) D.M. Williams and Round(1987)	2
Staurosira_venter	(Ehrenberg) Grunow in Pantocsek 1889	5
Stephanodiscus_hantzschii	Grunow in Cleve and Grunow 1880	1
Surirella_angusta	Kützing 1844	3
Surirella_brebissonii	Krammer and Lange-Bertalot 1987	7
Tabellaria_flocculosa	(Roth) Kützing 1844	8
Thalassiosira_pseudonana	Hasle and Heimdal 1970	1
Thalassiosira_weissfloggii	(Grunow) Fryxell and Hasle 1977	3
Tryblionella_debilis	Arnott in O'Meara 1873	1
Ulnaria_acus	(Kützing) Aboal in Aboal, Alvarez Cobelas, Cambra and Ector 2003	6
Ulnaria_ulna	(Nitzsch) P.Compère in Jahn et al. 2001	12

Appendix 5: Diatom taxa whose identities were inferred by comparing NGS and LM outputs

Species	Authority	Number Strains
Platessa_conspicua	(A. Meyer) Lange-Bertalot 2004	1
Achnanthes_oblongella	Østrup 1902	1
Achnanthidium_minutissimum	(Kützing) Czarnecki 1994	1
Actinocyclus_sp.	Ehrenberg 1837	1
Diatoma_sp.	Bory 1824	1
Eolimna_minima	(Grunow) Lange-Bertalot 1998	5
Eunotia_cf_formica	Ehrenberg 1843	1

Appendix 6: Diatom barcodes added from published sources

Species	Authority	Number Strains	Source [1]
Achnanthes coarctata	(Brébisson) Grunow in Cleve and Grunow 1880	1	R-SYST
Actinocyclus_sp.	Ehrenberg 1837	1	GenBank
Amphora_pediculus	(Kützing) Grunow in Schmid et al. 1875	1	GenBank
Asterionella formosa	Hassall 1855	1	R-SYST
Aulacoseira granulata	(Ehrenberg) Simonsen 1979	1	R-SYST
Bacillaria paxillifer	(Müller) Hendey 1951	1	R-SYST
Caloneis limosa	(Kützing) R.M.Patrick in R.M.Patrick and Reimer 1966	1	R-SYST
Craticula accomoda	(Hustedt) D.G.Mann in Round et al. 1990	1	R-SYST
Ctenophora pulchella	(Ralfs ex.Kützing) D.M.Williams and Round 1986	1	R-SYST
Cyclostephanos dubius	(Fricke) Round 1982	1	R-SYST
Cyclotella_distinguenda	Hustedt 1927	1	GenBank
Cymatopleura solea	(Brébisson) W.Smith 1851	1	R-SYST
Cymbopleura naviculiformis	(Auerswald) Krammer 2003	1	R-SYST
Denticula kuetzingii	Grunow 1862	1	R-SYST
Denticula_sp.	Kützing 1844	1	GenBank
Didymosphenia geminata	(Lyngbye) M.Schmidt 1899	1	R-SYST
Diploneis subovalis	Cleve 1894	1	R-SYST
Ellerbeckia sp.	R.M.Crawford 1988	1	R-SYST
Eolimna_minima	(Grunow) Lange-Bertalot 1998	3	GenBank
Eolimna_sp_Styx	Lange-Bertalot and W. Schiller in W. Schiller and Lange-Bertalot 1997	1	GenBank
Epithemia sorex	Kützing 1844	1	R-SYST
Eucocconeis laevis	(Østrup) Lange-Bertalot 1999	1	R-SYST
Eunotia_formica	Ehrenberg 1843	1	GenBank

