Environment Agency

Further development of a DNA based metabarcoding approach to assess diatom communities in rivers

Chief Scientist’s Group report

February 2020

Version: SC160014/R

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

We help people and wildlife adapt to climate change and reduce its impacts, including flooding, drought, sea level rise and coastal erosion.

We improve the quality of our water, land and air by tackling pollution. We work with businesses to help them comply with environmental regulations. A healthy and diverse environment enhances people's lives and contributes to economic growth.

We can’t do this alone. We work as part of the Defra group (Department for Environment, Food & Rural Affairs), with the rest of government, local councils, businesses, civil society groups and local communities to create a better place for people and wildlife.

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

www.gov.uk/environment-agency

ISBN: 978-1-84911-447-9

Further copies of this report are available from our publications catalogue: www.gov.uk/government/publications or our National Customer Contact Centre: 03708 506 506

Email: research@environment-agency.gov.uk

Author(s):
Martyn Kelly, Neil Boonham, Steve Juggins, David Mann & Rachel Glover

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

Environment Agency’s Project Manager:
Kerry Walsh, Environment & Business Directorate, Research, Analysis & Evaluation

Project Number:
SC160014

Citation:
Environment Agency (2020). Further development of a DNA based metabarcoding approach to assess diatom communities in rivers.
Environment Agency, Bristol.

Research 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 Chief Scientist's 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
Chief Scientist

Executive summary

An earlier project (SC140024) demonstrated the potential for using next generation sequencing (NGS) for the analysis of the composition of benthic diatom assemblages in rivers. This opened the possibility that, for the first time, ecological assessment of an element of the freshwater biota whose assessment is required under the Water Framework Directive could be performed using molecular, rather than traditional morphology based, taxonomy. However, the project report identified a number of areas where additional work was necessary to bring the method to a point where it was ready for implementation. This report summarises the outcomes of that additional work.

An important issue highlighted by project SC140024 was the relationship between outputs from the current (TDI4) light microscopy (LM) based method and the prototype NGS method. Although agreement between the 2 approaches was good, there were still often differences in the output for individual samples that could not be explained. Likely reasons include:

- gaps in the barcode database
- differences in quantification between the 2 methods
- general stoichiometric factors

Both the LM and NGS methods were therefore recalibrated in this project to optimise their relationships with inorganic nutrients and with each other. This exercise used a larger dataset (1,367 paired samples) and a larger barcode database (346 taxa). Different statistical approaches were adopted to obtain the strongest-possible relationship; however, there was only a small overall improvement, even with the best of these (response curves) and, in light of the reduced parsimony and general transparency associated with such methods, a straightforward weighted averaging approach was retained.

The relationships between the optimised LM and NGS versions and the inorganic nutrient gradient were of a similar strength. The prototype version of the NGS tool was considerably less stringent than the current LM TDI4 model (DARLEQ2), with 21% of sites classified more stringent. In contrast, after recalibration, both the optimised LM metric (TDI5LM) and the NGS variant derived from this (TDI5NGS) had much lower bias (-0.8% for the whole dataset and 3.4% when applied solely to sites with lower alkalinity, that is, <120mgL^{-1} $CaCO_3$). The classification results based on outputs from 666 sites were strongly aligned with 64% of samples classifying in the same ecological status class – 31% within one status class and 4% greater than one class. TDI5NGS has a tendency to be slightly more stringent than TDI5LM.

Concerns about the reference model underlying DARLEQ2 prompted some exploration of the consequences of changing this, as well as shifting from the use of LM to NGS data. When an alternative reference model (still under development) was applied, the bias increased substantially. While decisions about the most appropriate reference model are beyond the scope of this project, this helps to place the bias observed between LM and NGS into context.

Current practice for collecting samples for analysis by LM involves the use of the same toothbrush (after cleaning) at several sites, with river water to wash the biofilm off the substrates. The greater sensitivity of NGS analysis necessitated a rethink of this approach and experiments were conducted to compare the use of brand new versus used toothbrushes and river water versus distilled water for washing. Contamination from both used toothbrushes and river water was detected, albeit at a low level in each case. The possibility of occasional more significant contamination occurring cannot be ruled out and it is recommended that samples are collected with single use toothbrushes and distilled water from now on.

