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.

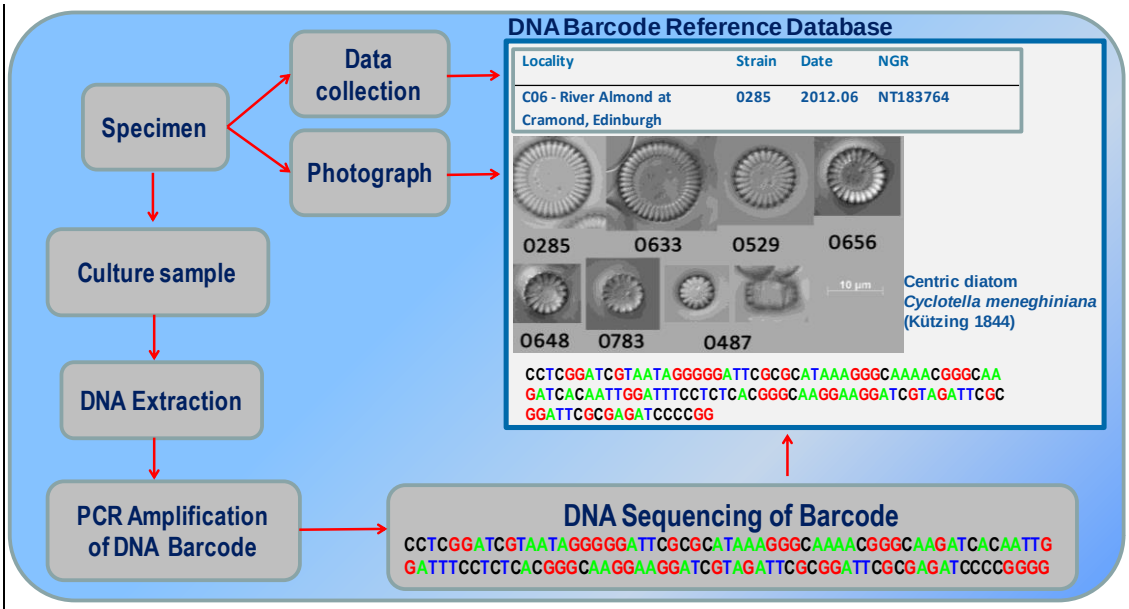


Figure 1.1 Steps involved in creating a DNA barcode reference database

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.

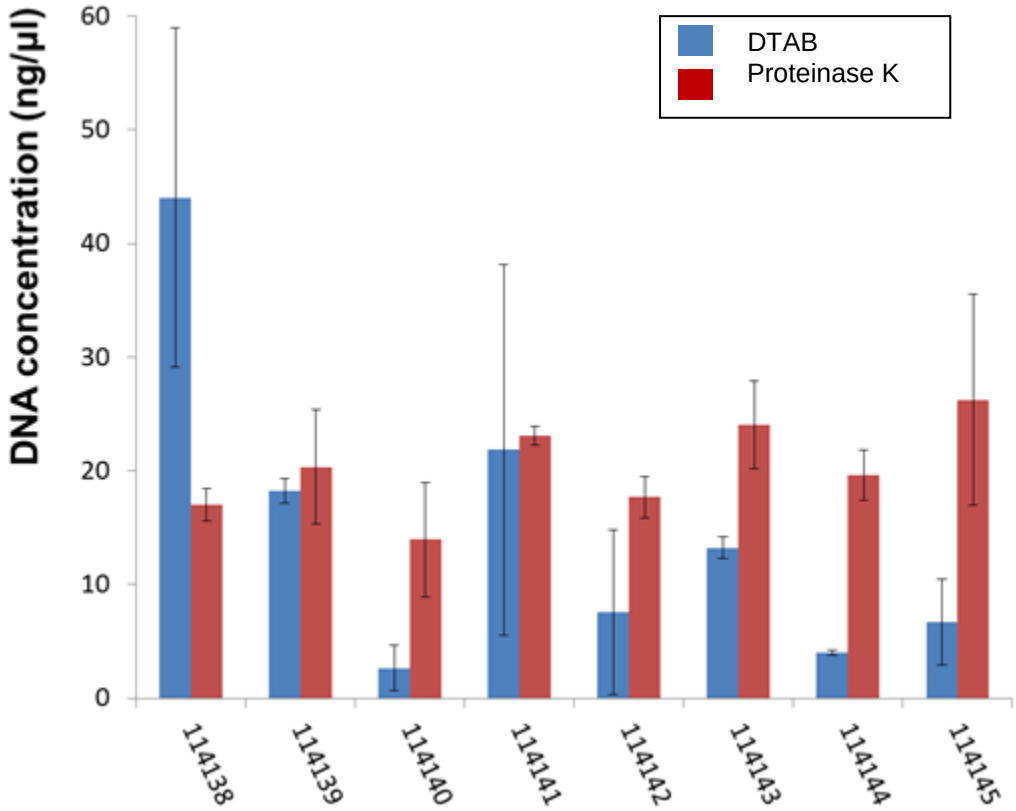


Figure 4.1 DNA concentrations from the extracts of 8 diatom samples

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.

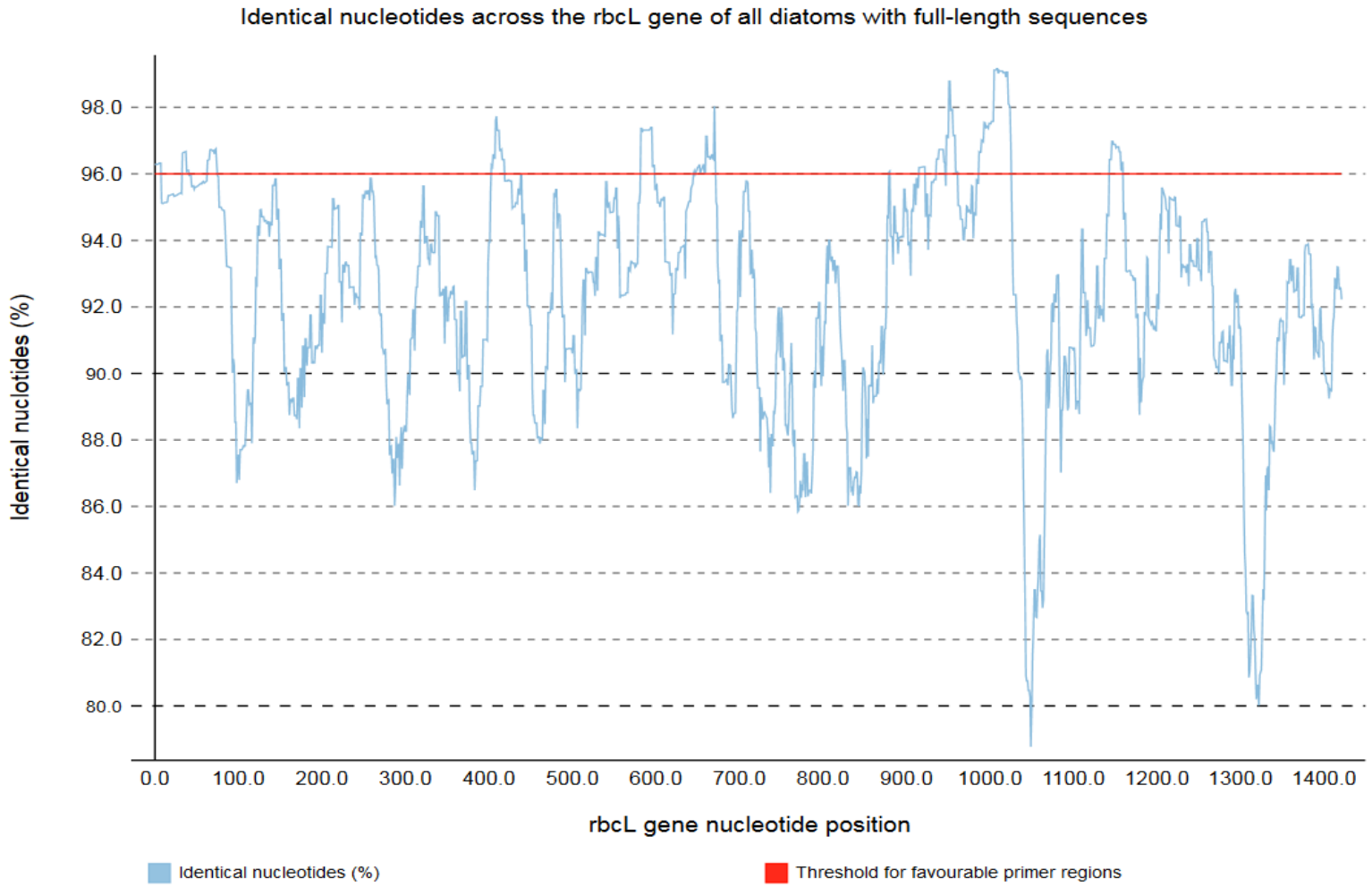


Figure 4.2 Percentage of identical nucleotides plotted along the length of an alignment of full length diatom rbcL sequences

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).

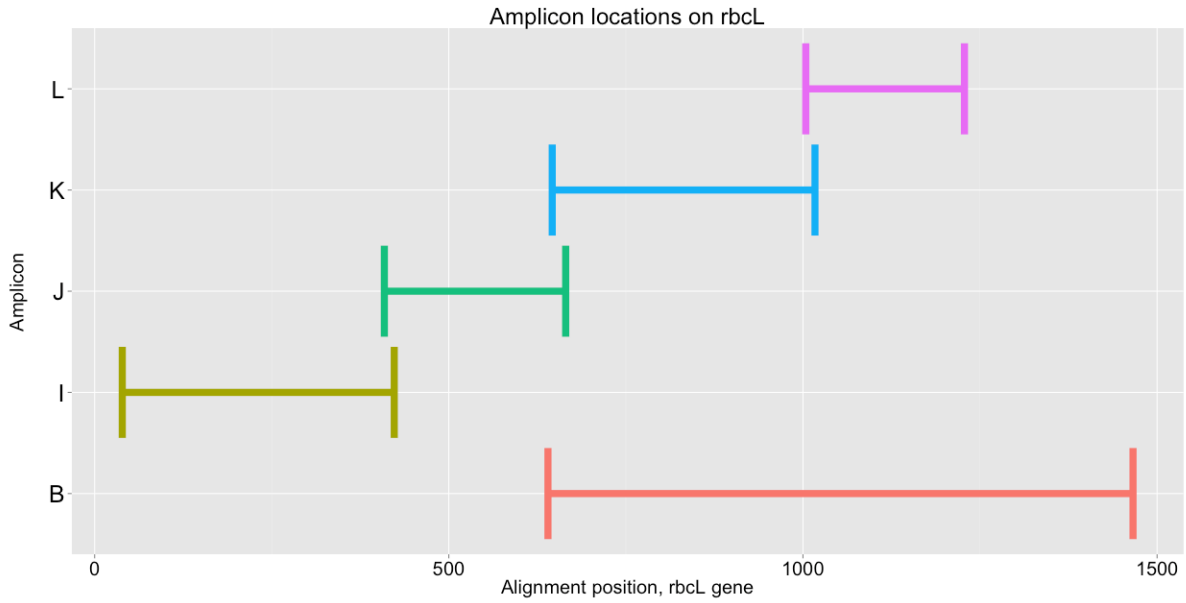


Figure 4.3 Locations of each hypothetical amplicon region (fragment) along the length of the rbcL gene

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.

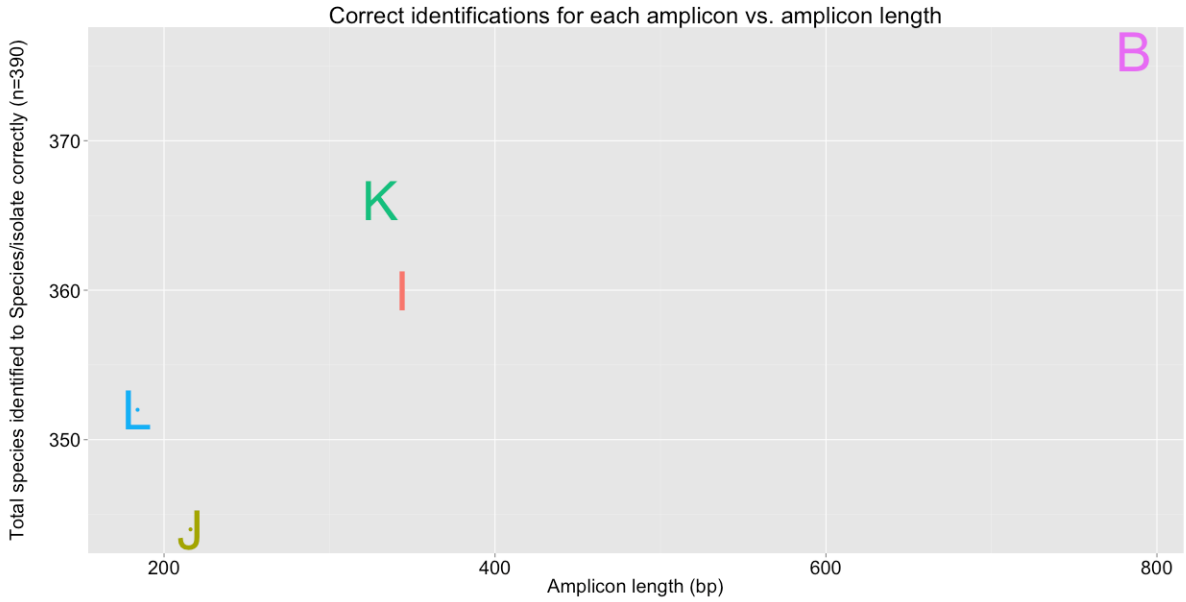


Figure 4.4 Correct species level taxonomic assignments plotted against the length of the amplicon fragment

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.

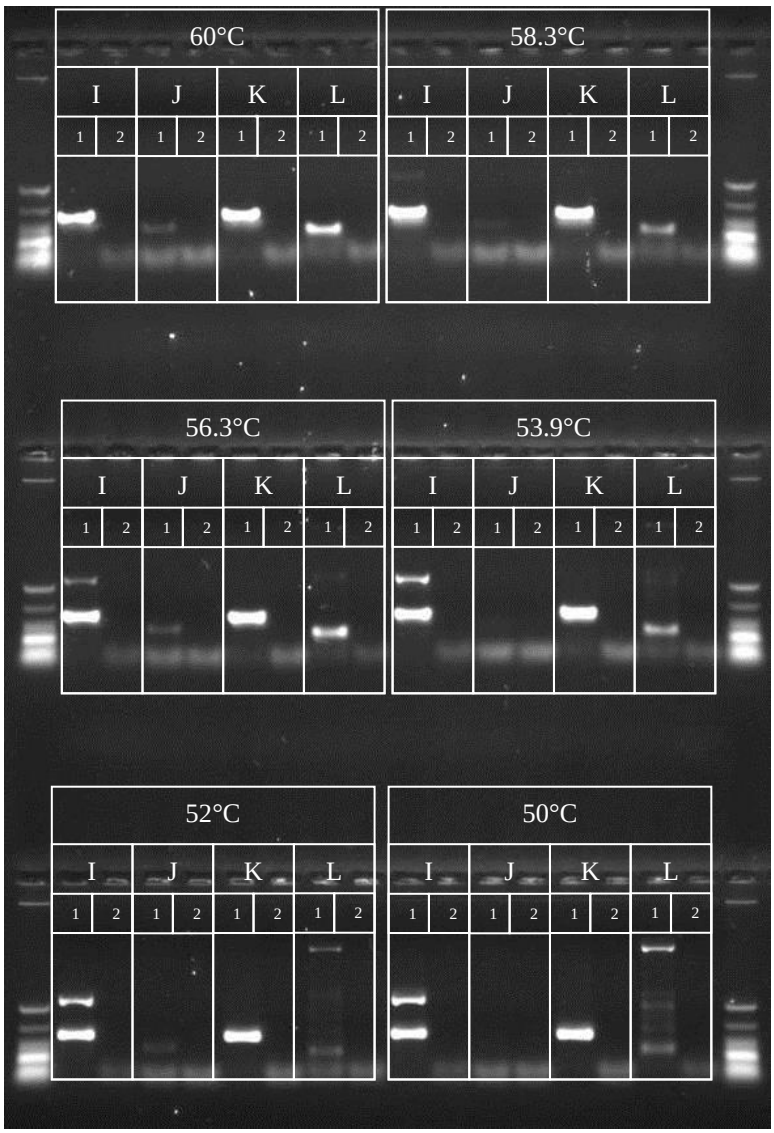


Figure 4.5 Gel electrophoresis of PCR products post amplification performed at different annealing temperatures (between 50 and 60°C) using newly designed primer sets (I, J, K and L) tested on DNA from a diatom sample and a no template control

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.