Species	Authority	Number Strains	Source [1]
Fallacia pygmaea	(Kützing) Stickle and D.G.Mann in Round et al. 1990	1	R-SYST
Fistulifera_pelliculosa	(Brébisson ex Kützing) Lange-Bertalot 1997	4	GenBank
Fistulifera_saprophila	(Lange-Bertalot and Bonik) Lange-Bertalot 1997	2	GenBank
Fragilariforma virescens	(Ralfs) D.M.Williams and Round 1988	1	R-SYST
Geissleria decussis	(Hustedt) Lange-Bertalot and Metzeltin 1996	1	R-SYST
Halamphora montana	(Krasske) Levkov 2009	1	R-SYST
Kareyevia ploenensis	(Hustedt) L. Bukhtiyarova 1999	1	R-SYST
Mastogloia sp.	G.H.K.Thwaites in W.Smith 1856	1	R-SYST
Navicula_tripunctata	(O.F.Müller) Bory 1822	2	GenBank
Neidium affine	(Ehrenberg) Pfitzer 1871	1	R-SYST
Nitzschia_inconspicua	Grunow 1862	68	GenBank
Nitzschia_soratensis	E. Morales and Vis 2007	10	GenBank
Parlibellus hamulifer	(Grunow) E.J. Cox 1988	1	R-SYST
Placoneis clementis	(Grunow) E.J.Cox 1987	1	R-SYST
Rhopalodia gibba	(Ehrenberg) O.Müll. 1895	1	R-SYST
Staurosira_construens	Ehrenberg 1843	1	GenBank
Staurosira_elliptica	(Schumann) D.M. Williams and Round 1987	1	GenBank
Staurosirella martyi	(Héribaud-Joseph) E.A.Morales and K.M.Manoylov 2006	1	R-SYST
Staurosirella pinnata	(Ehrenberg) D.M.Williams and Round 1987	1	R-SYST
Tabularia fasiculata	(Agardh) D.M.Williams and Round 1986	1	R-SYST
Tryblionella constricta	Gregory 1855	1	R-SYST

Notes: [1] R-SYST = www.rsyst.inra.fr, GenBank = www.ncbi.nlm.nih.gov/genbank

Appendix 7: Xanthophyta barcodes added to the barcode database

Taxon	Authority	Number Strains
Asterosiphon dichotomus	(Kützing) Rieth 1962	1
Botrydiopsis alpina	Vischer 1945	2
Botrydiopsis callosa	Trenkwalder 1975	1
Botrydiopsis constricta	Broady 1976	3
Botrydiopsis intercedens	Pascher 1939	2
Botrydiopsis pyrenoidosa	Trenkwalder 1975	1
Botrydium becherianum	Vischer 1938	2
Botrydium cystosum	Vischer 1938	1
Botrydium granulatum	(Linnaeus) Greville 1830	3
Botrydium stoloniferum	Mitra	3
Botryochloris sp	Borzí 1889	1
Bumilleria exilis	Klebs 1896	2
Bumilleria klebsiana	Pascher 1932	1
Bumilleria sicula	Borzí 1888	2
Bumilleria sp	Borzí 1888	3
Bumilleriopsis filiformis	Vischer 1945	2
Bumilleriopsis cf. filiformis	Vischer 1945	1
Bumilleriopsis peterseniana	Vischer et Pascher 1936	2
Bumilleriopsis pyrenoidosa	(Deason and Bold) Ettl 1978	1
Bumilleriopsis sp	Printz 1914	8
Chlorellidium pyrenoidosum	A.Begum and P.A.Broady 2002	1
Chlorellidium sp.		1
Chlorellidium tetrabotrys	Vischer and Pascher 1937	2
Excentrochloris sp	Vischer and Pascher 1937	5
Goniochloris sculpta	Geitler 1928	1
Heterococcus brevicellularis	Vischer 1945	1
Heterococcus caespitosus	Vischer 1936	5