In addition to looking at the performance of the NGS method in rivers, LM and NGS data from lakes in England were compared to provide an insight into the scale of differences between the 2 methods. Application of the current lake metric (LTDI2) to NGS data without any modification resulted in a good linear fit when compared with its use with LM data; as for the river study, both approaches had a similar strength of relationship with the inorganic nutrient gradients in the lake dataset. It should therefore be possible to develop a functional lake variant of the NGS tool with relatively little extra work.

Finally, the new approach was applied to an investigation of the source of water quality issues in a catchment in Devon. Both LM and NGS methods gave ambiguous results, reflecting a catchment experiencing a variety of ecological stresses. Although there was not complete agreement between results from the LM and NGS approaches, both indicated that there were issues both above and below a sewage treatment works that was the primary focus of concern. Important lessons can be learned from intensive studies such as this, which emphasise that, however good the agreement between 2 metrics, interpretation of the behaviour of individual taxa in NGS samples is not necessarily the same as interpretation based on LM data. The shift from LM to NGS as the primary means of collecting data needs to be accompanied by an adjustment of the understanding of how individual taxa combine to give an indication of the condition of a water body.

Acknowledgements

We are extremely grateful to Tim Jones (Environment Agency) for his invaluable and continuous support throughout the project. Thanks too go to operational staff from the Environment Agency for carrying out the light microscopy analysis on the calibration dataset samples and, in particular, to Michael Thomas for providing background information on Polly Brook in south Devon. We also thank Sarah Pritchard (Beacon Biological) for help with preparing diatom samples and Geoff Phillips for discussion about alternative approaches to establishing expected values of the TDI.

Contents

List of tables and figures

1 Introduction

The development of a metabarcoding approach to the use of diatoms to assess ecological status is described in Environment Agency (2018). The outcome of that project was a prototype next generation sequencing (NGS) metric that showed good agreement with the current analytical method based on light microscopy (LM). Behind this lay a substantial body of work that identified a short region of the ribulose bisphosphate carboxylase large chain (rbcL) gene that was suitable for high throughput sequencing using the Illumina MiSeq platform. This bioinformatics pipeline processed the Illumina output and a barcode database that allowed the appropriate Linnaean binomial to be assigned to each read. Together, these offer a viable alternative to the current approach and offer the potential of the first nationwide application of NGS technologies to routine assessments for the Water Framework Directive (WFD).

During the project (SC140024), a few issues were identified which required further work before the approach could be made fully operational. This report describes progress towards those goals. In addition to the data-driven improvements discussed in the report, a number of 'back room' improvements have been incorporated. These are summarised below, along with some comments on preparation for implementation.

1.1 Further development of NGS diatom assessment tool

At the end of the previous phase of work, the barcode database contained 702 barcode sequences representing 170 out of approximately 2,800 diatom species recorded from Britain and Ireland. Although project SC140024 showed that this selection was adequate to capture the main patterns of variation in the dataset, there was still considerable scatter in relationships between analyses performed using LM and NGS data. As a large number of NGS reads were not assigned to a species, there seems to be potential for better performance via an expanded barcode database. This was a desk based activity, adding newly published barcodes as they became available via online databases, rather than by isolating and culturing new strains. The barcode database now contains 1,232 barcode sequences representing 346 species. This includes:

- some taxa not yet recorded from the UK, but which should improve the overall efficiency of the bioinformatics at assigning barcodes
- 29 planktic taxa that are not used for calculation of the Trophic Diatom Index (TDI) but will ensure that as many reads as possible are assigned to taxa (see Appendix 1)

The previous project also highlighted problems with the quantification of NGS data, and in particular with a few taxa (for example, *Melosira varians*) dominating assemblages even when present in LM analyses at relatively low percentages. The potential for improving post-bioinformatics data handling to address this is discussed in Section 2.

The high sensitivity of NGS methods generally necessitates a new look at aspects of sample collection. Although project SC140024 showed the current sampling method of scrubbing the upper surfaces of cobbles with a toothbrush was an effective means of collecting samples for NGS analysis, current practice permitted reuse of the toothbrush at several sites whereas it is standard practice for samples for NGS analyses to be collected using single use and/or aseptic equipment. Understanding the extent to which the sampling method itself might introduce uncertainty to analyses may yield insights into the relationships observed in SC140024. It will also inform decisions about whether single use equipment should be used when collecting samples. This is considered in Section 3.

Having developed a metric for ecological assessment of rivers for the WFD using data generated by NGS, the potential for developing a companion metric for ecological assessment of lakes is considered in Section 4. This investigation used samples collected from lakes in England during 2014 and 2016, and offers an insight into the scale of modifications that would be necessary if the approach were also to be adopted for lakes.