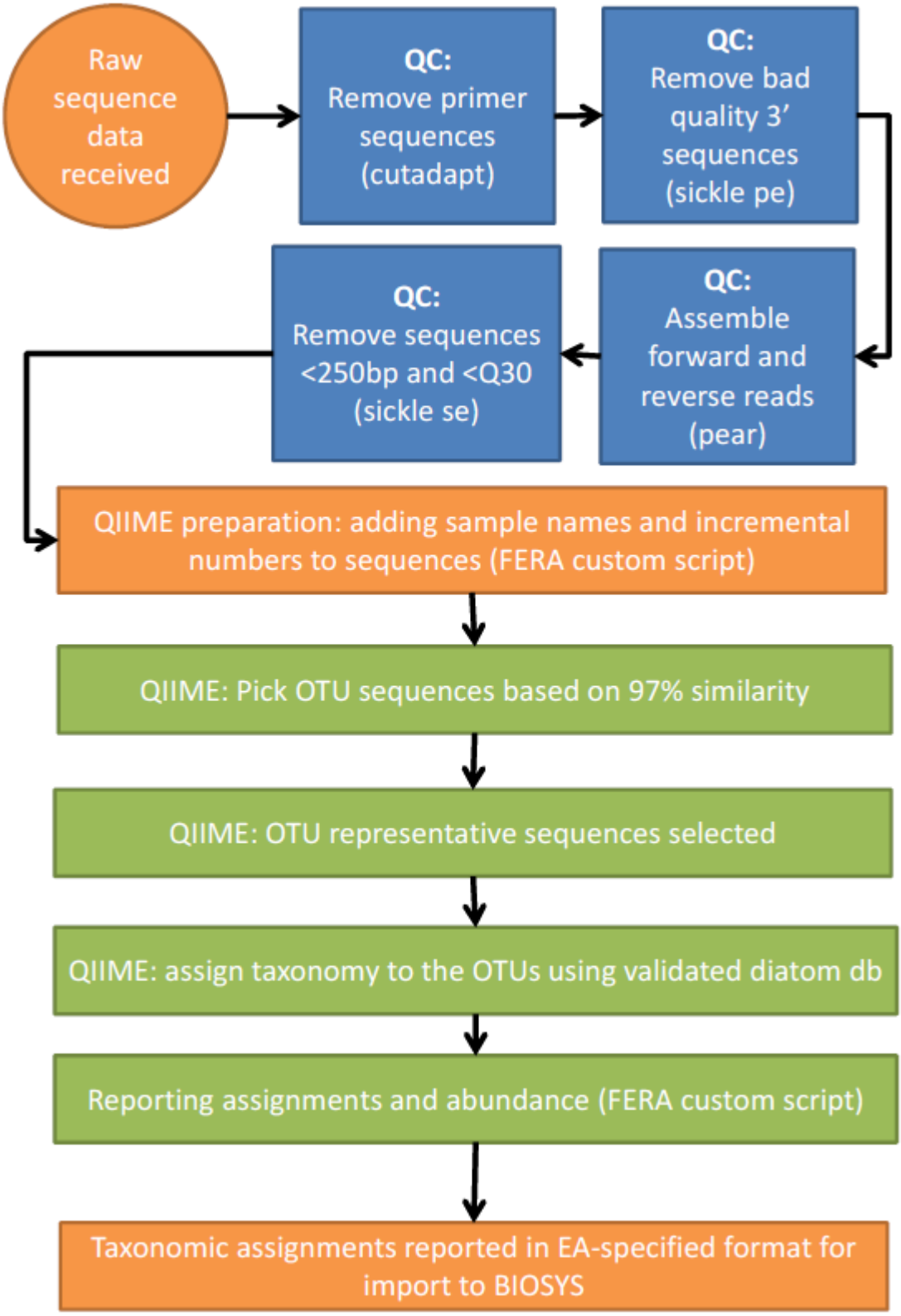


Figure 5.1 Quality control and QIIME pipeline for analysis of diatom NGS data

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.

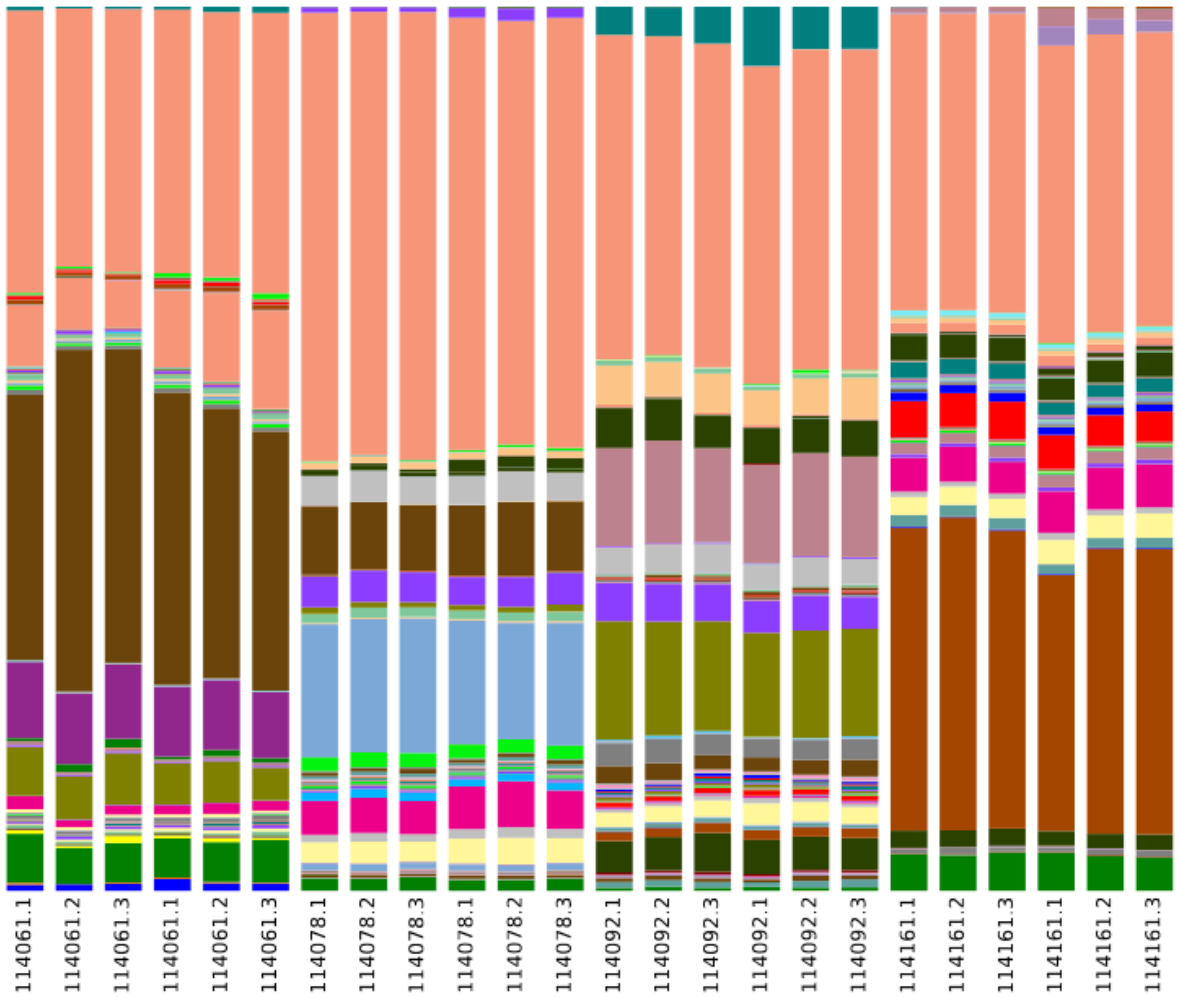


Figure 5.2 Stacked bar chart showing each of the 4 samples with 6 PCR replicates

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)

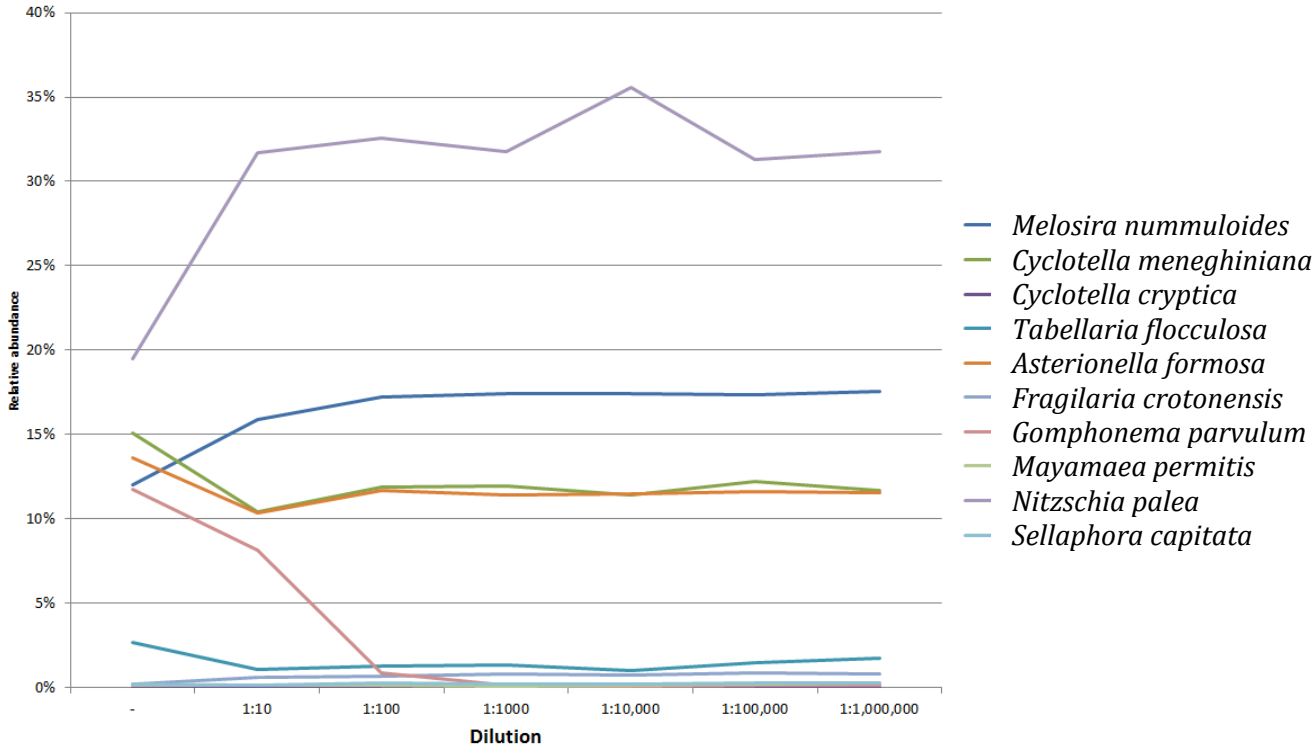


Figure 5.3 Relative abundance of each species in the mock community

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.

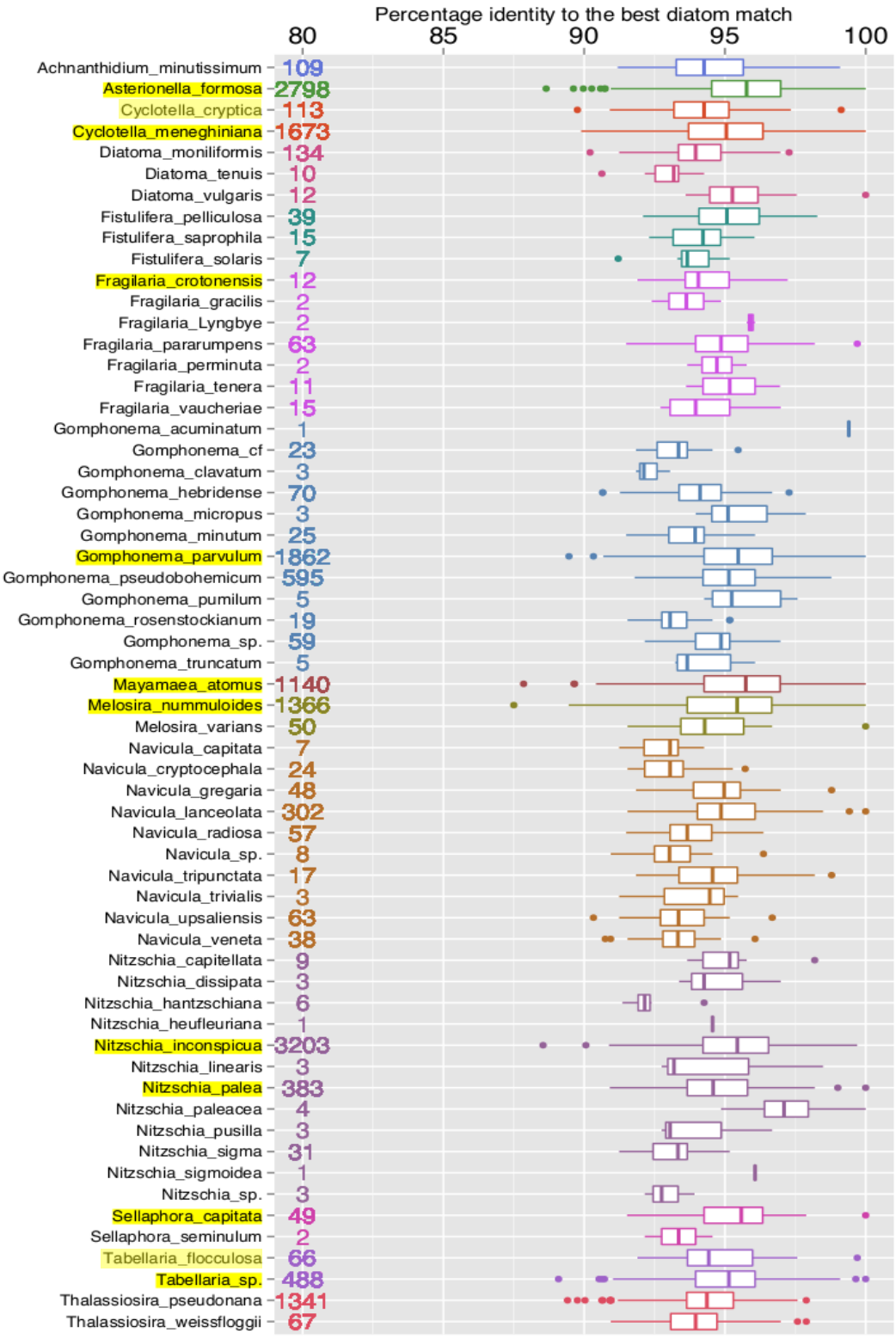


Figure 5.4 Box and whisker plots for all species detected in the mock community sample, showing the number of OTUs assigned to the species (right of the name) and boxplots showing the percentage similarities of all the representative sequences to the best match in the database that resulted in assignment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

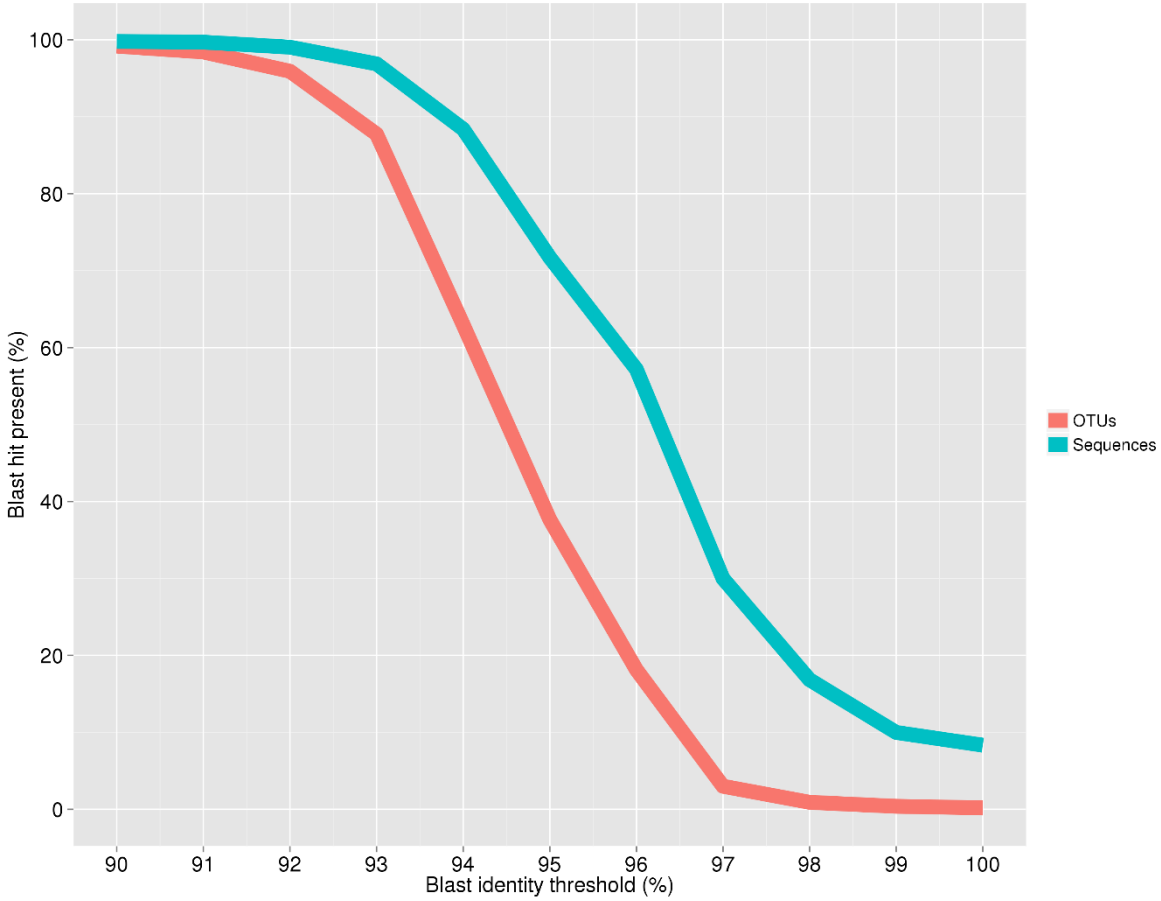


Figure 5.5 Number of OTUs (red) and overall proportion of sequences in samples (blue) having a hit in the diatom database within increasing BLAST identity threshold

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.