Taxon	Authority	Number Strains
Heterococcus chodatii	Vischer 1937	1
Heterococcus conicus	Pitschmann 1963	4
Heterococcus crassulus	Vischer 1945	1
Heterococcus fournensis	Vischer 1945	2
Heterococcus cf. fuornensis	Vischer 1945	1
Heterococcus leptosiroides	Pitschmann 1963	1
Heterococcus mainxii	Vischer 1937	2
Heterococcus moniliformis	Vischer 1937	1
Heterococcus pleurococcoides	Pitschmann 1963	3
Heterococcus protonematoides	protonematoides Vischer 1945	3
Heterococcus ramosissimus	Pitschmann 1963	2
Heterococcus sp	Chodat 1908	3
Heterococcus viridis	Chodat 1908	7
Heterothrix debilis	Vischer 1936	1
Mischococcus sphaerocephalus	Vischer 1932	2
Monodus unipapilla	H.Reisigl 1964	1
Ophiocytium capitatum	Wolle 1887	2
Ophiocytium majus	Nägeli 1849	2
Ophiocytium parvulum	(Perty) A.Braun 1855	2
Pleurochloris meiringensis	Vischer 1945	3
Pseudobumilleriopsis pyrenoidosa	Deason and Bold 1960	1
Pseudopleurochloris antarctica	C.Andreoli, I.Moro, N.La Rocca, F.Rigoni, L.Dalla Valle and L.Bargelloni 1999	1
Sphaerosorus composita	L.Moewus	2
Tribonema aequale	Pascher 1925	4
Tribonema affine	(G.S.West) G.S.West 1904	7
Tribonema elegans	Pascher 1925	1
Tribonema intermixtum	Pascher	8
Tribonema microchloron	Ettl	2
Tribonema minus	(Wille) Hazen 1902	3
Tribonema cf. minus	(G.A.Klebs) Hazen 1902	2
Tribonema missouriense		1

Taxon	Authority	Number Strains
Tribonema regulare	Pascher 1939	16
Tribonema sp	Derbès and Solier 1856	11
Tribonema ulotrichoides	Pascher 1925	2
Tribonema utriculosum	(Kützing) Hazen 1902	13
Tribonema viride	Pascher 1925	8
Tribonema vulgare	Pascher 1923	10
Vaucheria aversa	Hassall 1843	1
Vaucheria borealis	Hirn 1900	1
Vaucheria bursata	(O.F.Müller) C.Agardh	5
Vaucheria canalicularis	(Linnaeau) T.A.Christensen 1968	1
Vaucheria compacta	(Collins) Collins in Taylor 1937	1
Vaucheria conifera	T.A.Christensen 1987	1
Vaucheria cornonata	Nordstedt 1879	1
Vaucheria dichotoma	(Linnaeus) Martius 1817	2
Vaucheria dilwynii	(F.Weber et D.Mohr) C.Agardh 1812	1
Vaucheria erythrospora	T.A.Christensen 1956	2
Vaucheria frigida	(Roth) C.Agardh 1824	4
Vaucheria geminata	(Vaucher) de Candolle in Lamarck et de Candolle 1805	2
Vaucheria hamata	(Vaucher) De Candolle in Lamarck and De Candolle 1805	1
Vaucheria litorea	C.Agardh 1823	3
Vaucheria medusa	T.A.Christensen 1952	1
Vaucheria prona	T.A.Christensen 1970	3
Vaucheria pseudogeminata	P.A.Dang. 1939	1
Vaucheria repens	Hassall 1843	2
Vaucheria schleicheri	De Wildeman 1895	1
Vaucheria synandra	Woronin 1869	1
Vaucheria terrestris	(Vaucher) De Candolle in Lamarck and De Candolle 1805	1
Vaucheria walzii	Rothert 1896	1
Vaucheria zapotecana	Bonilla-Rodriguez, Garduno-Solorzano, Martinez-Garcia, Campos, Monsalvo-Reyes and Quintanar-Zuniga 2013	1