Finally, the method is applied to an ongoing operational investigation to evaluate the effectiveness of the method for detecting ecological changes in response to point and diffuse pollution in a small catchment (see Section 5).

1.2 Preparation for implementation

A number of additional steps have been taken to prepare the method for implementation. These include:

- ensuring that each taxon in the barcode database has an appropriate code to allow input of NGS data to the Environment Agency’s biological database, BIOSYS
- updating of the classification software (Diatoms for Assessing River and Lake Ecological Quality, DARLEQ) and associated guidance
- knowledge transfer and staff training in implementation of the new method

2 Development of NGS based variant of TDI (TDI5)

2.1 Introduction

Section 6 of Report SC140024 (Environment Agency 2018) described the development of the NGS based metric (TDI5) to assess ecological status using diatoms. Improvements to the bioinformatics, expansion of the barcode database and the availability of more samples, along with lessons learned during the development of the prototype NGS metric, offer an opportunity to revisit that process in order to develop a stronger model.

Many of the stages described in SC140024 are repeated here, but several new steps have been added. As the NGS model was developed by an ordination technique that calibrated NGS data to obtain the best fit possible to the current LM model, some initial optimisation of the LM model was undertaken at the outset. Following this, the new models – LM and NGS – were compared to models developed using ordinations against the pressure gradients.

Finally, in light of concerns that the reference model currently used for predicting ecological status using diatoms may be flawed, an approach using a plausible alternative reference model is also considered. This is, in part, an exercise to put the scale of differences between LM and NGS into perspective, as it demonstrates that a shift to a new reference model is likely to have much greater implications for classification than the switch from LM based to NGS based approaches.

An important message that emerged from project SC140024, and which is reinforced in this report, is that NGS data are fundamentally different to LM data, with the relative proportions of taxa in a sample often differing substantially between the 2 methods.

To understand this, it is first necessary to appreciate that current LM based methods have a number of inbuilt biases, albeit biases which those analysing and interpreting the data can accommodate. This means that changes in ecological status along a stream or over time should be explainable in terms of fluxes of individual taxa.

NGS data too have some biases, though end users – at least at first – will be less familiar with these. The highly automated nature of high throughput NGS is that outputs could, potentially, be packaged into a ‘black box’ that produced standardised status assessments. This could in turn open the door to statistically complicated models hidden behind a user-friendly ‘front-end’. However, there are benefits to having a relatively straightforward model behind the assessments as, even if users have to ‘recalibrate’ their understanding of the fluxes of taxa along ecological gradients, there is a transparency to the assessment process.

The discussion around choice of models below therefore needs to be considered in terms of:

- statistical strength

- the ability to discriminate between different levels of ecological status
- the issue of transparency/explainability

When the gains in statistical power are trivial, the general approach adopted was to opt for the most parsimonious of the approaches available.

2.2 Dataset summary

The combined dataset contains LM and NGS samples collected from a variety of sources. These include:

- samples collected in 2014 and early 2015 as part of the studies documented in Environment Agency (2018)
- samples collected as part of the Environment Agency's routine surveillance monitoring program in 2016

NGS and LM data were harmonised against the DARLEQ master taxon dictionary, which has been updated to reflect new taxa recorded in the LM and NGS datasets and nomenclature changes since the release of DARLEQ 2.0. 2014 and 2016 biological data were also matched against environmental data from England, Wales, Scotland and Northern Ireland.

Environmental variables included are:

- phosphate-P (P-PO_4)
- nitrate-N (N-NO_3)
- ammonium-N (N-NH_4)
- alkalinity
- conductivity
- pH

Biochemical oxygen demand (BOD) data were only available for 111 samples and so this variable was not included in subsequent analyses.

Environmental data are expressed as either the mean (alkalinity and pH) or geometric mean (all other variables) of all available data for the period 2012 to 2016.

Detection limit information was not always available and so measurements below the detection limit were taken as the detection limit. This may overestimate actual values at low concentrations. However, the water chemistry data were used primarily to validate diatom metrics and used only to modify the indicator values of a few, rare taxa (see Section 2.4); they will therefore have negligible effect on metric calculations.