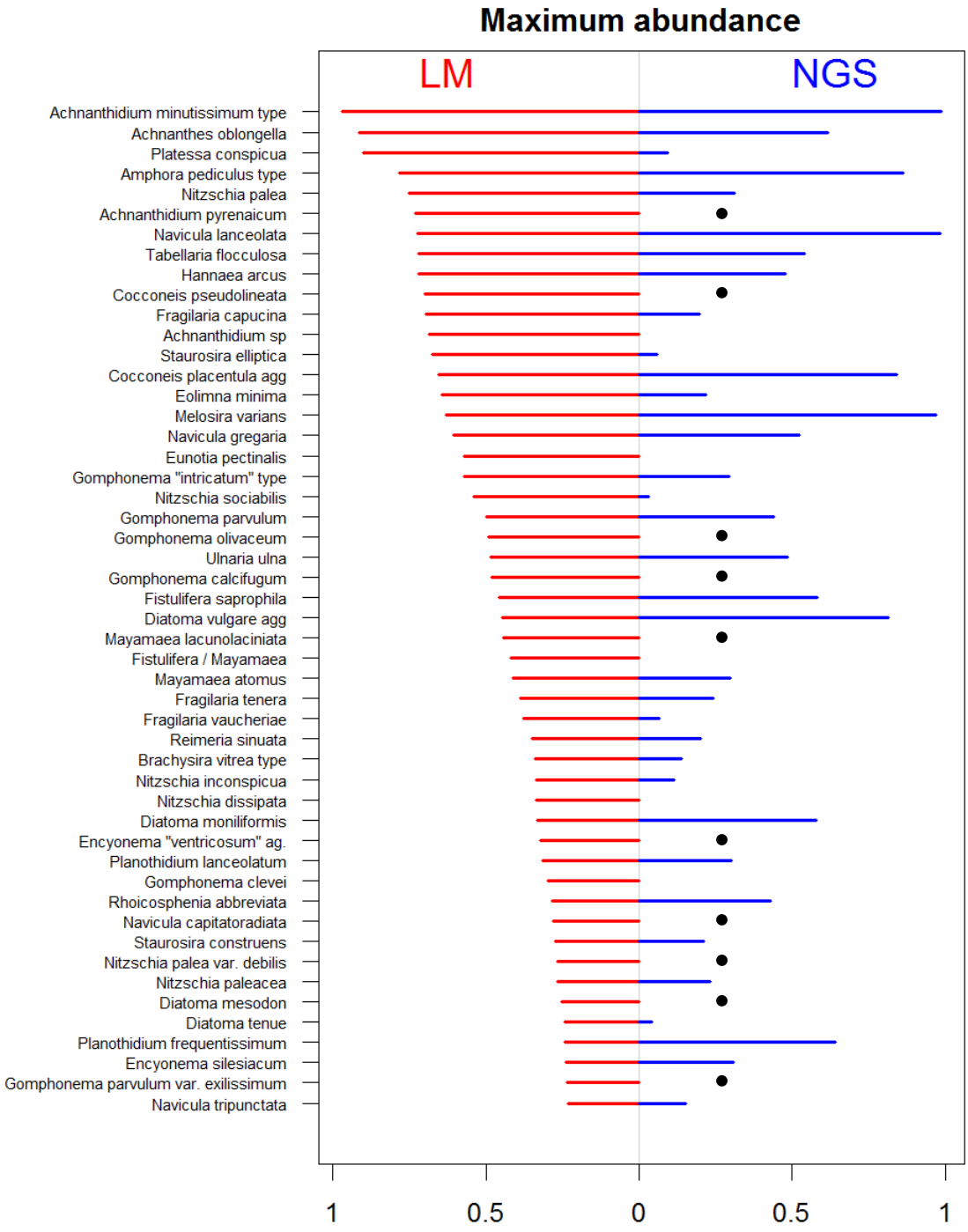


Figure 6.1 Differences in maximum abundance of the 50 most common diatom taxa in 628 samples as recorded by LM to show comparison with NGS data

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.

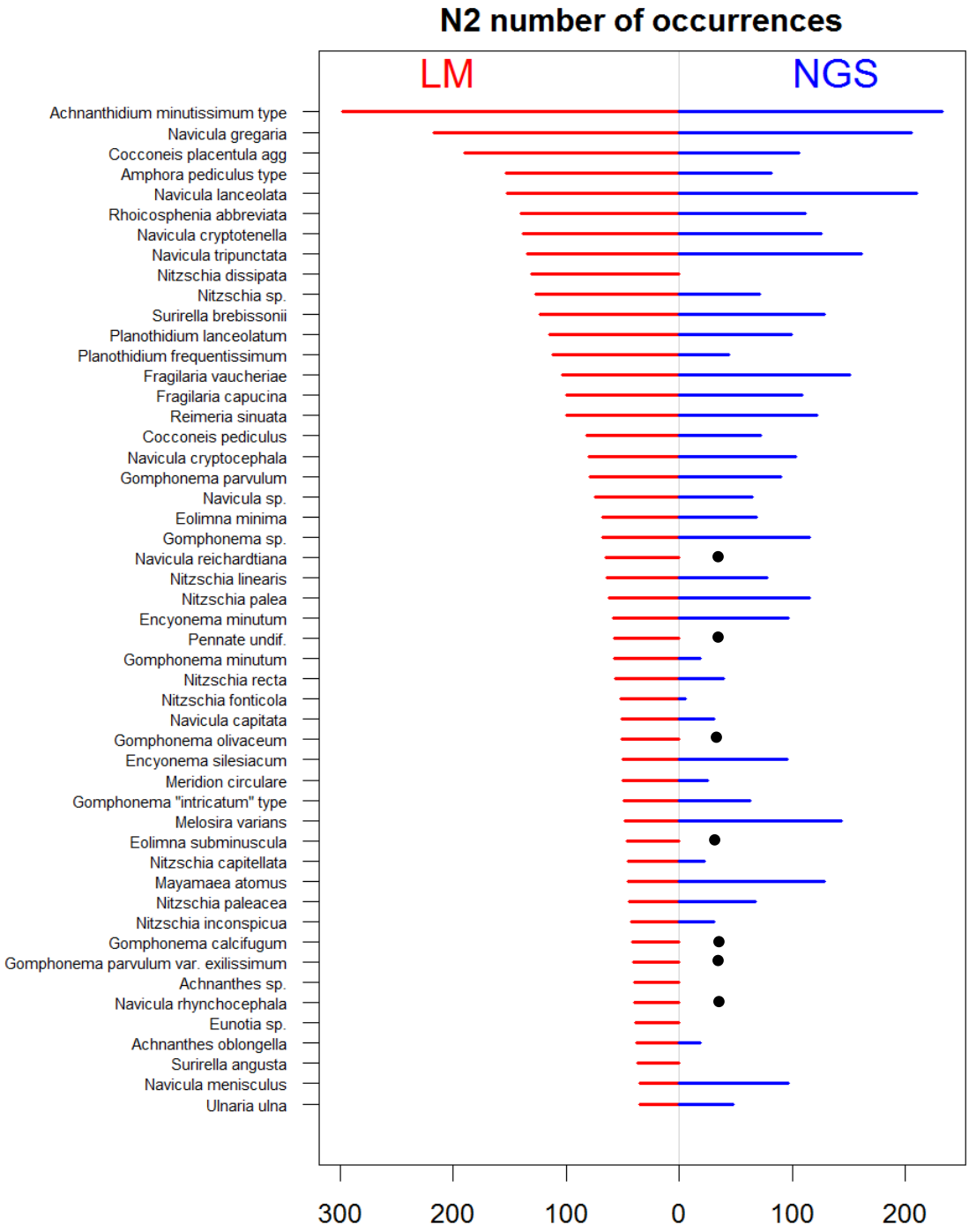


Figure 6.2 Differences in the total number of times that a taxon was recorded for the 50 most frequently occurring diatom taxa in samples as recorded by LM compared to NGS in the 628 sample dataset.

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.

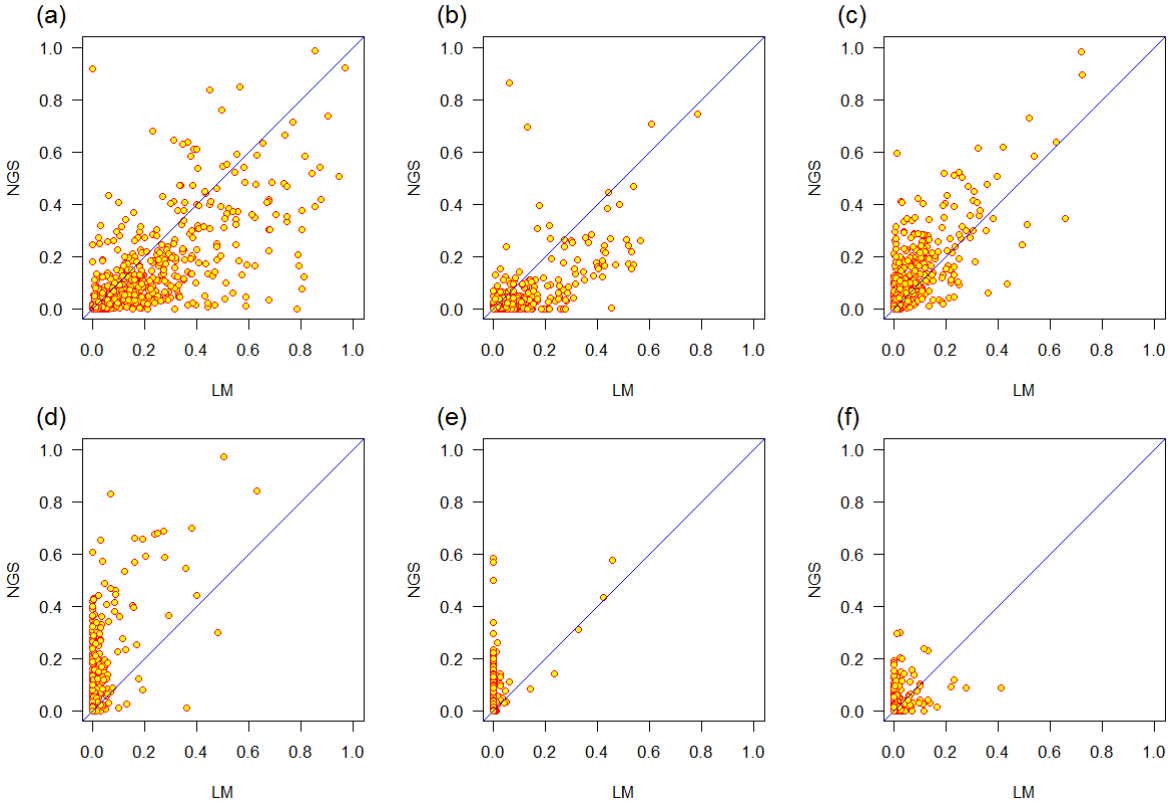


Figure 6.3 Differences between representation of common taxa in LM and NGS analyses of selected diatom species: (a) *Achnanthidium minutissimum* type (small, 1 chloroplast); (b) *Amphora pediculus* (small, 1 chloroplast); (c) *Navicula lanceolata* (medium sized, 2 chloroplasts); (d) *Melosira varians* (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'.

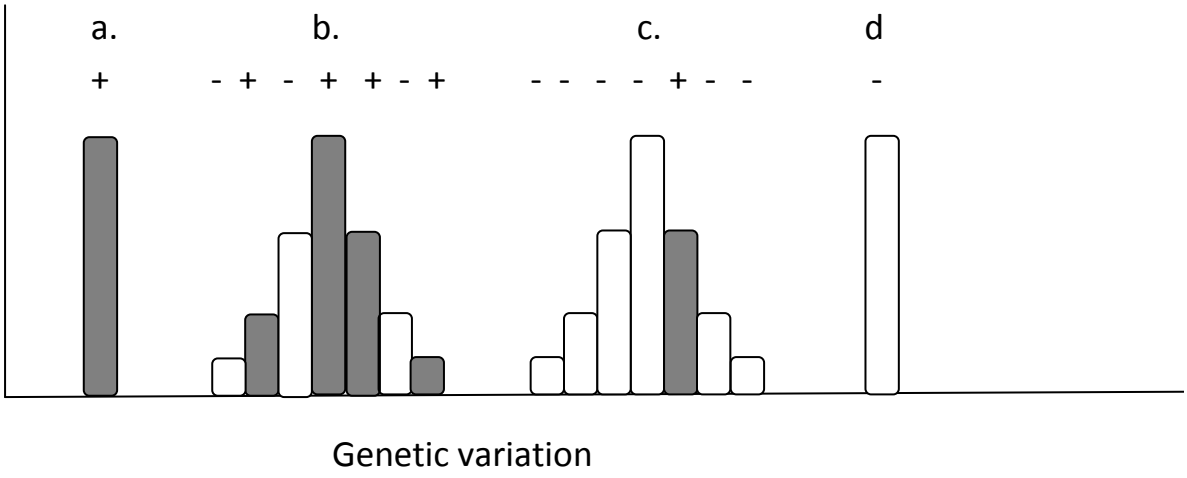


Figure 6.4 Conceptual diagram of relationship between LM and NGS outputs for 4 different scenarios: (a) clearly defined taxon aligns with barcode; (b) species complex with several different barcodes represented in the barcode database; (c) species complex poorly represented in the barcode database; and (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)

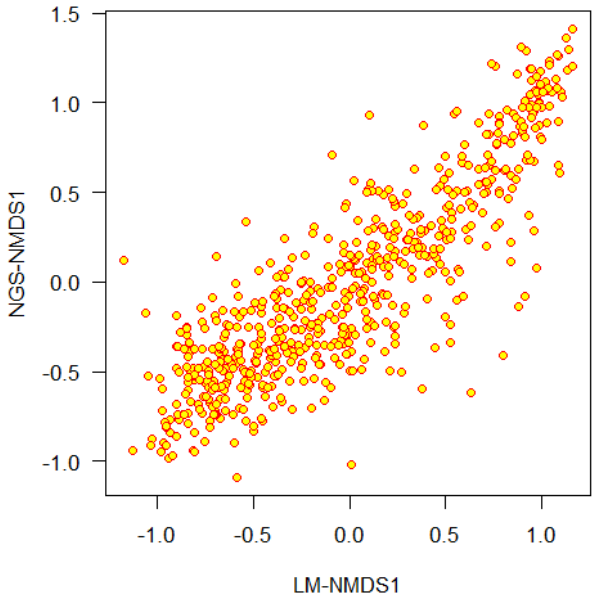


Figure 6.5 Comparison of the first axes of NMDS ordinations performed using LM and NGS data (r = 0.87)

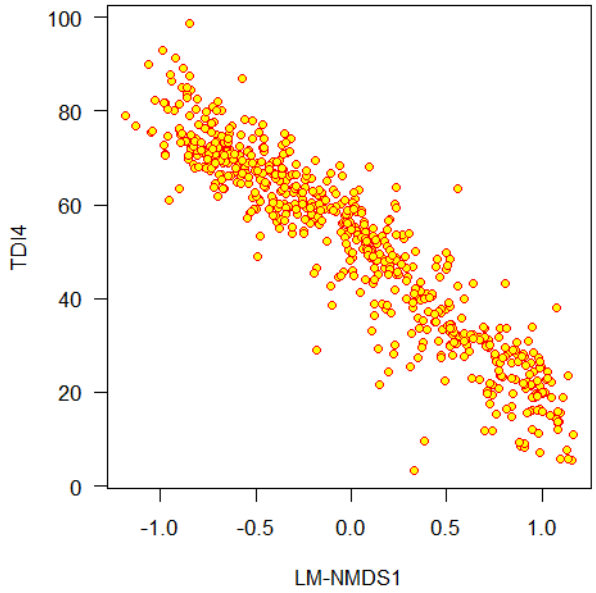


Figure 6.6 Axis 1 of NMDS of LM data versus TDI4 (r = -0.94)

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.

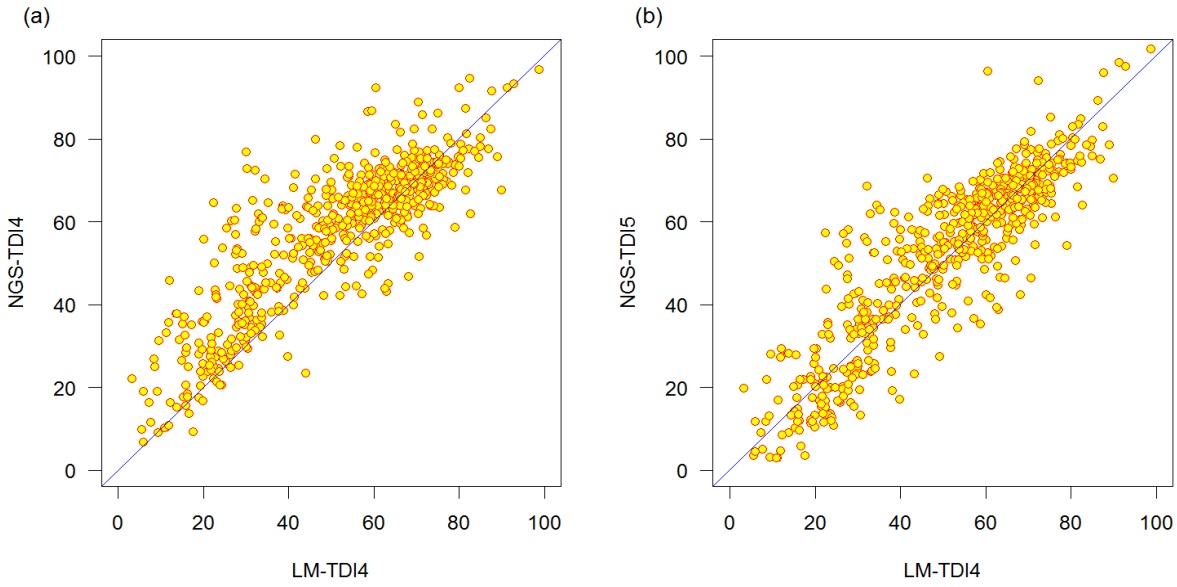


Figure 6.7 Comparison between the TDI calculated on LM and NGS data for 628 samples from UK rivers: (a) using TDI4 (LM) weights to calculate TDI for NGS data (Pearson's *r* = 0.86, Lin's *r* = 0.81; and (b) using NGS specific weights ('TDI5', Pearson's *r* = 0.90, Lin's *r* = 0.89; RMSE = 9.3)

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.

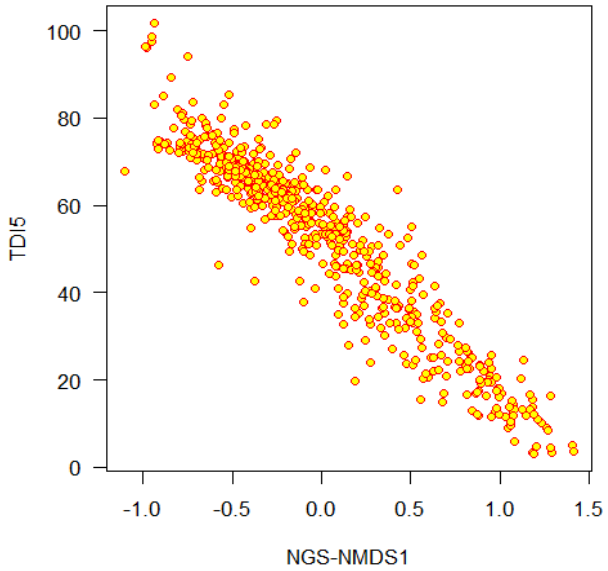


Figure 6.8 Axis 1 of NMDS of NGS data versus TDI5 (*r* = -0.95).