Taxon	Authority	Number Strains
Xanthonema bristolianum	Xanthonema bristolianum (Pascher) P.C.Silva 1979	2
Xanthonema cf. bristolianum	Xanthonema bristolianum (Pascher) P.C.Silva 1979	1
Xanthonema debile	(Vischer) P.C.Silva 1979	4
Xanthonema cf. debile	(Vischer) P.C.Silva 1979	3
Xanthonema exile	(G.A.Klebs) P.C.Silva 1979	3
Xanthonema cf. exile	(G.A.Klebs) P.C.Silva 1979	1
Xanthonema hormidioides	(Vischer) P.C.Silva 1979	4
Xanthonema cf. hormidioides	(Vischer) P.C.Silva 1979	1
Xanthonema montanum	(Vischer) P.C.Silva 1979	3
Xanthonema mucicolum	(Ettl) Ettl	2
Xanthonema sessile	(Vinatzer) Ettl and Gärtner 1995	2
Xanthonema solidum	(Vischer) P.C.Silva 1979	3
Xanthonema sp	P.C.Silva 1979	17
Xanthonema tribonematoides	Pascher) P.C.Silva 1979	2
Xanthonema cf. tribonematoides	Pascher) P.C.Silva 1979	1
‘Botryidiopsidaceae’ sp		2
‘Uncultured xanthophyte’		12
‘Xanthophyceae’		2