After taxonomic harmonisation, the LM and NGS datasets contained a total of 1,412 and 1,515 samples respectively, giving a combined dataset of 1,367 paired LM and NGS samples. This combined dataset was further screened to remove NGS samples with a read count of <500 reads of non-planktic taxa in the barcode database, resulting in a final paired LM and NGS dataset of 1,337 samples.

Some NGS and/or LM samples could clearly be considered outliers in that their species composition lies outside the range of variation expected from the LM and NGS methods. This issue is discussed further in Section 6, but at this stage no attempt was made to identify or remove such outliers.

2.2.1 Species profiles

After taxonomic harmonisation, the LM and NGS datasets contains a total of 493 and 306 non-planktic taxa (that could be identified using the barcode database) respectively. The distribution of total number of reads once planktic taxa had been excluded is shown in Figure 2.1a and the distribution of the relative abundance of unassigned reads in the NGS dataset in Figure 2.1b. The average number of reads is 41,048, with unknown taxa accounting for over half the total read count in 354 samples.

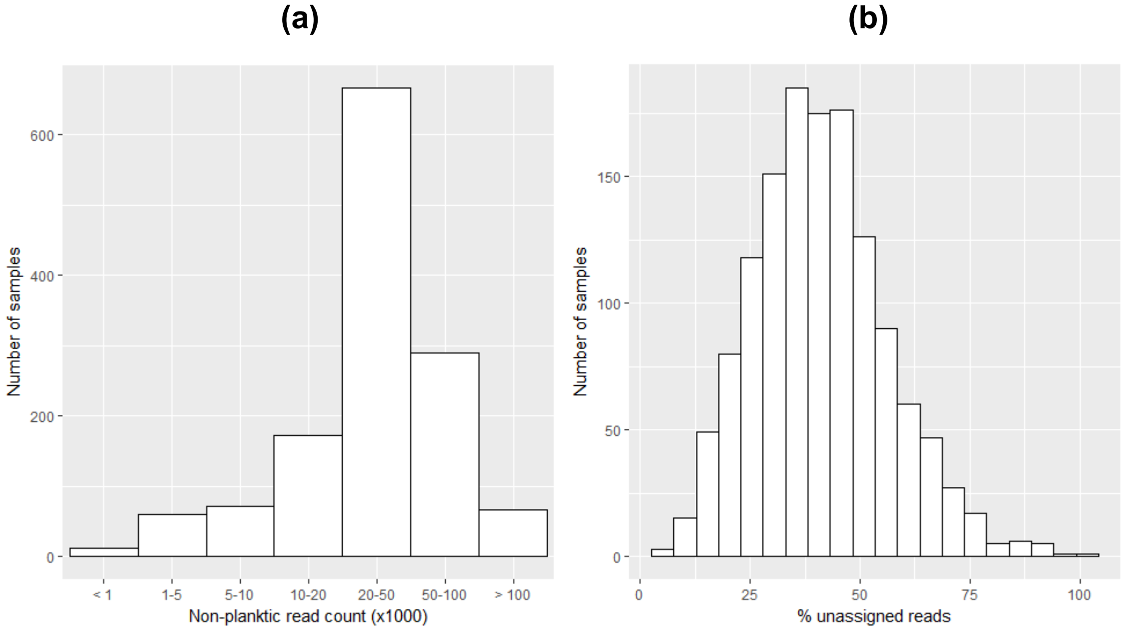

Figure 2.1 (a) Number of reads of non-planktic taxa in the NGS dataset. (b) Proportion of reads that could not be assigned to Linnaean binomials

Notes: Thirty samples with <500 reads are excluded from these plots.

Once unknowns are removed and the proportions of other taxa recalculated to give a total of 100%, species profiles can be compared (Figure 2.2). There are some similarities, but also striking differences. For example, *Melosira varians* (ME015A) and *Navicula lanceolata* (NA009A) are far more abundant (on the right hand side of the bottom graph in Figure 2.2) in the NGS dataset, while *Achnanthidium minutissimum* (ZZZ835) is more abundant in the LM than in the NGS dataset. This taxon is at the top right for LM but approximately at the intersection of Max = 75, N2 = 200 for NGS.

Although omitting unassigned reads from calculations of relative abundance means that reported values are probably higher than the ‘true’ relative abundances, there is no way of knowing which of the unassigned reads relate to other benthic and which to planktic species (which are omitted from the total count for LM analyses). Most importantly, the metrics only ‘see’ those taxa that are represented in the barcode database. Excluding unassigned reads from the total count therefore means that the abundance of each taxon in a sample is expressed in the way that it is weighted in the metric calculations.