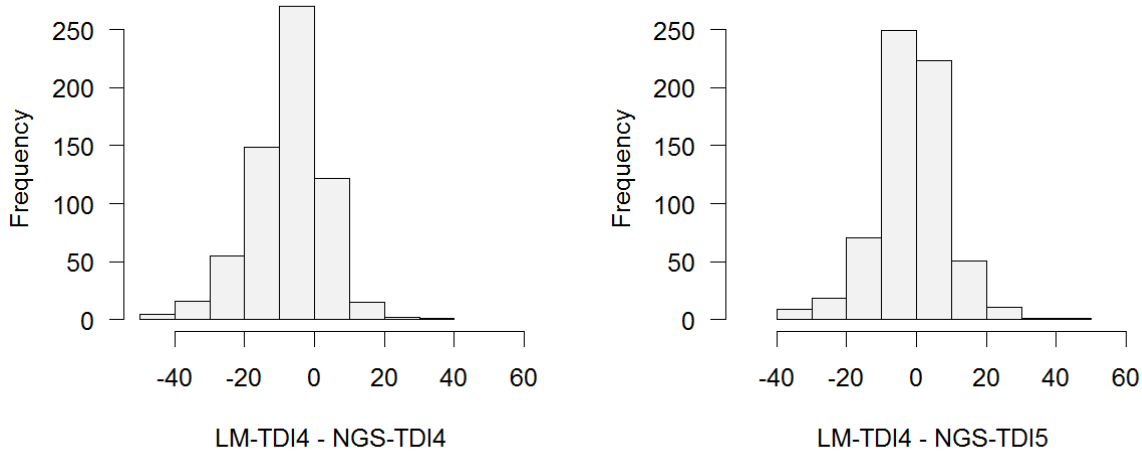


Figure 6.9 Histograms showing agreement between TDI calculated with LM and NGS data for 628 samples from UK rivers, calculated using NGS data and TDI4 weights (left) and calculated using NGS specific weights (right)

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.

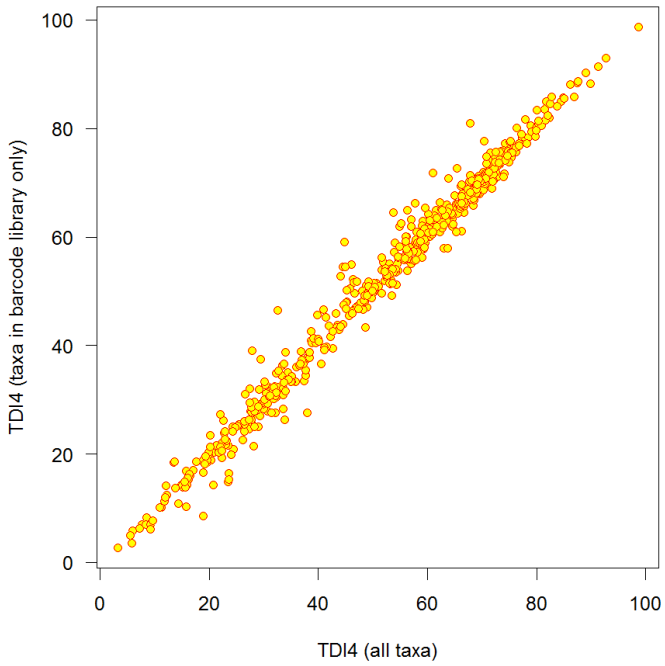


Figure 6.10 Difference between TDI4 based on LM data calculated with all taxa and with just those taxa represented in the barcode database

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.

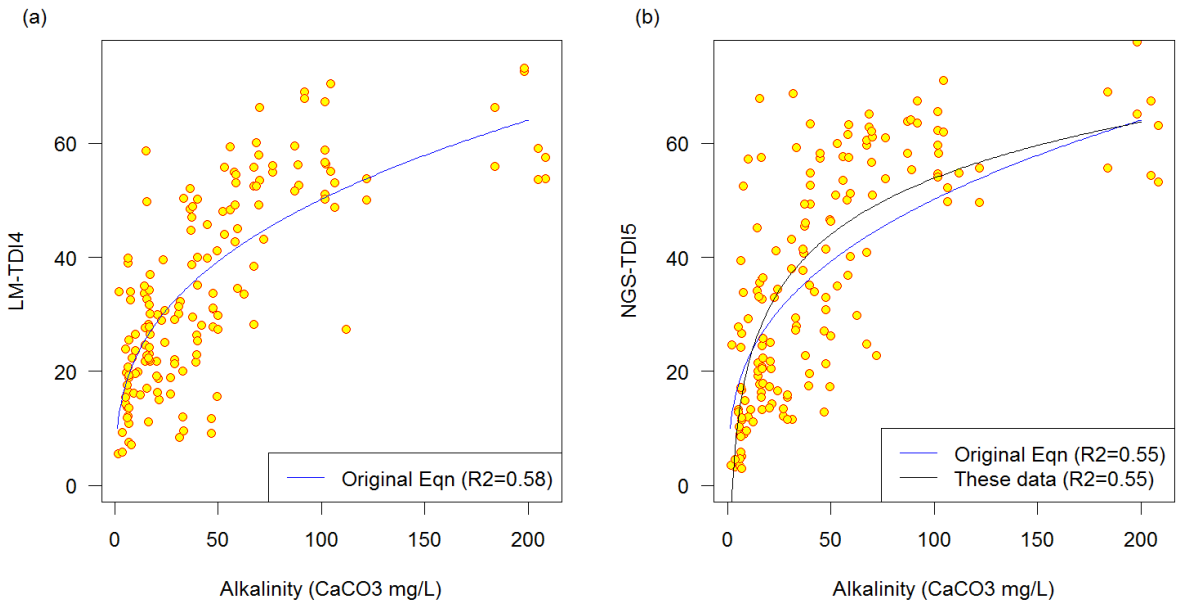


Figure 6.11 Relationship between alkalinity and TDI for 171 samples from reference sites throughout the UK: (a) based on LM results and TDI4 calculation (Equation 6.5); and (b) based on NGS results and TDI5 calculation (see Equation 6.6).

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

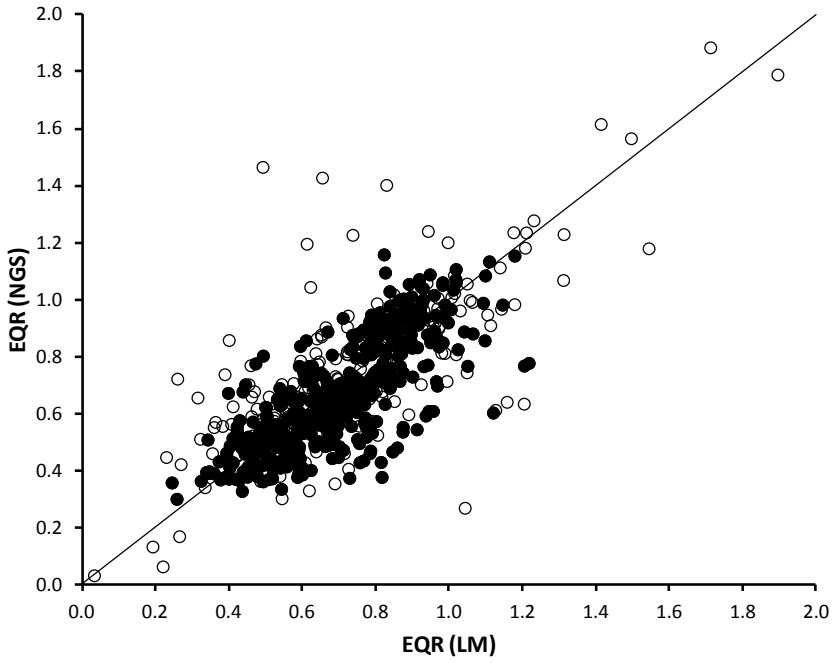


Figure 6.12 Comparison between EQR calculated on LM and NGS data for 620 samples from UK rivers for which alkalinity data were available

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.

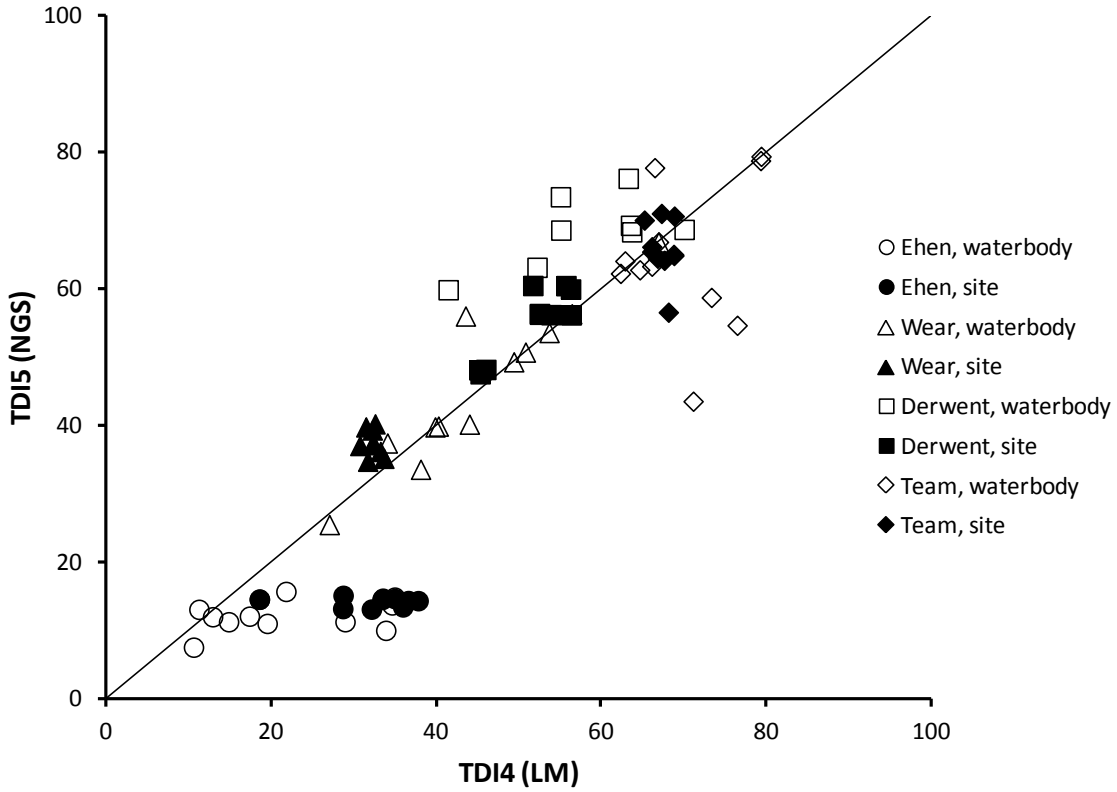


Figure 7.1 Within water body and within site variation in LM and NGS analyses of diatom samples from 4 contrasting river sites in England

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.

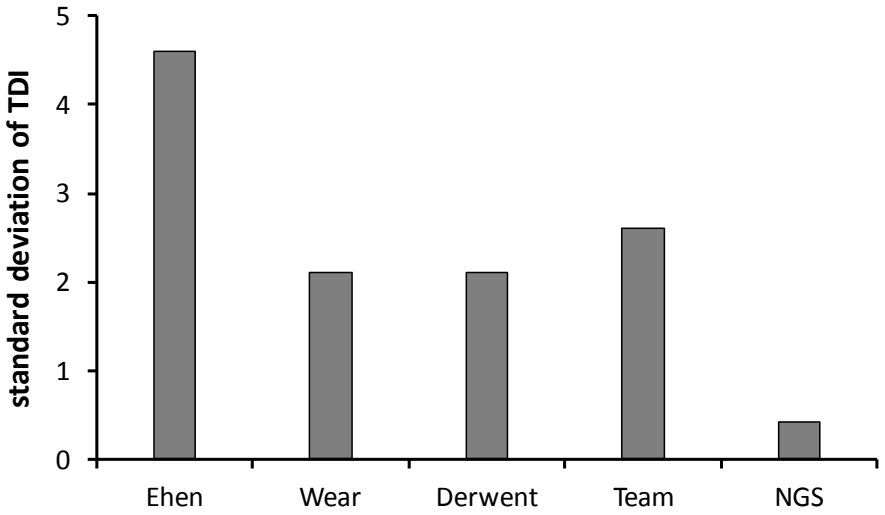


Figure 7.2 Variation (as standard deviation of TDI) between analytical results (LM) from experienced analysts for one sample from each water body reported in Table 7.2 alongside results from tests of analytical specificity for NGS

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.

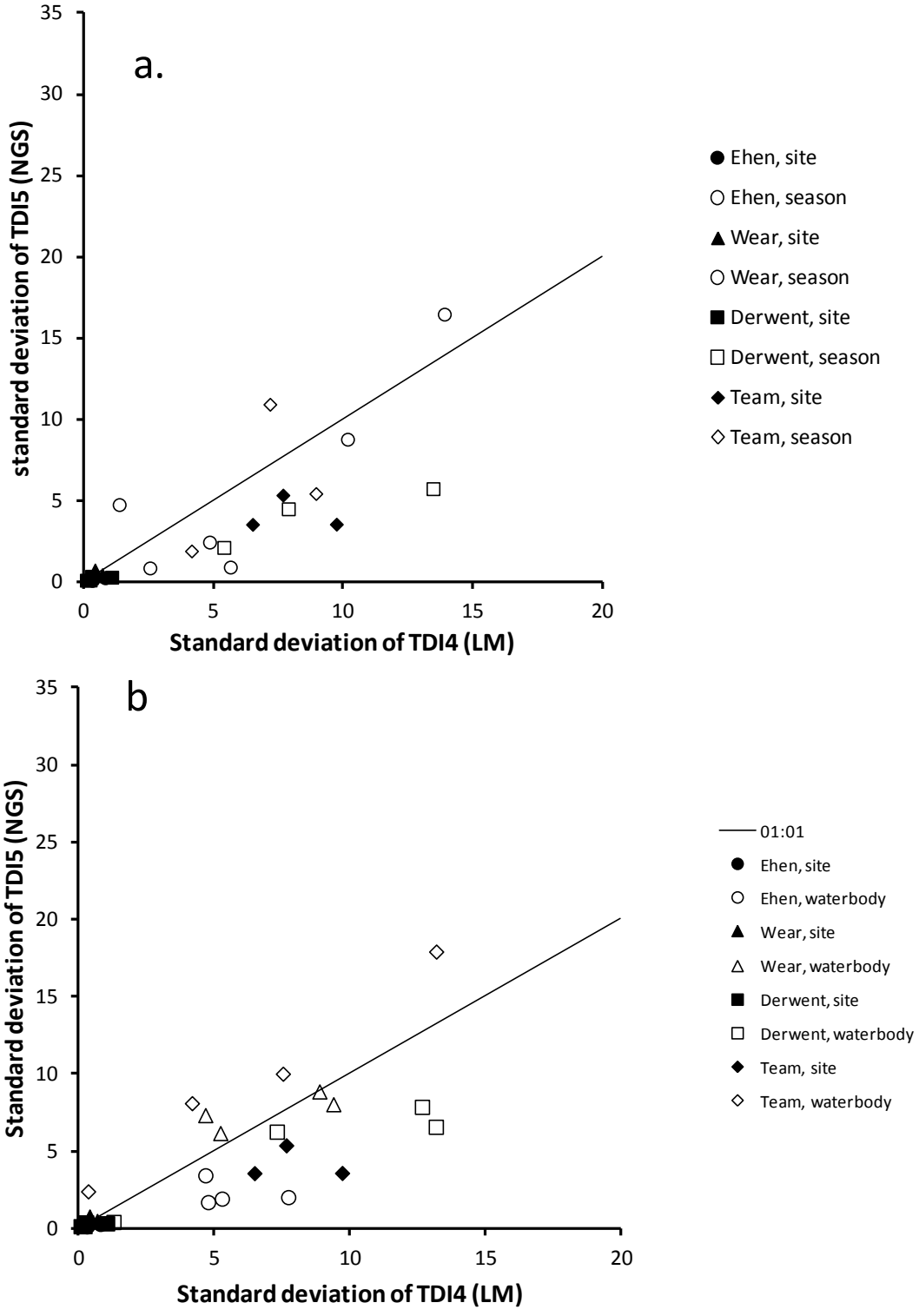


Figure 7.3 Within site and within waterbody variation in LM and NGS analyses of diatom samples from 4 contrasting river sites in England expressed as standard deviation: (a) water body variation expressed as spatial variation within the water body (*n* = 3) on 4 separate occasions; and (b) water body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