Appendix 8: Python code written for this project

This code was written in order to calculate the number of correct taxonomic assignments made for each hypothetical amplicon region.

```
# For the diatom alignment taxonomy assessments
# To be run on QIIME server so no biopython

import argparse
import sys
from collections import defaultdict

def main():
    options = parseArguments()
      # The otus and tax files are indexed by the OTU number.
      # Gives an output of each actual sequence in the alignment slice and its taxonomic
assignment.
    sequences = defaultdict(lambda: defaultdict(dict))

    # Grab the names of the sequences in the alignment file.
    alignment_sequences = []
    for line in open(options.alignment,"rU"):
      if line.startswith(">"):
          line = line.rstrip()
          sample = line.strip(">")
          alignment_sequences.append(sample)


    # Load in all the correct taxonomies to the sequences dict
    for line in open(options.alltax,"rU"):
      line = line.rstrip()
      if (line.startswith("Strain")):
          #This is the first line
          pass
      else:
          #Process
          linelist = line.split('\t')
          seqname = linelist[0]
          if seqname in alignment_sequences:
            taxonomy = linelist[1]
            taxonomylist = taxonomy.split(';')
            sequences[seqname]["correct_taxonomy"]["full"] = taxonomy
            sequences[seqname]["correct_taxonomy"]["class"] = taxonomylist[0]
            sequences[seqname]["correct_taxonomy"]["family"] = taxonomylist[1]
            sequences[seqname]["correct_taxonomy"]["genus"] = taxonomylist[2]
            sequences[seqname]["correct_taxonomy"]["species"] = taxonomylist[3]
            sequences[seqname]["correct_taxonomy"]["strain"] = taxonomylist[4]
            #Note: "strain" for the DTM_composite is the seqname

    # Now go through the OTU taxonomy and create a lookup.
    otu_taxonomies = {}
    for line in open(options.tax,"rU"):
```

```
        line = line.rstrip()
        linelist = line.split('\t')
        otu = int(linelist[0])
        taxonomy = linelist[1]
        otu_taxonomies[otu] = taxonomy

    # Now go through the otus and go through each of the samples and assign the
actual taxonomy.
    for line in open(options.otus,"rU"):
        line = line.rstrip()
        linelist = line.split('\t')
        otu = int(linelist[0])
        linelist.pop(0)#linelist now only contains sequence ids.
        # Grab the taxonomy for this otu
        otu_tax = otu_taxonomies[otu]
        if (otu_tax.startswith("No blast hit")):
            otu_tax = "NULL;NULL;NULL;NULL;NULL;"
        # Print otu_tax
        otu_taxlist = otu_tax.split(';')
        # Assign this OTU taxonomy to all sequences associated with this OTU.
        for seq in linelist:
            sequences[seq]["actual_taxonomy"]["full"] = otu_tax
            sequences[seq]["actual_taxonomy"]["class"] = otu_taxlist[0]
            sequences[seq]["actual_taxonomy"]["family"] = otu_taxlist[1]
            sequences[seq]["actual_taxonomy"]["genus"] = otu_taxlist[2]
            sequences[seq]["actual_taxonomy"]["species"] = otu_taxlist[3]
            sequences[seq]["actual_taxonomy"]["strain"] = otu_taxlist[4]

    # Not all DTM taxonomy sequences will have been in the original alignment
    for seq in sequences:
        try:
            actual = sequences[seq]["actual_taxonomy"]["full"]
        except:
            sequences[seq]["actual_taxonomy"]["full"] = "not-in-slice"
            sequences[seq]["actual_taxonomy"]["class"] = "not-in-slice"
            sequences[seq]["actual_taxonomy"]["family"] = "not-in-slice"
            sequences[seq]["actual_taxonomy"]["genus"] = "not-in-slice"
            sequences[seq]["actual_taxonomy"]["species"] = "not-in-slice"
            sequences[seq]["actual_taxonomy"]["strain"] = "not-in-slice"

    for seq in sequences:
        match = "no_match"
        #sequentially get more specific on the match between correct/actual
        if (sequences[seq]["correct_taxonomy"]["class"] ==
sequences[seq]["actual_taxonomy"]["class"]):
            match = "class"
        if (sequences[seq]["correct_taxonomy"]["family"] ==
sequences[seq]["actual_taxonomy"]["family"]):
            match = "family"
        if (sequences[seq]["correct_taxonomy"]["genus"] ==
sequences[seq]["actual_taxonomy"]["genus"]):
            match = "genus"
        if (sequences[seq]["correct_taxonomy"]["species"] ==
sequences[seq]["actual_taxonomy"]["species"]):
            match = "species"
```

```
      if (sequences[seq]["correct_taxonomy"]["strain"] ==
sequences[seq]["actual_taxonomy"]["strain"]):
          match = "strain"
      print seq, sequences[seq]["correct_taxonomy"]["full"],
sequences[seq]["actual_taxonomy"]["full"],match

def parseArguments():
    parser = argparse.ArgumentParser()
    parser.add_argument('-otus', help='The picked otus TEXT file from QIIME. This is
the output of pick_otus.py', required=True)
    parser.add_argument('-tax', help='The OTU taxonomy assignments from QIIME.
This is the output of assign_taxonomy.py', required=True)
    parser.add_argument('-alltax', help='All the sequence taxonomy assignments from
the main taxonomy file input used in assign_taxonomy.py')
    parser.add_argument('-alignment', help='Original alignment slice', required=True)
    args = parser.parse_args()
    return args

if __name__ == '__main__':
    main()
```

Appendix 9: DNA extraction procedure using enzymatic lysis and spin column purification

The methodology given below outlines the extraction procedure for DNA from diatom samples with a manual method using the Qiagen DNeasy® Blood and Tissue kit, and an automated method using the BioRobot® Universal with the QIAamp® Investigator BioRobot® kit.

Note: Samples are received in preservative and stored at -30°C until extraction.

Before beginning the procedure, preheat an incubator to 56°C.

A9.1 Preparation of samples

1. Thaw sample thoroughly.
2. Vortex to create a homogenous mixture.
3. Spin down samples at 3,000*g* for 15 minutes at 5°C.
4. Remove promptly from the centrifuge and check that all material has pelleted.
5. Remove lid and gently tip buffer into waste container, being careful not to disturb the pellet. Then without re-inverting the tube, take a 1ml pipette and remove all excess buffer from the inside of the rim of the tube.
6. Re-invert and wait for the liquid to pool round the pellet and remove the last of the liquid. If at any point the pellet in disturbed re-spin using conditions in step 3.