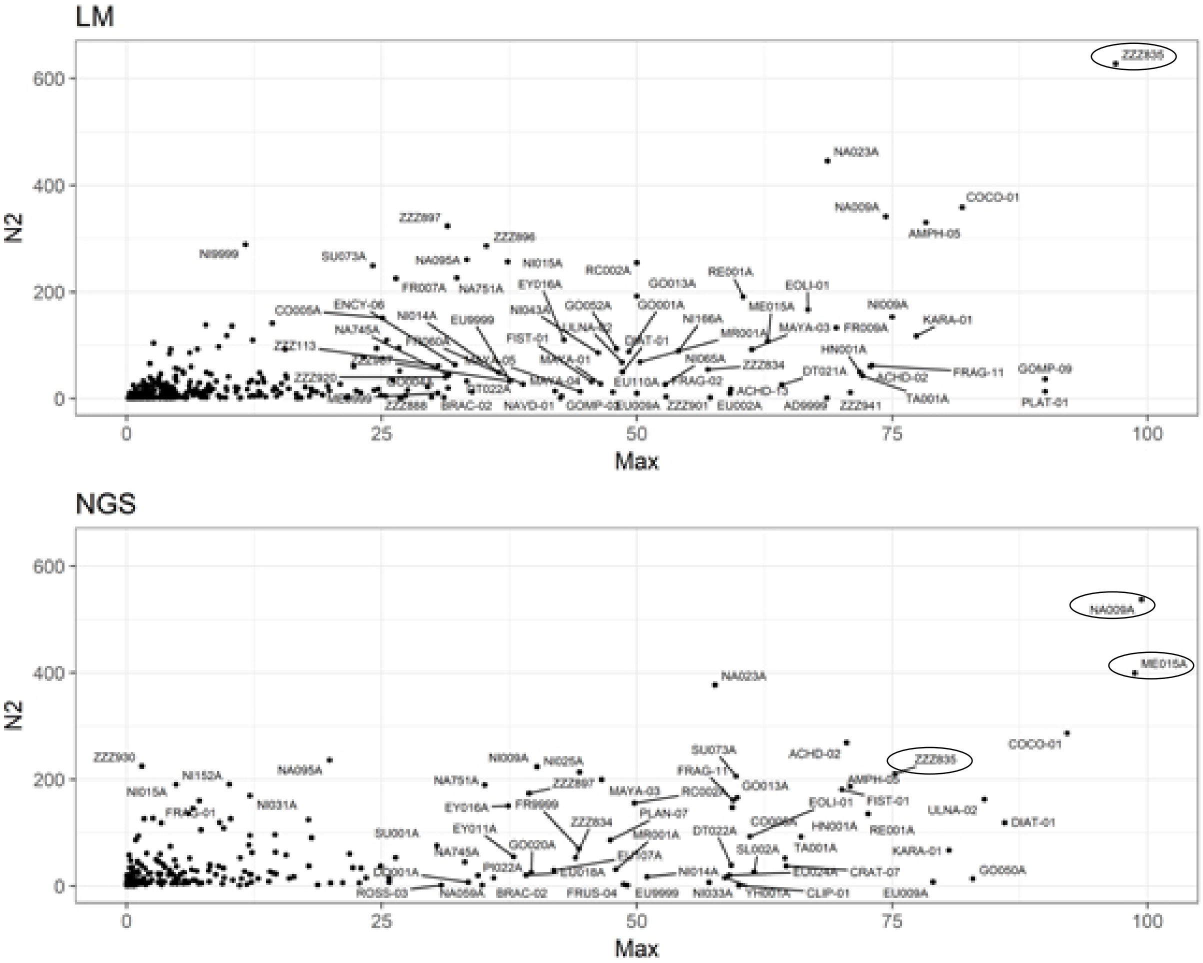

Figure 2.2 Species profiles (maximum abundance versus Hill’s N2 diversity) for the LM (top) and NGS (bottom) datasets

Notes: Hill's N2 diversity encapsulates the likelihood of a species being found in a sample (high values: very commonly found; low values: rare).
Species codes are given in Appendix 2.

2.2.2 Environmental data

Over a thousand of the paired NGS and LM samples could be matched to water chemistry; Table 2.1 and Figure 2.3 summarise the coverage of the most important variables.