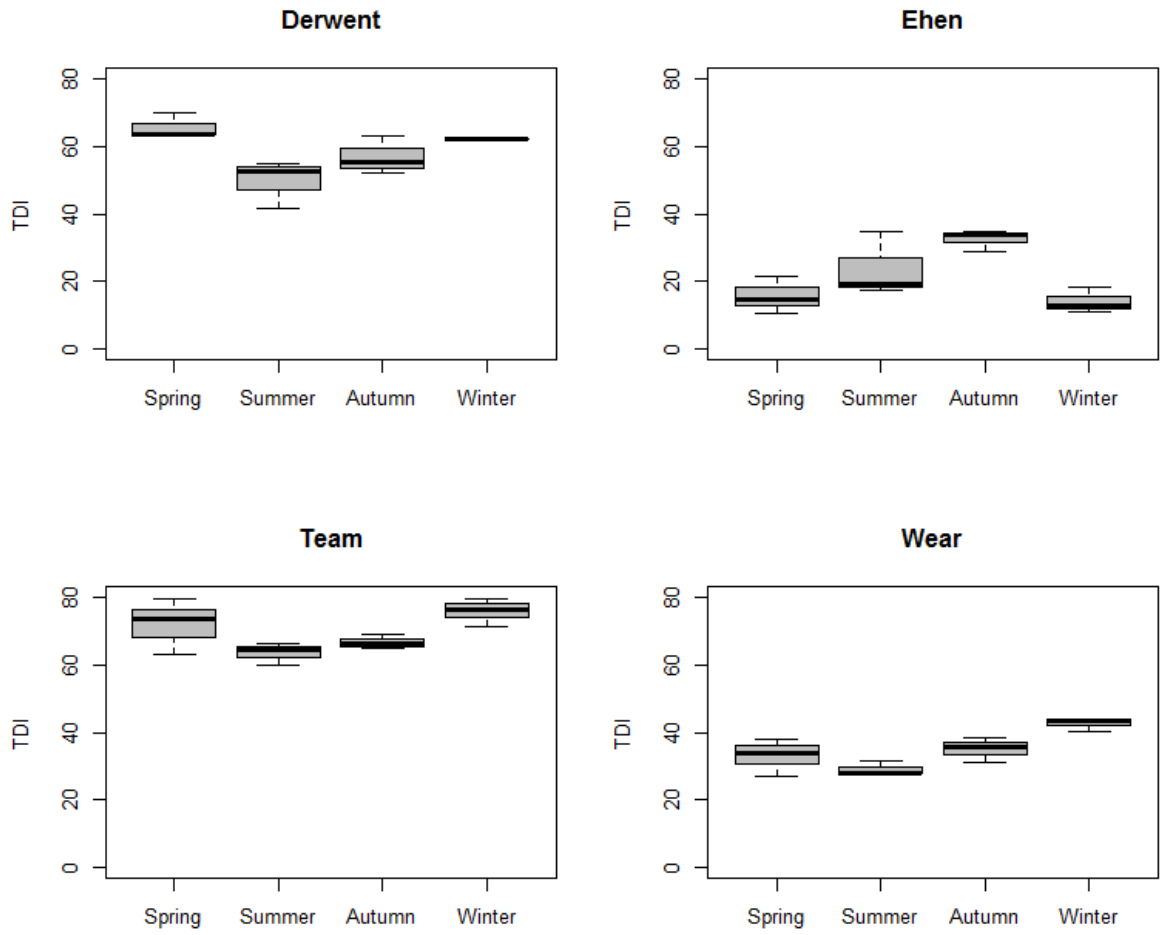


Figure 7.4 Seasonal variation in TDI4 (LM analyses) in the Rivers Ehen, Wear, Derwent and Team

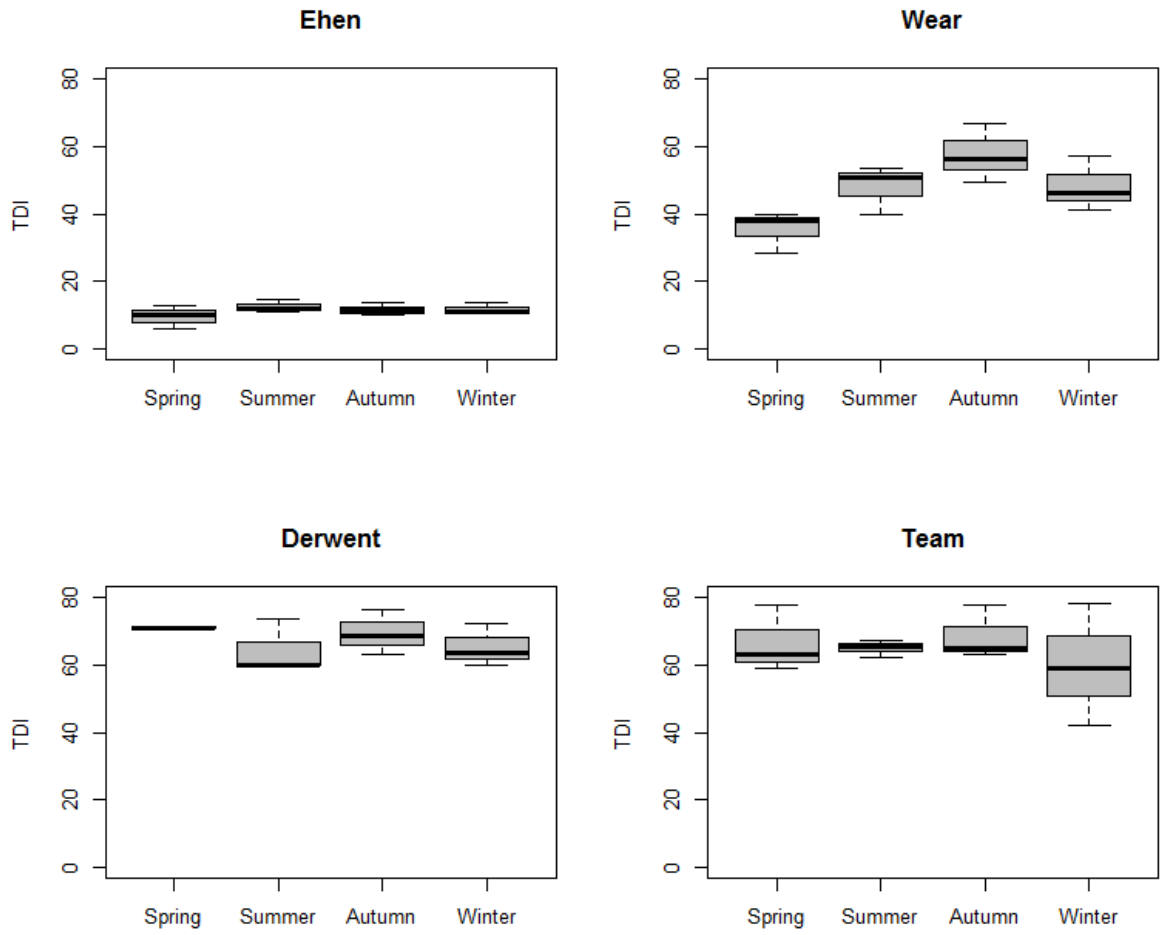


Figure 7.5 Seasonal variation in TDI5 (NGS analyses) in the Rivers Ehen, Wear, Derwent and Team

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.

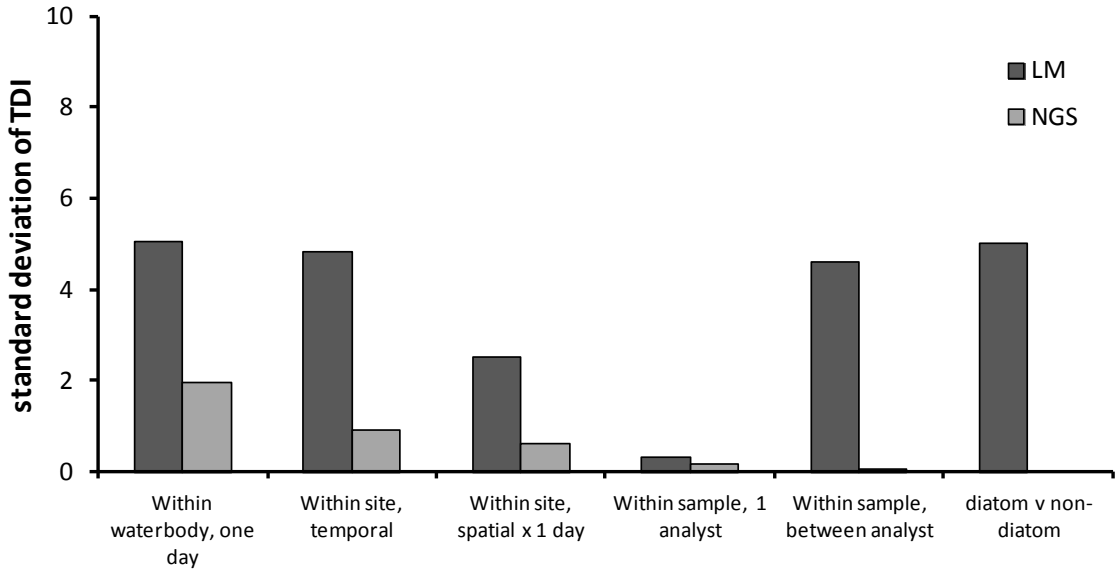


Figure 7.6 Sources of uncertainty in TDI calculated using LM and NGS data from samples collected from the River Ehen (high status)

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

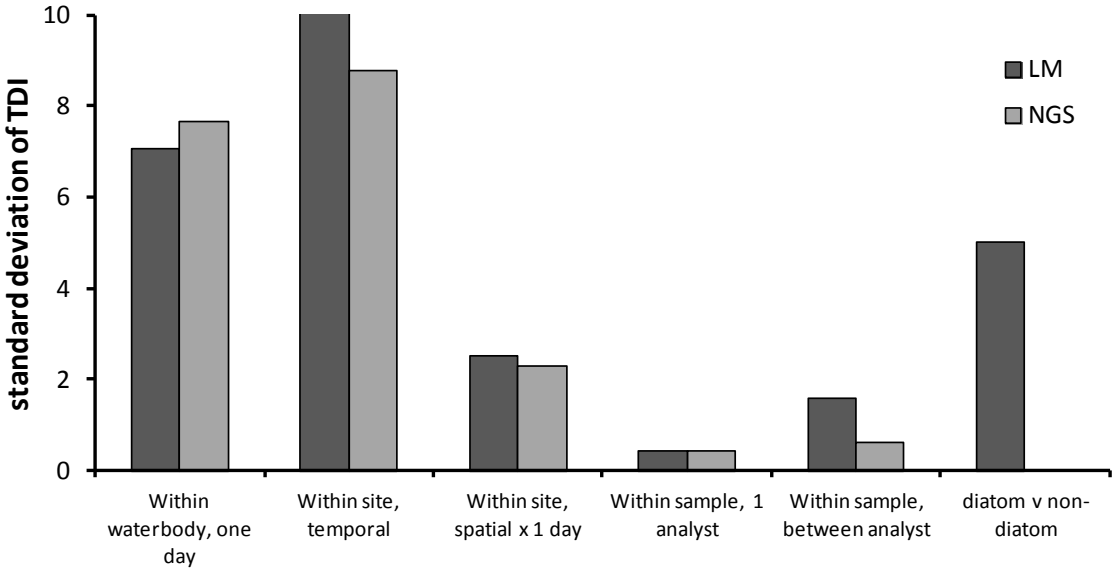


Figure 7.7 Sources of uncertainty in TDI calculated using LM and NGS data from samples collected from the River Wear (good status)

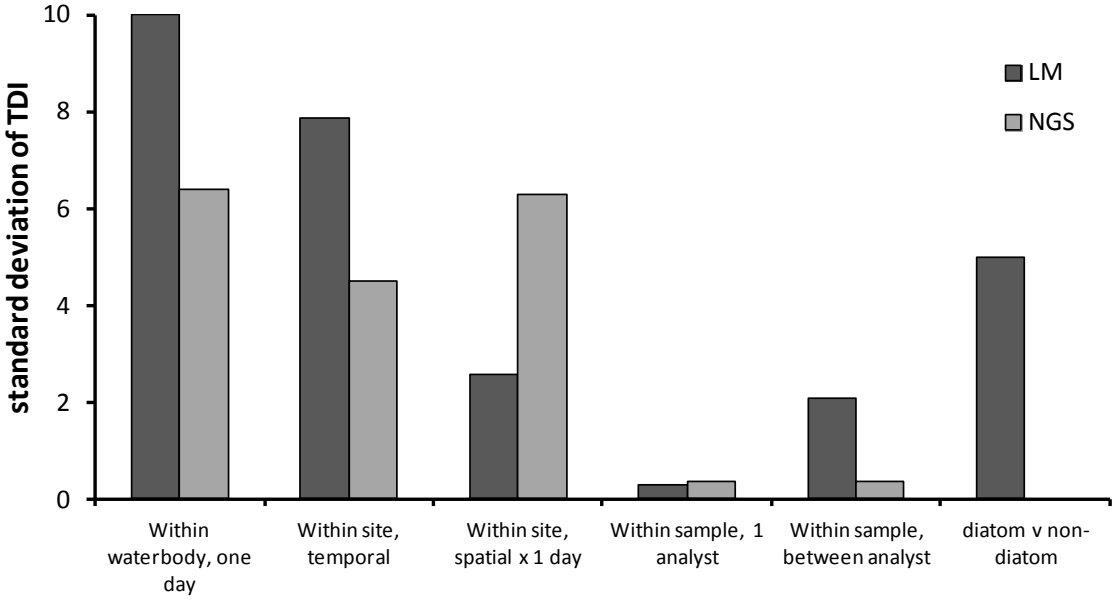


Figure 7.8 Sources of uncertainty in TDI calculated using LM and NGS data from samples collected from the River Derwent (moderate status)

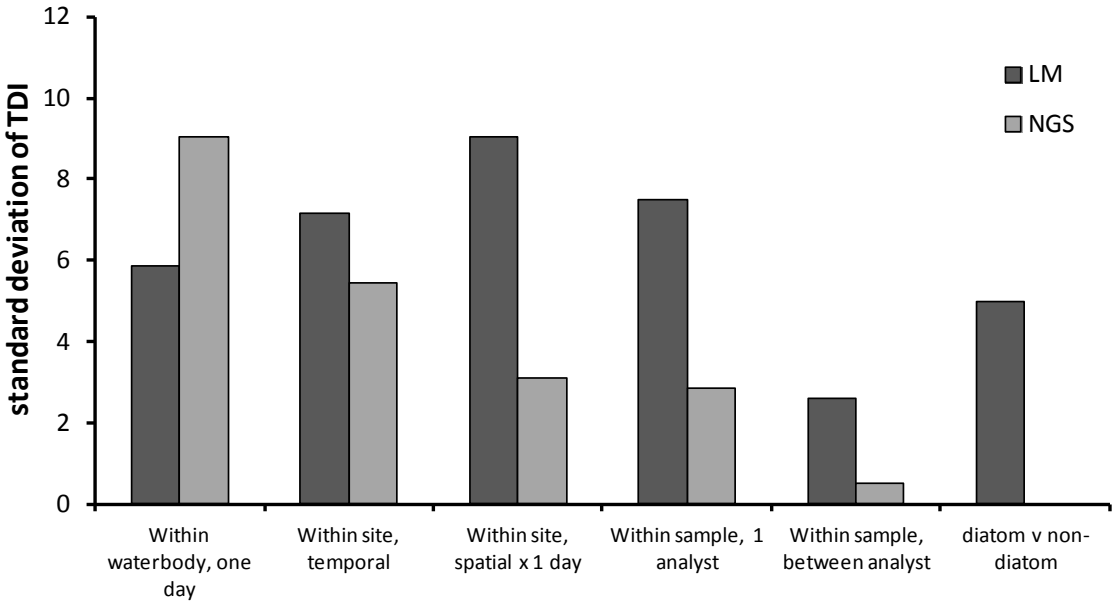


Figure 7.9 Sources of uncertainty in TDI calculated using LM and NGS data from samples collected from the River Team (poor/bad status)

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.

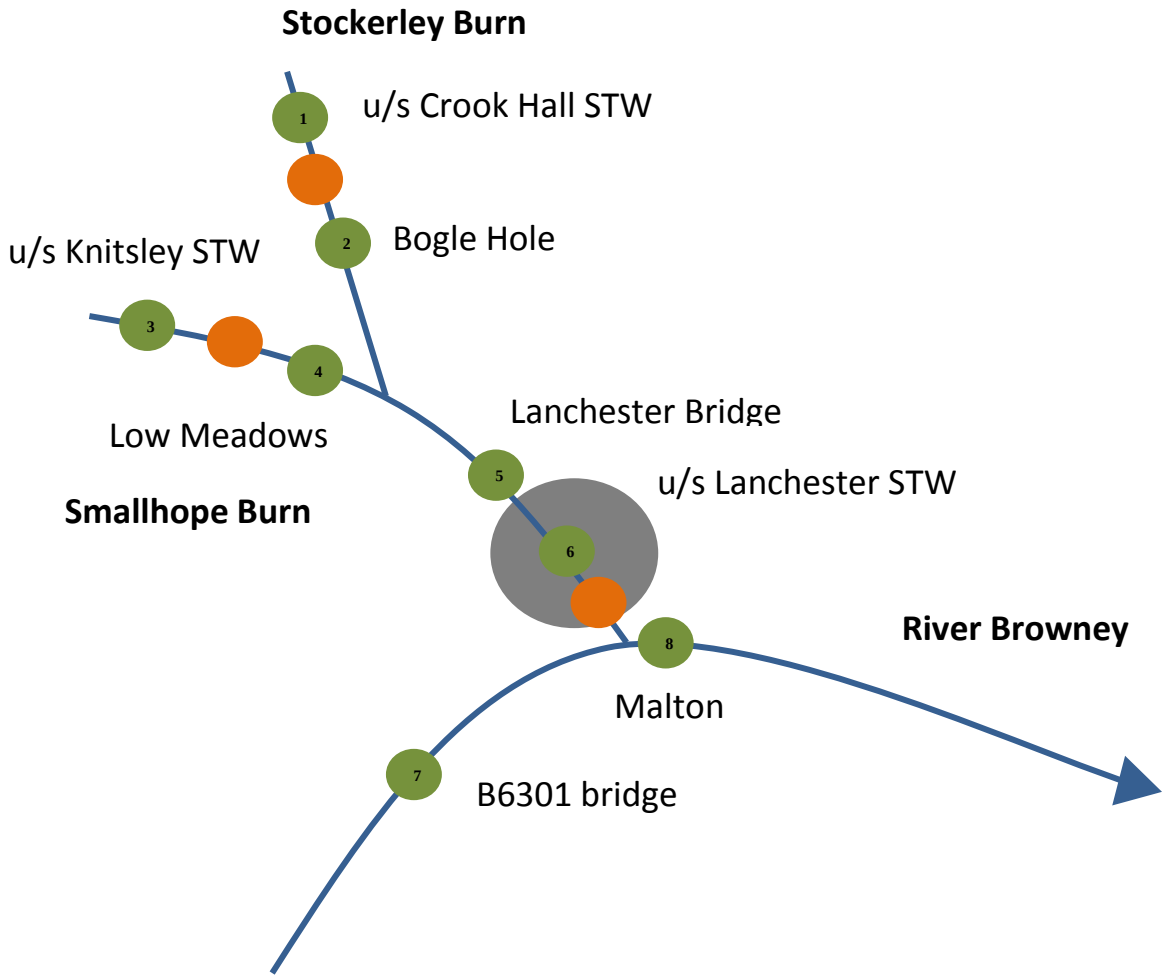


Figure 8.1 Schematic map of the upper River Browney and tributaries showing the location of STWs (orange circles), sampling sites (green circles) and the town of Lanchester (grey circle)

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.