A9.2 For extraction by hand using Qiagen DNeasy® Blood and Tissue kit:

1. Place approx. 0.05g of the pellet from above procedure into appropriately labelled 1.5ml tube. Repeat for each sample.
2. Add 180µl Buffer ATL and 20µl Proteinase K. Vortex thoroughly.
3. Incubate at 56°C shaking at 100 rpm for 5 hours (or overnight).
4. Vortex for 15 seconds.
5. Add 200µl Buffer AL to the sample. Mix thoroughly by vortexing.
6. Add 200µl ethanol (96–100%). Mix again thoroughly by vortexing.
7. Pipette the mixture from step 6 (including any precipitate) into the DNeasy mini spin column placed in a 2ml collection tube.
8. Centrifuge at 6000*g* (8000 rpm) for 1 minute. Discard flow-through and collection tube.

9. Place the DNeasy mini spin column in a new 2ml collection tube. Add 500µl Buffer AW1 and centrifuge for 1 minute at 6,000*g* (8,000 rpm). Discard flow-through and collection tube.

10. Place the DNeasy mini spin column in a new 2ml collection tube. Add 500µl Buffer AW2 and centrifuge for 3 minute at 20,000*g* (14,000 rpm) to dry the DNeasy membrane. Discard flow-through and collection tube.

11. Place the DNeasy mini spin column in a clean 1.5ml or 2ml microcentrifuge tube. Pipette 200µl Buffer AE directly onto the DNeasy membrane.

12. Incubate at room temperature for 2 minutes before centrifuging for 1 minute at 6,000*g* (8,000 rpm) to elute.

13. If downstream processing is not happening straight away, store samples at -30°C.

A9.3 For extraction using Qiagen BioRobot Universal

1. Place approximately 0.05g of each sample into the appropriate corresponding well of the BioRobot S-Block, noting the appropriate sample number for each well on the sample sheet.

2. Add 300µl Buffer ATL and 20µl Proteinase K to each well of the plate – pipette up and down to mix.

3. Seal the plate using a plastic plate seal and incubate at 56°C shaking at 100 rpm for 5 hours or overnight. Note: if taking into a quarantine lab for incubation, ensure you double bag the samples. When the incubation has finished, remove one layer of protection before leaving the lab and dispose as quarantine waste.

4. Prepare the BioRobot®:

    a. Switch on the BioRobot using the ‘on/off’ switch on the front right of the machine.

    b. Switch on the associated computer and log on.