The dataset has good coverage of the alkalinity and conductivity gradients, with coverage of the latter falling off at about 1,000μS cm^{-1}, suggesting very limited coverage of brackish conditions.

Most samples have a pH of around neutral, with just a small number with a pH <7.

The distribution of phosphorus values reflects:

- the routine use of 3 different limits of detection for routine analyses within the Environment Agency (0.02, 0.01 and 0.001 mgL^{-1})
- the inclusion of data from other organisations in the dataset
- the subsequent amalgamation of monthly samples to compute averages

The nitrate-N dataset, by comparison, extends down to 0.1 mgL^{-1} and there are far fewer values at or below the limit of detection. Ammonium-N is included in this summary to show the relatively small number of samples in the dataset with evidence of elevated levels of organic pollution (the limited data for BOD show the same trend).

Table 2.1 Summary statistics of selected environmental variables for the combined LM and NGS dataset

	N	Mean	Standard deviation	Median	Minimum	Maximum
pH	1,043	7.8	0.3	7.8	5.8	8.4
Conductivity (µScm^{-1})	1,027	364	293	276	32	2,740
Alkalinity (mgL^{-1} $CaCO_3$)	1,167	86.5	73.4	58.7	1.7	381.6
P-PO_4 (mgL^{-1})	1,042	0.096	0.171	0.038	0.001	1.778
N-NO_3 (mgL^{-1})	1,040	2.744	3.039	1.781	0.041	27.254
N-NH_4 (mgL^{-1})	1,040	0.057	0.072	0.037	0.005	0.884

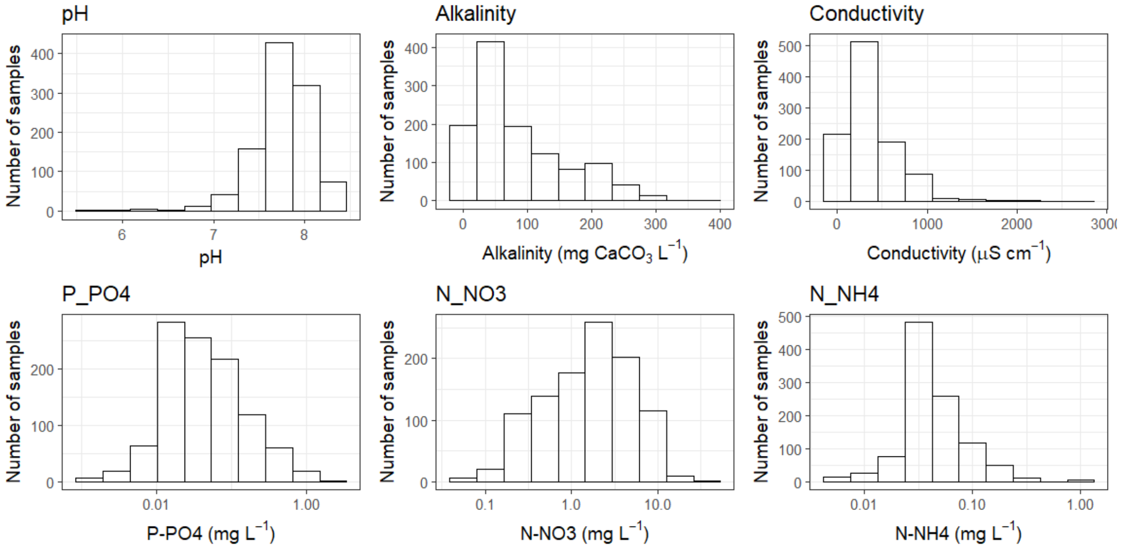

Figure 2.3 Distribution of samples along key environmental gradients

2.3 Optimising the LM metric

In project SC140024, the species indicator values (species nutrient sensitivity scores) for the prototype NGS metric were derived by weighted averaging using the current LM based variant of the TDI (TDI4) as the explanatory variable.

Some subsequent analyses led to a suspicion that some of the indicator values used in the current TDI4 may be overestimates or underestimates of their sensitivity to the nutrient pressure gradient. Before attempting to derive a new version of the NGS metric, the relationship between the TDI4 and the nutrient pressure gradient was therefore evaluated and, where appropriate, species indicator values adjusted.

The first step in this process was to derive a weighted average (WA) model that directly calculates species indicator values as the weighted mean of their distribution along the pressure gradient. These WA indicator values have been shown to be good estimates of species optima (Ter Braak 1986).