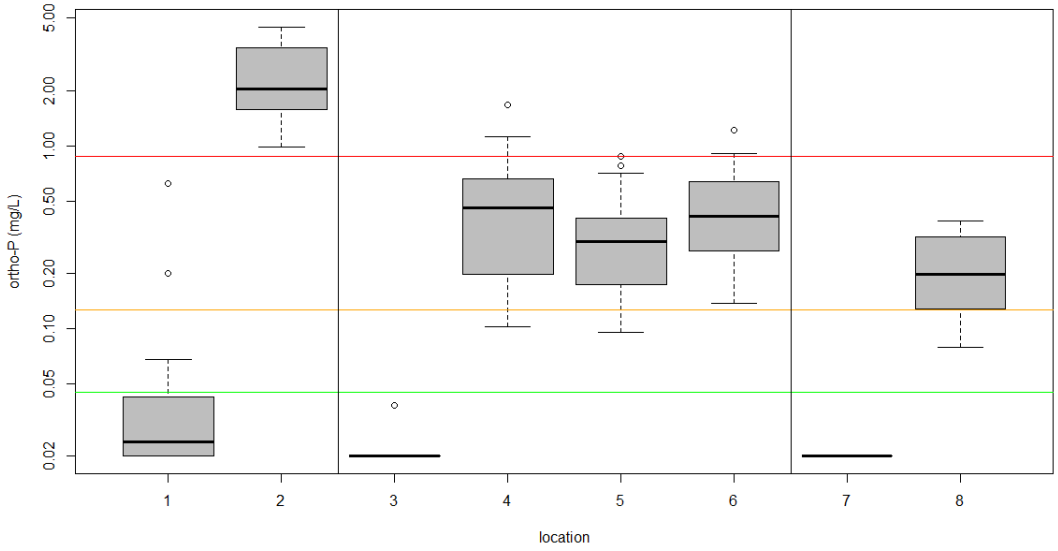


Figure 8.2 Variation in reactive phosphorus in Stockerley and Smallhope Burns and the upper River Browney

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

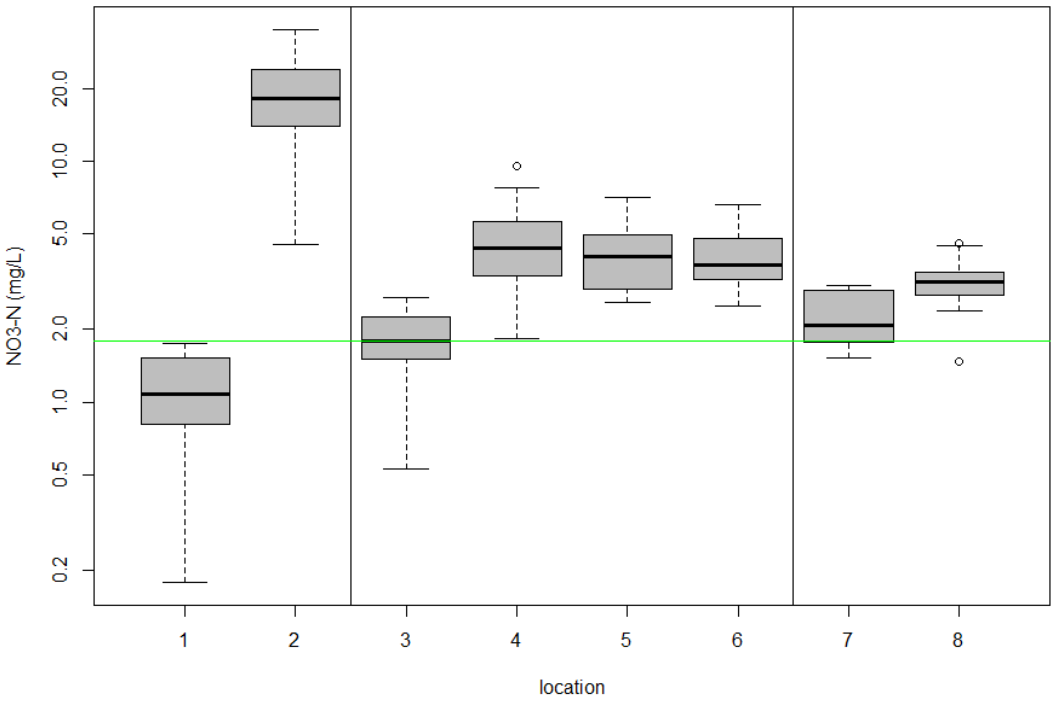


Figure 8.3 Variation in nitrate-N in Stockerley and Smallhope Burns and the upper River Browney

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 ($r = 0.753$, $p < 0.001$).

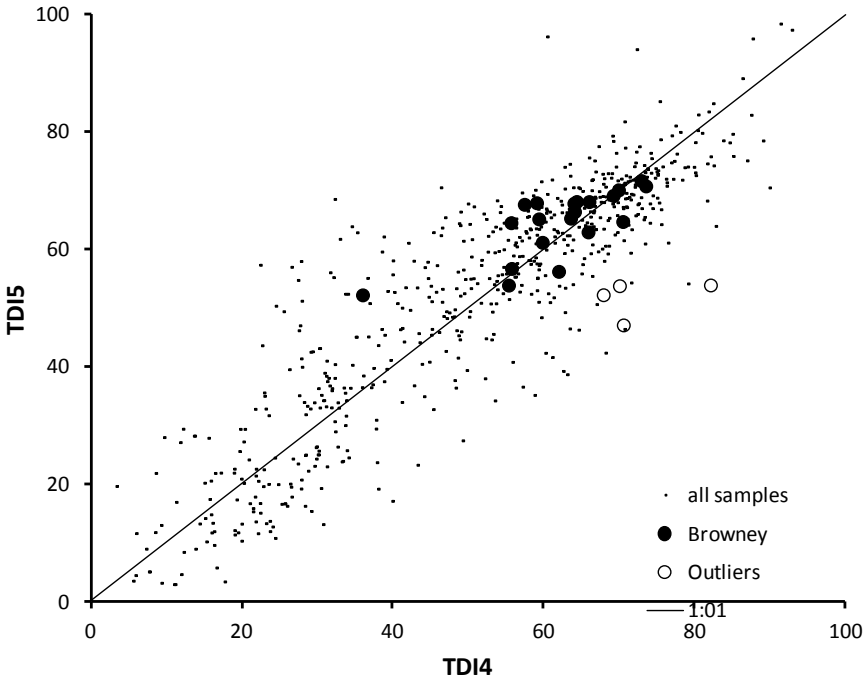


Figure 8.4 Relationship between TDI4 (LM) and TDI5 (NGS) with sites from River Browney subcatchments overlain

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.

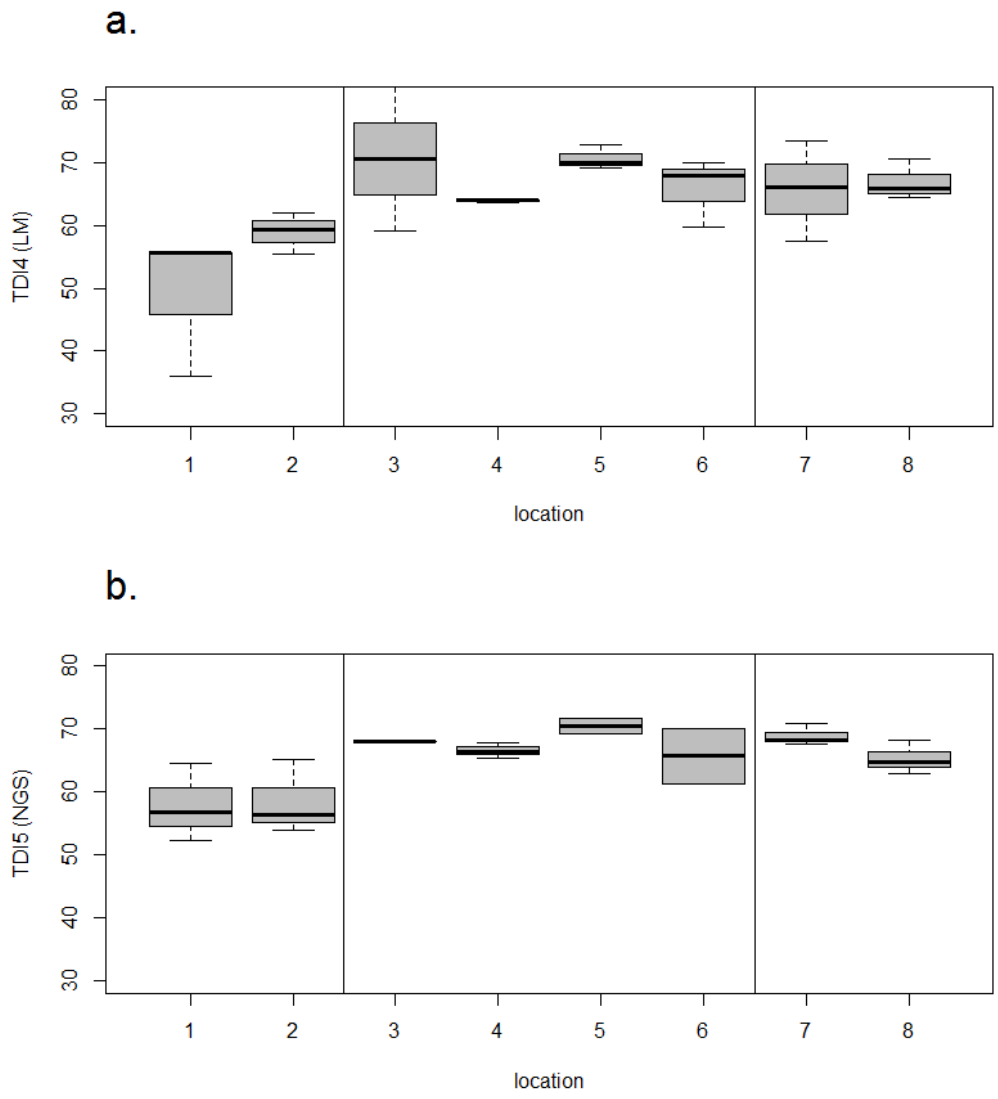


Figure 8.5 Variation inTDI4 (a) and TDI5 (b) in the Stockerley and Smallhope Burns and the upper River Browney

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).

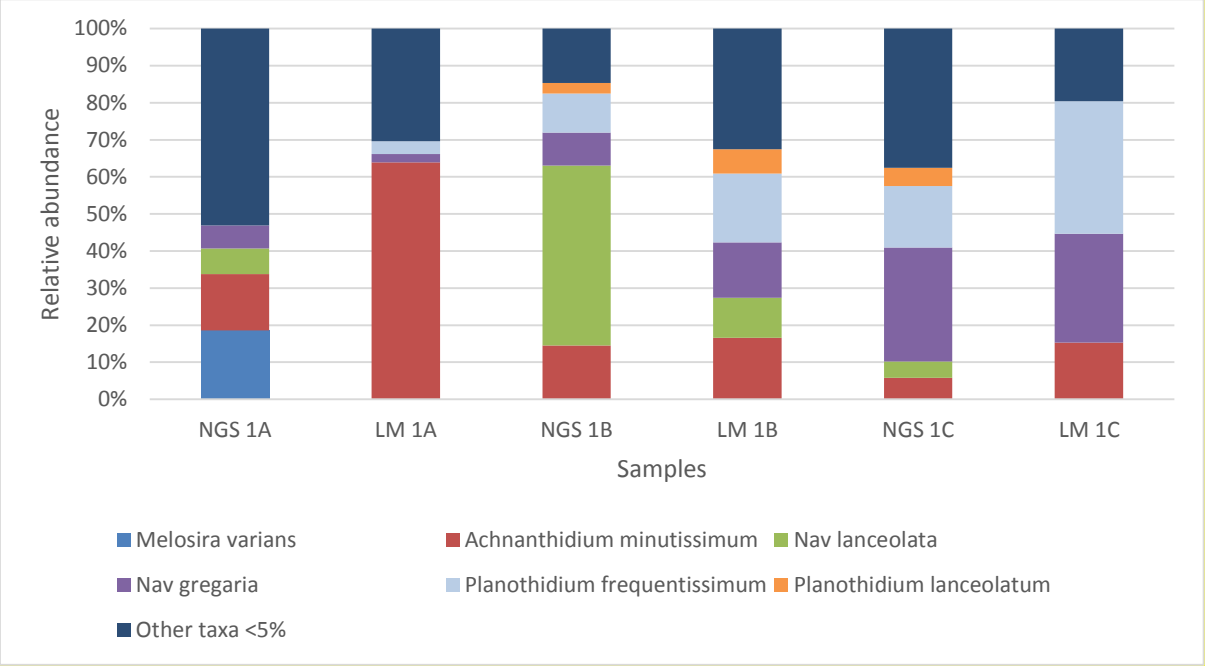


Figure 8.6 Variation in composition of taxa at site 1 between LM and NGS samples

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.

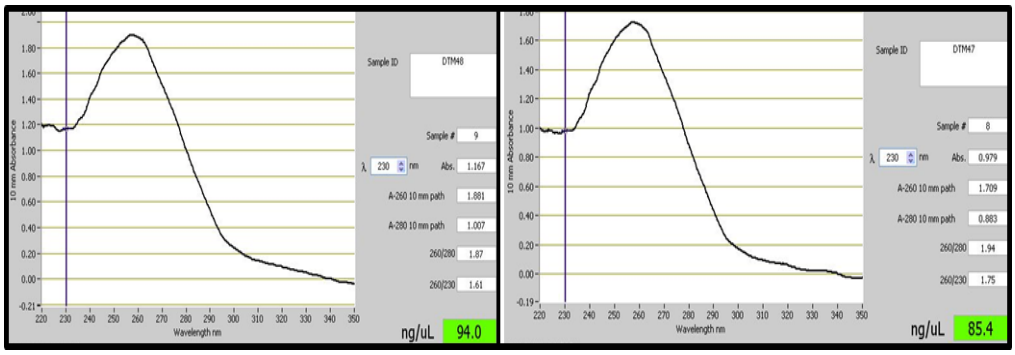


Figure A1.1 Representative spectrophometric analysis of DNA extracted from diatom samples

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.

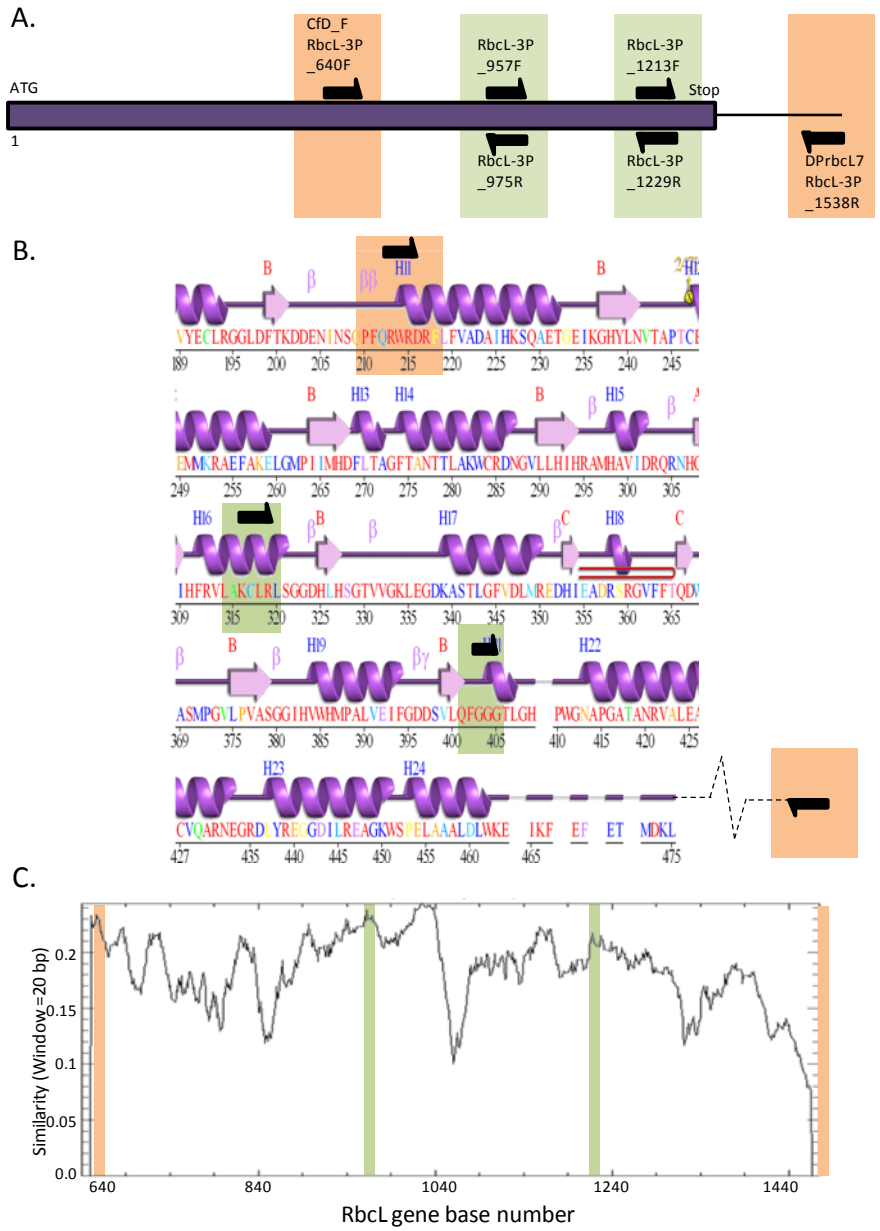


Figure A1.2 Design of rbcL NGS compatible primers

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.