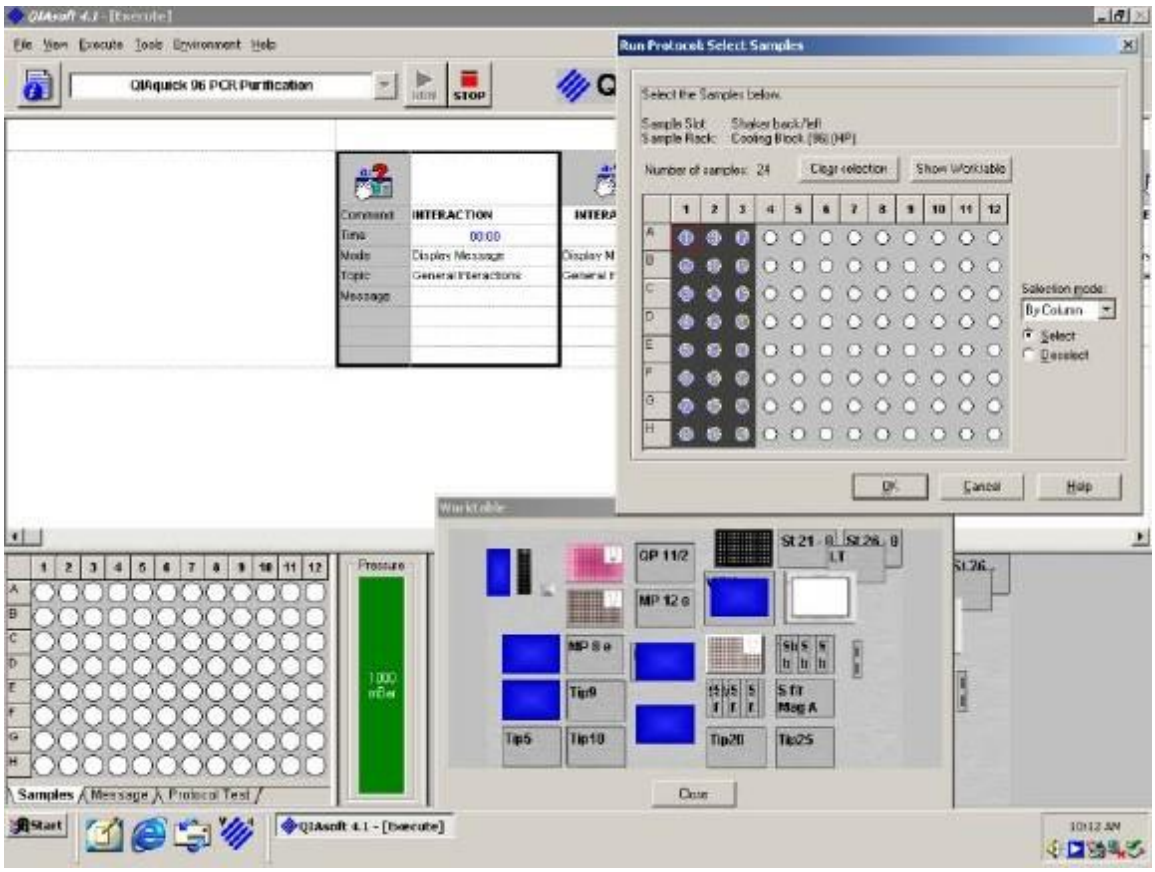
    c. Launch the QIAsoft 5 operating system.

    d. Enter the username ‘general operator’ and leave the password field blank. Press OK.

    e. Within the software, go to the dropdown menu (in red box in screenshot below) and select QIAamp Investigator BioRobot Kit > QIAamp DNA Casework (manual lysis) UNIV.

    f. Click ‘run’ (highlighted in green above) to start the setup process.

    g. The screen below shows and guides you through the setup of the BioRobot.



h. Follow the step by step instructions, pressing OK or Next once a step has been completed. NOTE: ensure that all reagents being placed on the machine do not contain any precipitate – if they do heat at 56°C for 5 minutes or until they have dissolved.