Models were derived by comparing TDI4 to P-PO$_4$, N-NO$_3$ and the first component of a principal components analysis of P-PO$_4$ and N-NO$_3$ (PC1). This, in effect, combines the phosphorus (P) and nitrogen (N) gradients into a single pressure variable. TDI4 was also compared with predictive models developed using WA and response curve (RC) fitting (Birks et al. 1990) for each pressure.

These models (TDI.WA.PO4, TDI.RC.PO4 and so on) give an estimate of the best possible fit of the diatom data to the pressure gradient using a unimodal species response model, fitted using either WA or maximum likelihood RC modelling (Myung 2003). Theory and empirical studies show that WA is heuristically and computationally much simpler and can approximate the RC solution. However, it can suffer from truncation problems at the gradient ends in which high values are underestimated and vice versa. RC overcomes this limitation, but at the expense of higher computational burden, less interpretable species indicator values, and a tendency to extrapolate for poorly fitted samples. All P and N variables were $\log_{10}$ transformed before analysis.

Figure 2.4 shows the relationship between these models and pressure gradients. The TDI4 is plotted against all 3 pressure gradients (PO_4-P, NO_3-N and PC1) (top row of Figure 2.4); in addition, 3 variants of the WA model were derived, each optimised to a different pressure gradient (middle row of Figure 2.4), along with 3 variants of the RC model (bottom row of Figure 2.4). Table 2.2 gives the Pearson correlation coefficients for each of these relationships.

There are 3 immediate conclusions from these results.

1. The relationship between any model and P-PO_4 is relatively weak, and less strong than the corresponding correlation with NO_3-N. This is interpreted as being due to noise in the P-PO_4 data rather than P being less important for determining composition of diatom assemblages than NO_3-N. The relationship is particularly weak at low P values, where there are measurement and detection limit issues. The correlation with PC1, which integrates the N and P signals, is uniformly strongest. PC1 was therefore used to represent the nutrient pressure gradient in all subsequent analyses.

2. Several of the relationships between models and pressures exhibit some non-linearity. If the models were to be used to develop predictive pressure–response relationships, then using a linear model to encapsulate the relationship would not be appropriate. In this case, however, the models all measure species turnover along a gradient for which a linear fit is appropriate. The non-linearity is informative insofar as it indicates points along the gradient where the model may be less sensitive to changes in pressure, but this does not compromise the quality of the model per se.

3. TDI4 performs slightly less well for all pressure variables than a corresponding model developed using WA or RC. Similarly, WA performs less well than RC for PC1, although the differences are small (0.75, 0.79 and 0.82 for correlations between PC1 and TDI, WA and RC respectively). The small improvement of RC over WA is offset by the lack of interpretability of the RC model; although it is possible to derive species optima for this model, they can describe fitted curves that lie outside the gradient ends and are therefore not necessarily ecologically plausible. Given these problems, RC was discounted as a candidate for an improved TDI4.