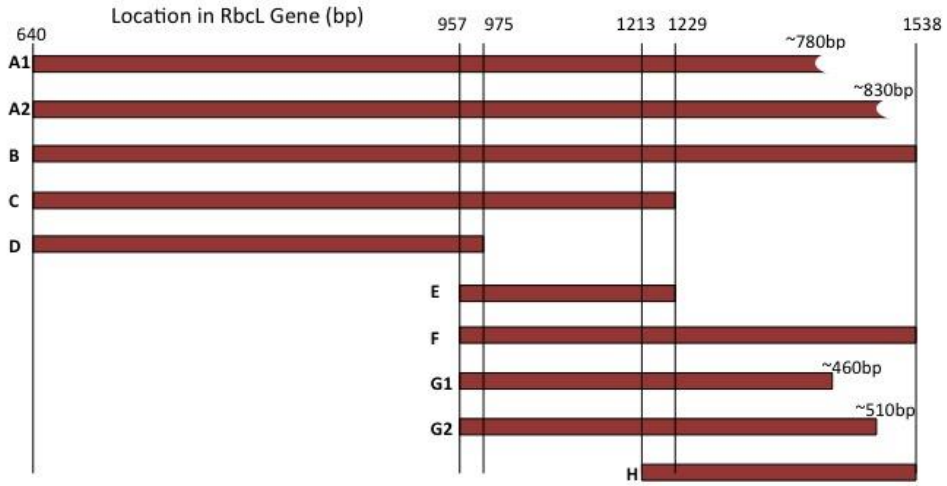


Figure A1.3 Regions of rbcL-3′ gene exploited for bioinformatic analysis

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

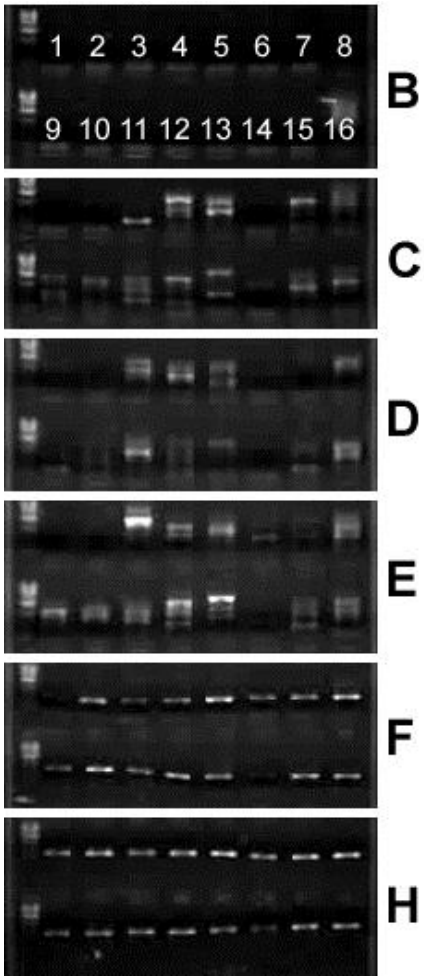
Primer region	No. Applicable reads from all GenBank records	No. OTUs @0.97	No. OTUs @0.95	No. OTUs @0.90
A1	383	202	139	48
A2	62	24	17	6
B	6	5	5	3
C	570	262	190	63
D	727	345	268	92
E	694	271	188	44
F	10	7	6	5
G1	520	214	152	48
G2	95	24	15	9
H	16	10	10	6

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.

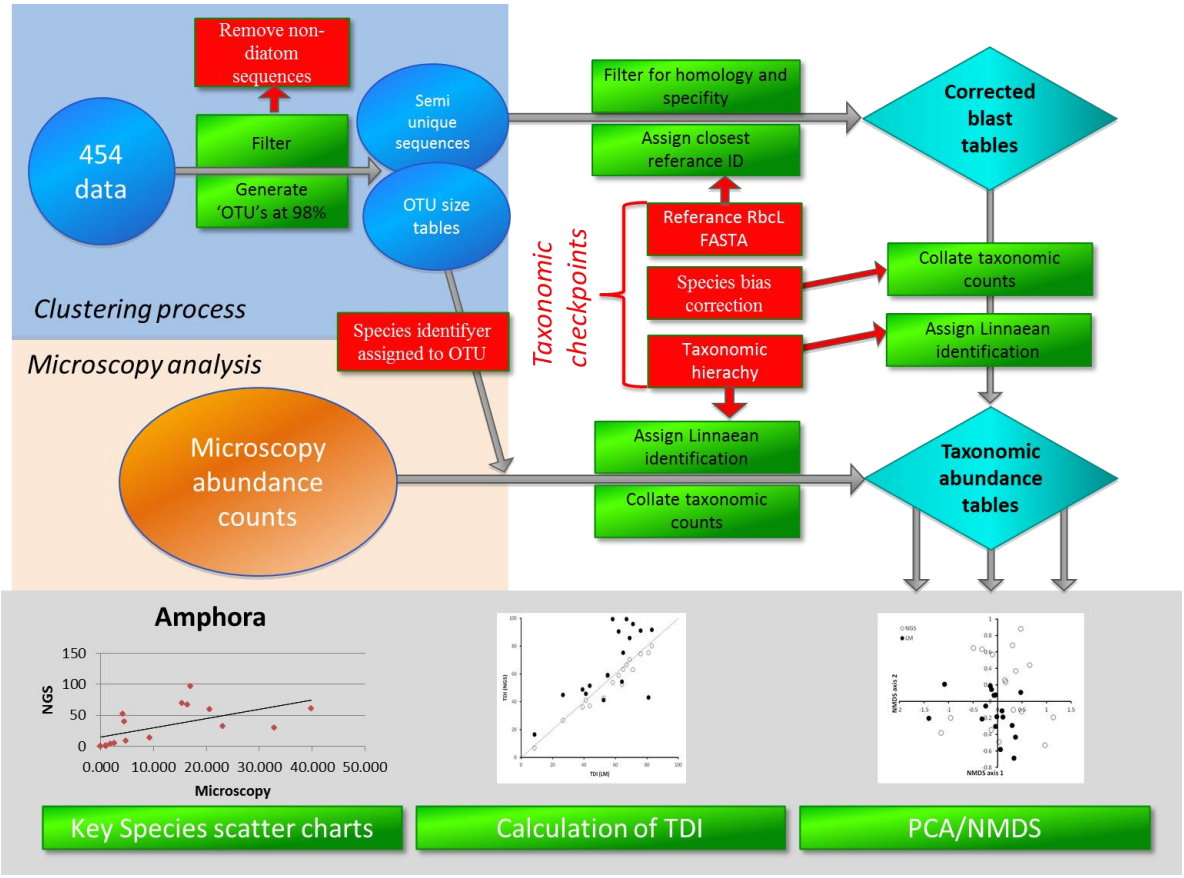


Figure A1.5 Overview of analytical workflow of PROMpT

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.

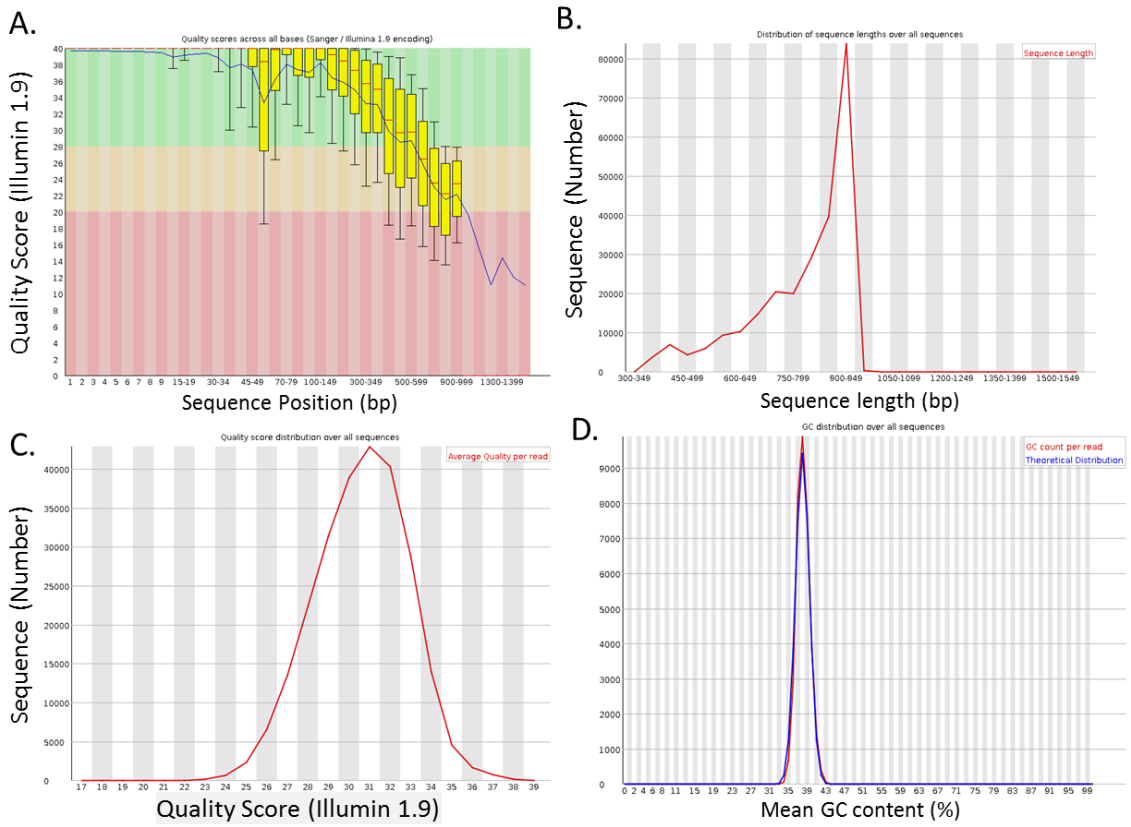


Figure A1.6 Quality analysis of raw GS FLX+ data

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.

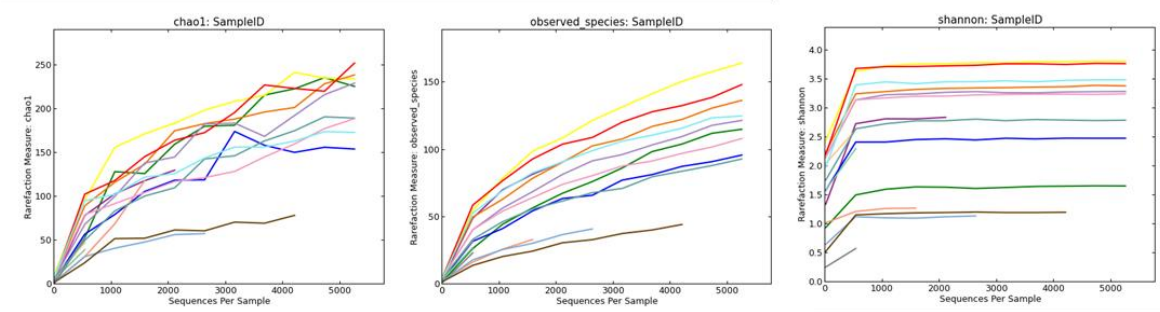


Figure A1.7 Diversity metric analysis of diatom community data

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.

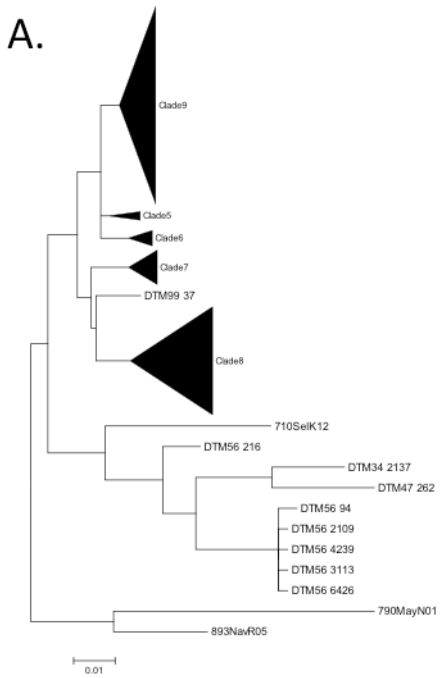
- Nomenclatural issues between reference databases were detected and harmonised, contributing to a significant improvement in species assignment.
- The inclusion of partial sequences (often the forward element of the rbcL-3′ fragment) significantly assisted species identification.
- Inclusion of specific 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

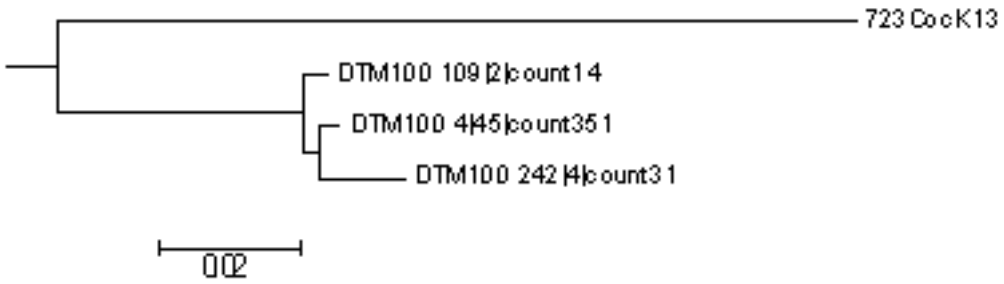
Figure A1.8 Phylogenetic analysis of *Eolima minima complex*: (A) maximum likelihood tree of *Eolima minima* OTUs; (B) estimates of average evolutionary divergence over sequence pairs within groups; and (C) estimates of evolutionary divergence over sequence pairs between groups

A1.2.4 Inferred taxa analysis: *Achnanthes oblongella*

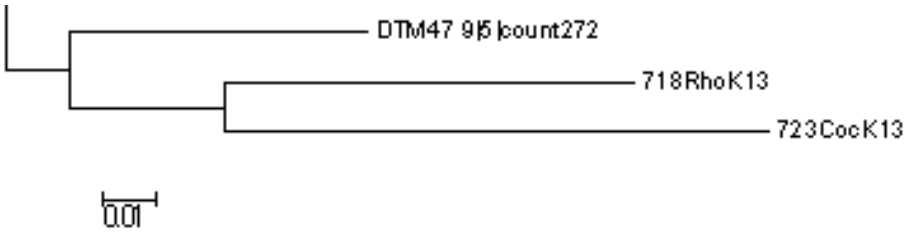
The dominant species occurring within samples were assembled in order to identify species that were missing from the barcode reference database and where there may be good evidence for species inference.

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)

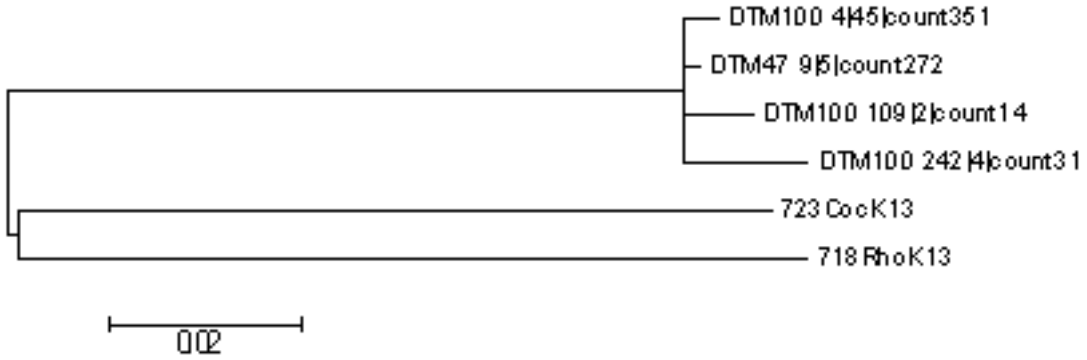


Figure A1.9 Clades of putative *Achnanthes oblongella*: orphan clades were identified individually from DTM100 (A), DTM47 (B) and then the relevant OTUs were combined into a single maximum likelihood guide tree (C)

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.

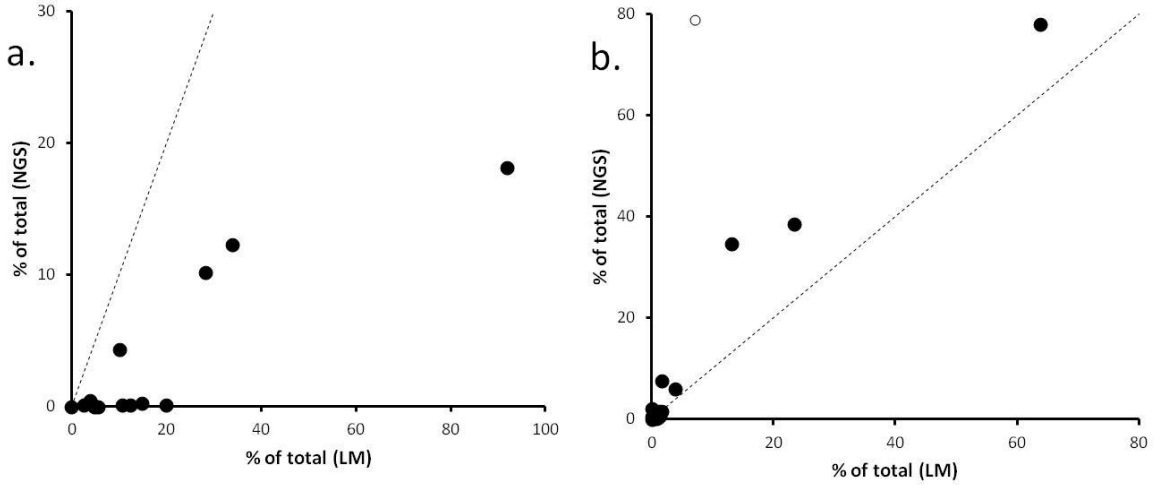


Figure A1.10 Comparison between representation of 2 taxa by traditional LM analysis and NGS: (a) *Achnanthidium*; and (b) *Eolimna*

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.