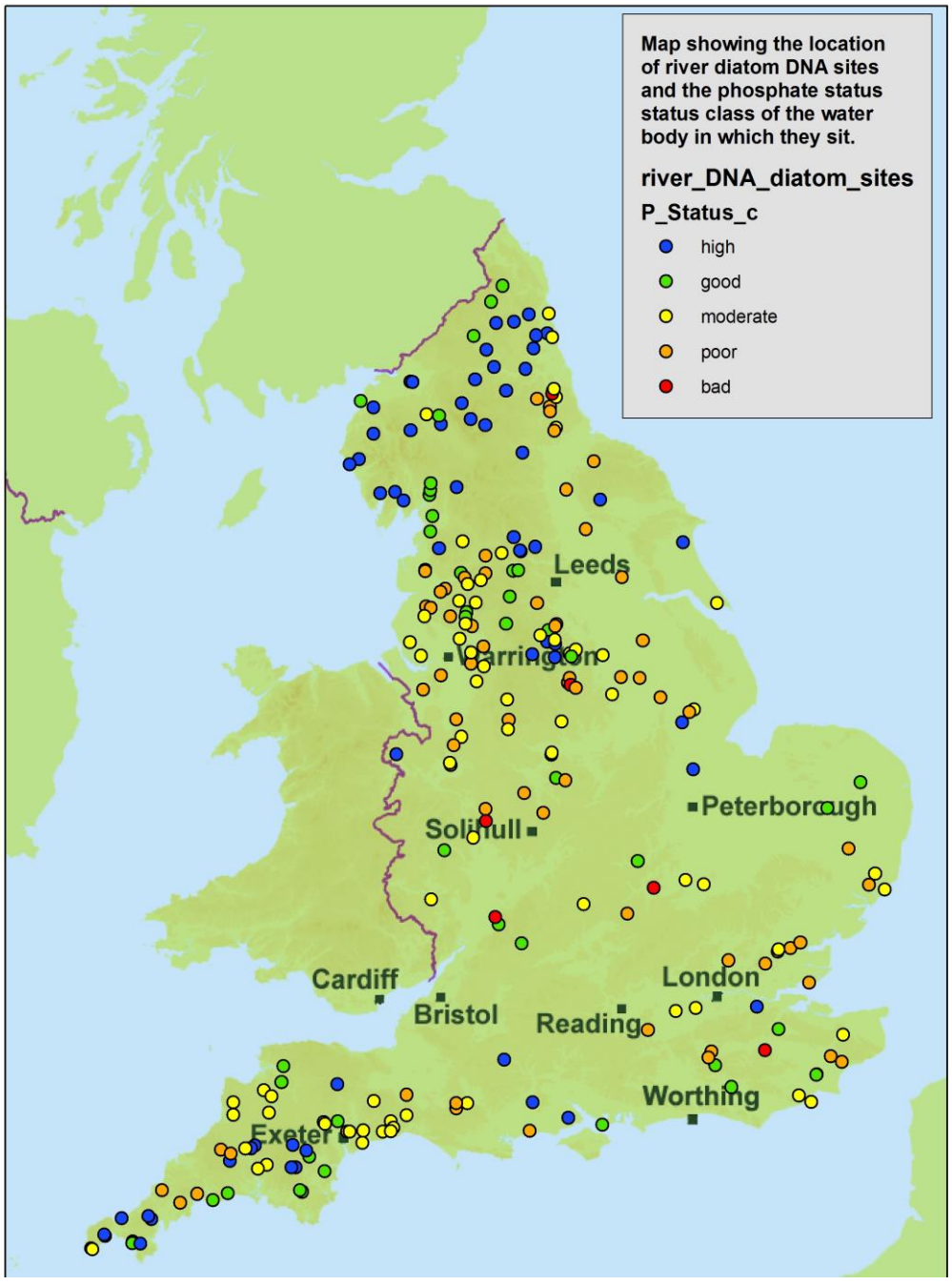
i. The final instruction before the program initialises will instruct you to place the S-Block containing your lysed samples onto the BioRobot. DO NOT press Next from this unless you are ready to proceed with the extraction! When ready press Next.

5. The BioRobot will now process your samples. This will take about 2.5 hours for a full 96-well plate.

6. At the end of the run, remove your samples in their 96-well plate.

7. If downstream processing is not happening straight away, store the samples at -30°C.

Appendix 10: Distribution of sites used to collect diatom samples for the calibration dataset



See Water Framework Directive UK TAG website for information on phosphorus standards (www.wfduk.org/resources/new-and-revised-phosphorus-and-biological-standards).

Would you like to find out more about us or about your environment?

Then call us on
03708 506 506 (Monday to Friday, 8am to 6pm)

email
enquiries@environment-agency.gov.uk

or visit our website
www.gov.uk/environment-agency

incident hotline 0800 807060 (24 hours)
floodline 0345 988 1188 / 0845 988 1188 (24 hours)

Find out about call charges: www.gov.uk/call-charges

Environment first: Are you viewing this on screen? Please consider the environment and only print if absolutely necessary. If you are reading a paper copy, please don't forget to reuse and recycle if possible.