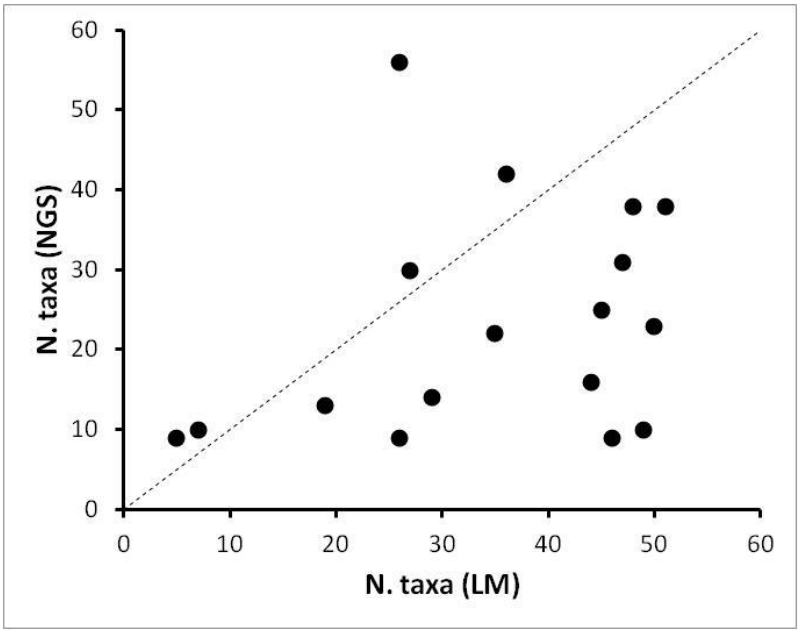


Figure A1.11 Comparison between number of taxa (N. taxa) recorded by LM and NGS

Notes: NGS taxa are based on OTUs; see text for more details.
The diagonal line indicates slope = 1.

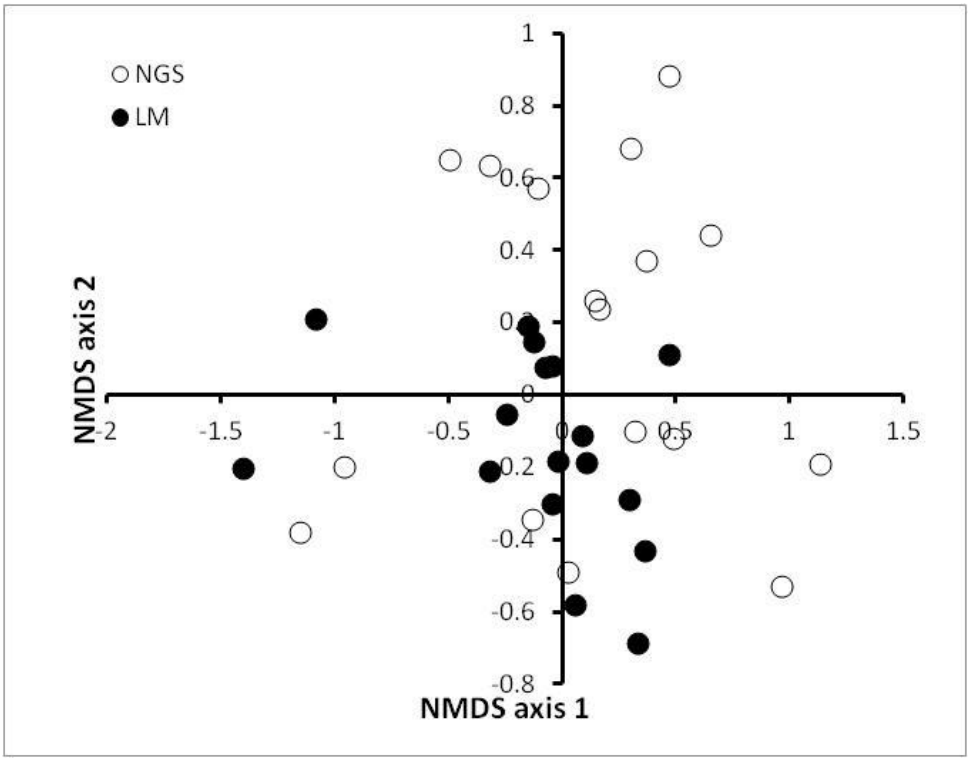


Figure A1.12 First 2 axes of NMDS analysis using combined data from samples analysed by LM and NGS

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.

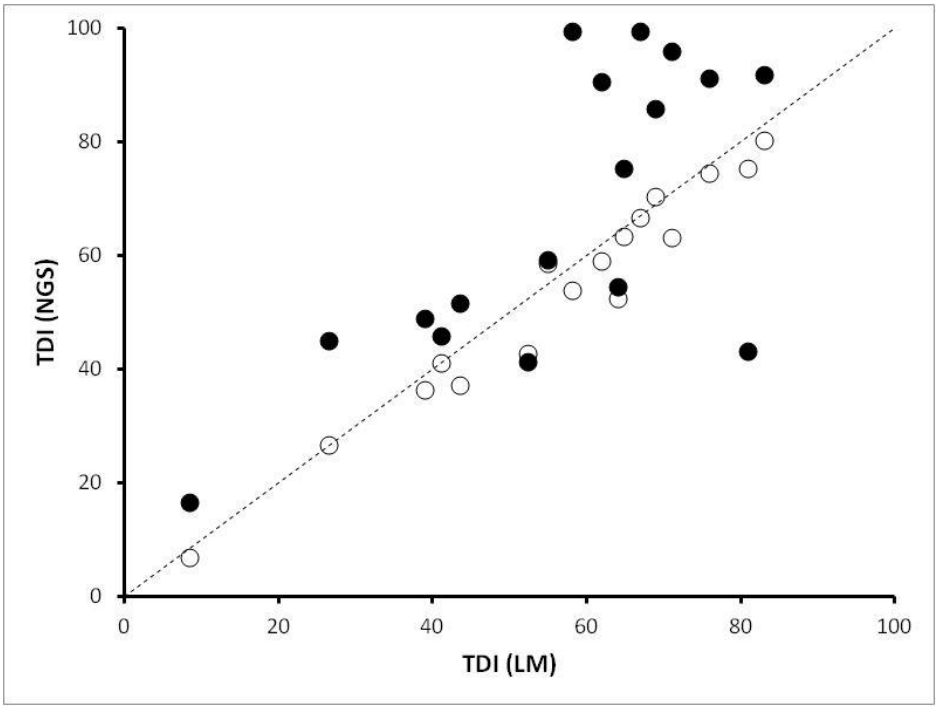


Figure A1.13 Comparison of TDI values computed using traditional LM analyses and NGS

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.

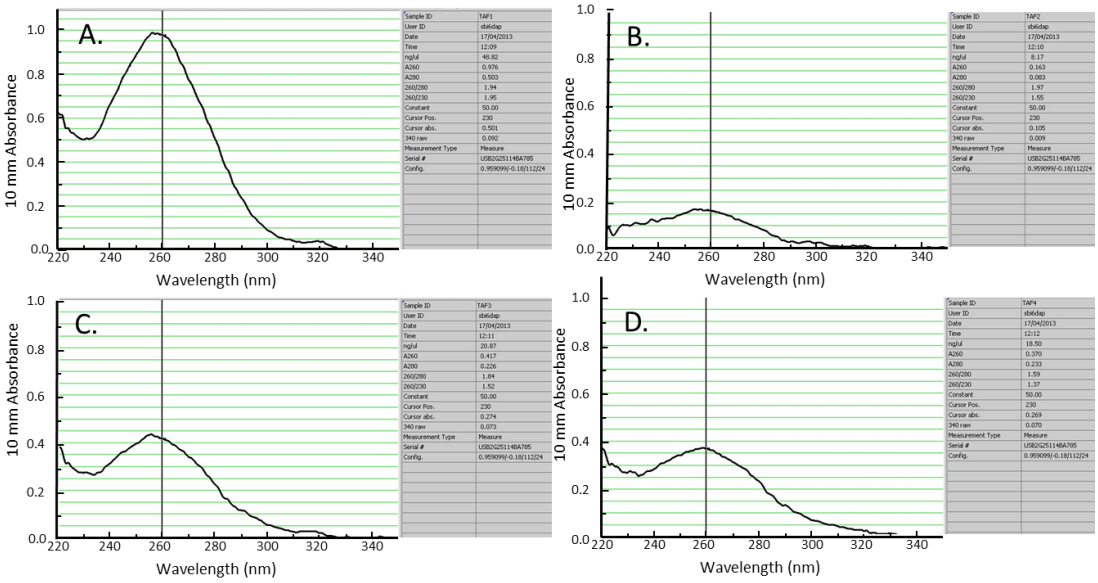


Figure A2.1 Compatibility test for diatom preservation with DNA extraction. DNA was extracted and analysed from diatoms subsampled from an individual community preparation and either immediately centrifuged and preserved at -20ºC (A) or maintained for 72 hours at room temperature with an equal volume 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}.

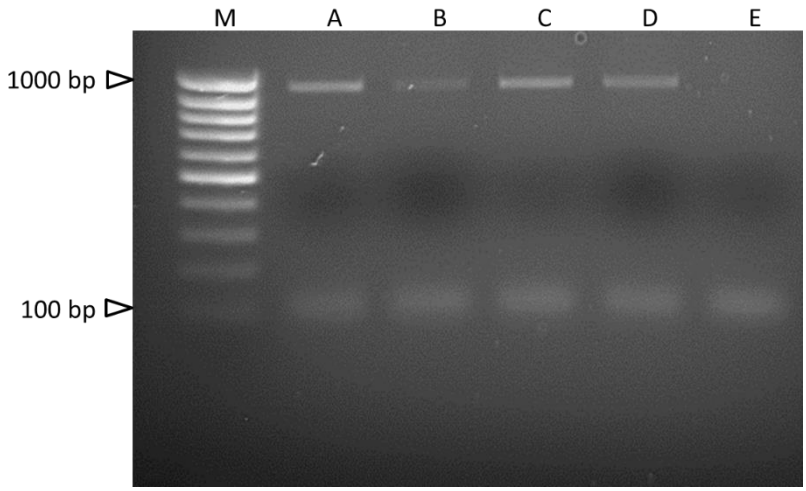


Figure A2.2 rbcL amplification from diatom assemblages after preservation treatments. DNA extracted from environmental samples after differential preservation were amplified using the rbcL-3′ primers previous reported by Hamsher et al. (2011). Lanes show the following samples: (M) 100 bp ladder; (A) 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.

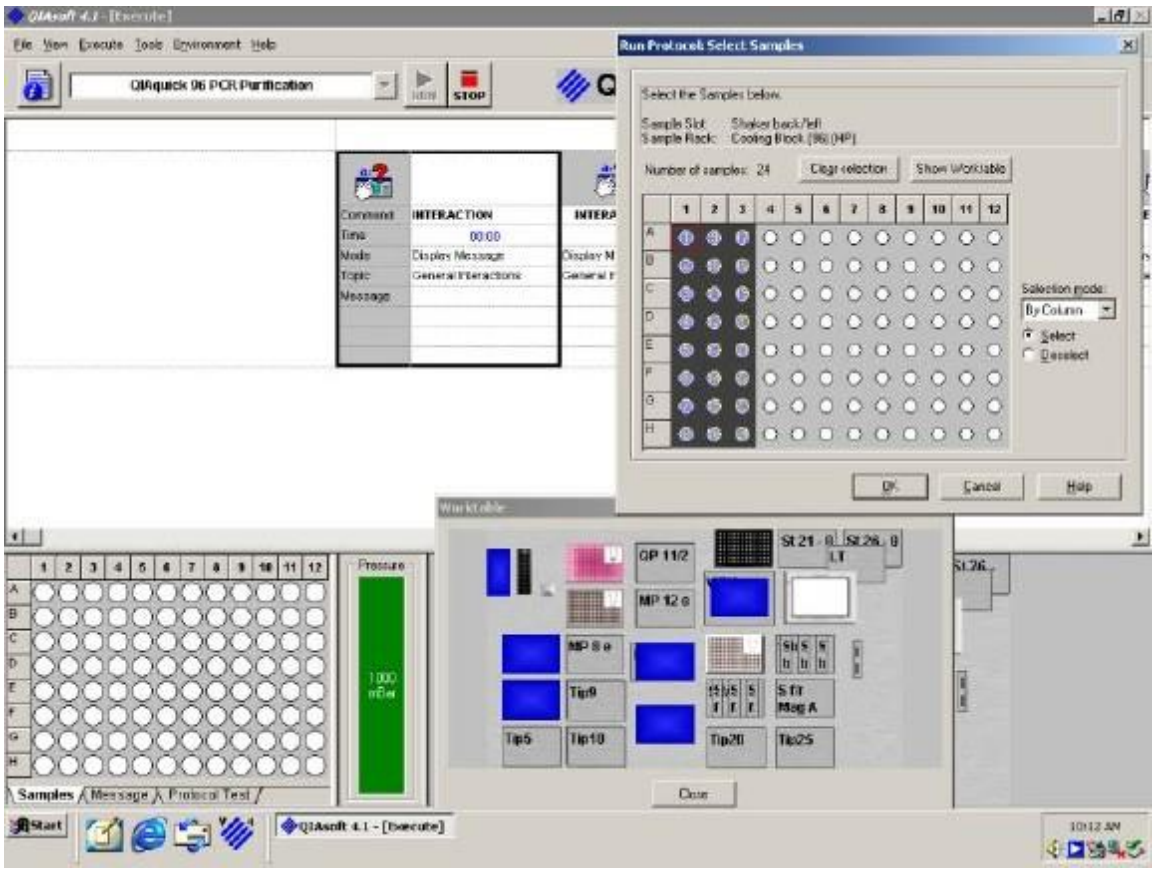
    c. Launch the QIAsoft 5 operating system.

    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.

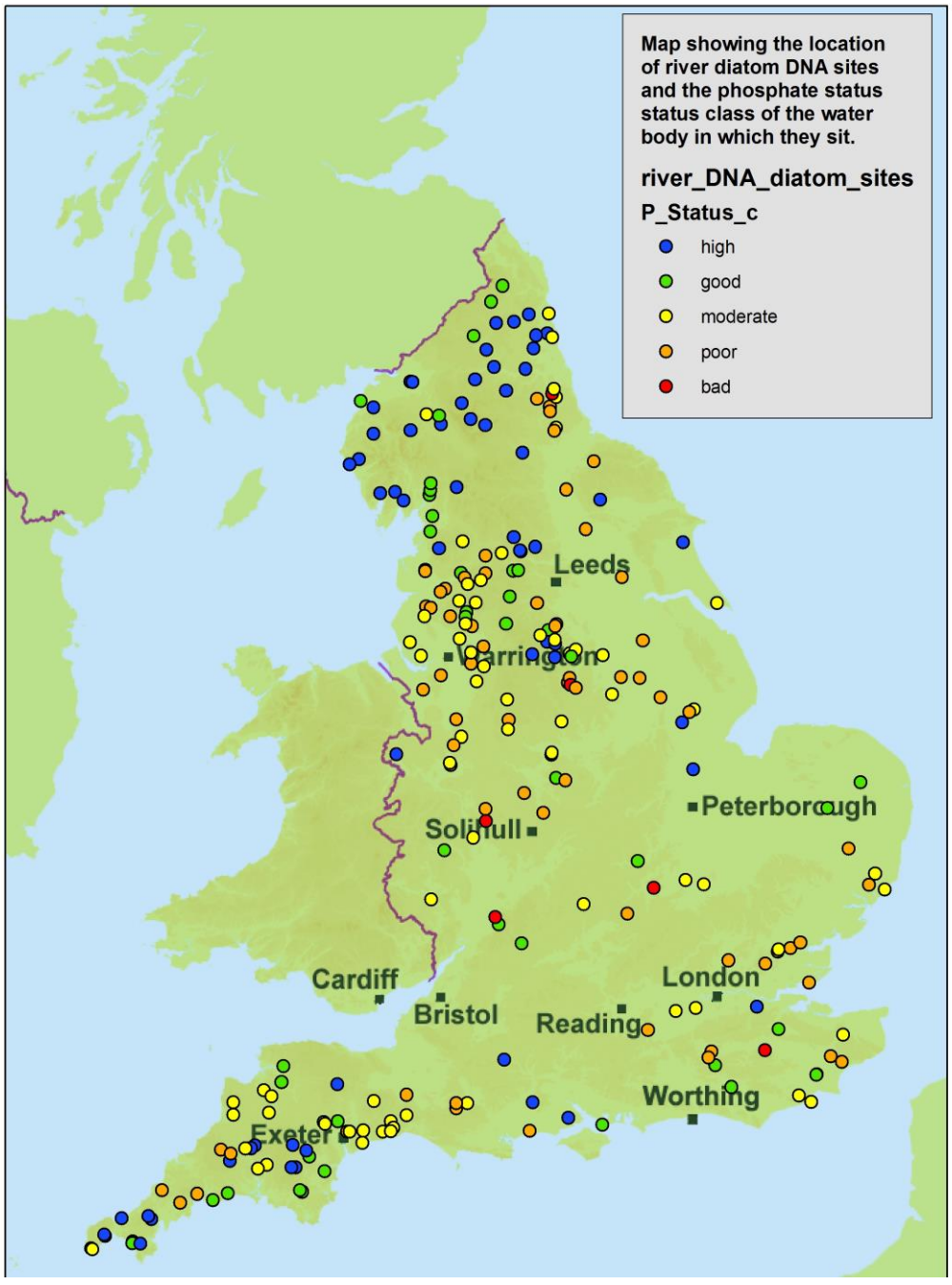
i. The final instruction before the program initialises will instruct you to place the S-Block containing your lysed samples onto the BioRobot. DO NOT press Next from this unless you are ready to proceed with the extraction! When ready press Next.

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.