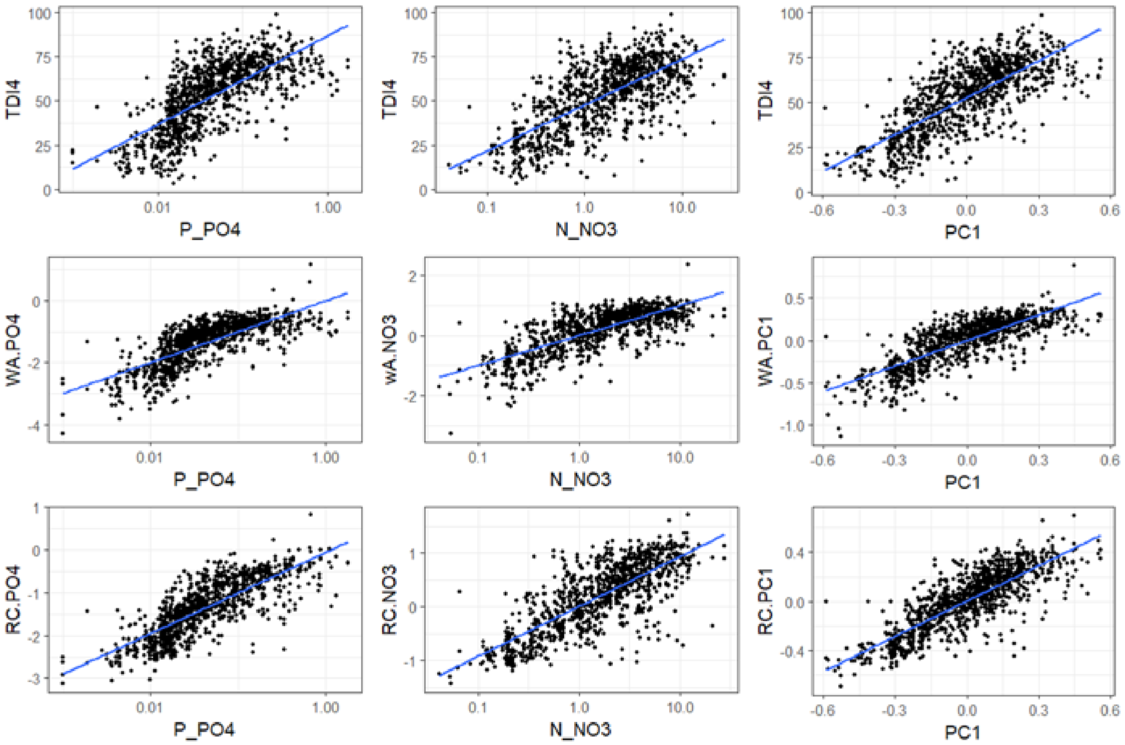

Figure 2.4 Relationship between TDI4 and the 3 nutrient pressure variables (top row), between a WA model and the pressure variables (middle row) and between an RC model and the pressure variables (bottom row)

Table 2.2 Pearson product–moment correlation coefficients (r) for the relationships shown in Figure 2.4

	TDI4	WA.PO4	WA.NO3	WA.PC	RC.PO4	RC.NO3	RC.PC
P-PO$_4$	0.36	0.41	0.36	0.39	0.49	0.42	0.46
N-NO$_3$	0.46	0.49	0.49	0.49	0.53	0.57	0.55
PC1	0.75	0.79	0.78	0.79	0.81	0.80	0.82

WA produces slightly stronger models than TDI4. This suggests that WA could be a candidate for a new TDI. However, WA derives species indicator values that are on a continuous scale relating to the pressure gradient. This means that the simple heuristic five-fold species indicator values of TDI4 are lost. Moreover, WA indicator values can only be calculated for taxa in the current dataset. It is not obvious how expert knowledge could be included in the metric or how new taxa could be added (without recalculating the metric with an expanded database at a future data). TDI4 performs only slightly less well than WA, indicating that TDI4 indicator values (which combine empirical distributions and expert knowledge) are broadly accurate and capture the main patterns of species variation along the nutrient pressure gradient.

2.4 Improving TDI4 to give TDI5LM

Further validation of the TDI4 was obtained by plotting TDI4 indicator values against the WA indicator values (that is, the so-called WA optima) of a model for the PC1 nutrient pressure gradient (Figure 2.5). While most taxa have TDI4 indicator values that are integers, a few have indicator values that are decimals. These arise from an earlier Environment Agency funded project in which groups of taxa that proved challenging to analysts were amalgamated into categories that were given the weighted mean sensitivity of the constituent species (Environment Agency 2012).

There was a generally strong relationship between TDI4 indicator values and species optima derived from a WA analysis with nutrient pressure PC1 (r = 0.69). There were, however, some misclassified taxa (for example, *Nitzschia brevissima* NI073A). A simple iterative algorithm was applied in which taxa that had a calculated WA optimum more or less than 0.2 units outside the median for a TDI4 class were re-allocated to an adjacent TDI4 class. This re-allocation of taxa was verified using expert judgement and repeated until there were no further obviously misclassified taxa. TDI indicator values after re-allocation are shown in Figure 2.6. The correlation of the new TDI scores with WA optima is somewhat improved (r = 0.87). The revised TDI is referred to as TDI5LM.

The correlation between TDI4 and TDI5LM is 0.99 and Lin's concordance correlation coefficient is 0.99. The difference between TDI4 and TDI5LM (TDI4 – TDI5LM) is small for most samples (Figure 2.7). Samples with a larger difference are explained by a dominance of taxa that have revised indicator values. Examples include *Diatoma vulgare* 5 -> 4, *Encyonopsis microcephela* 2 -> 1 and *Gomphonema 'intricatum'* 3.6 -> 2; *Epithemia adnata* 5 -> 2.

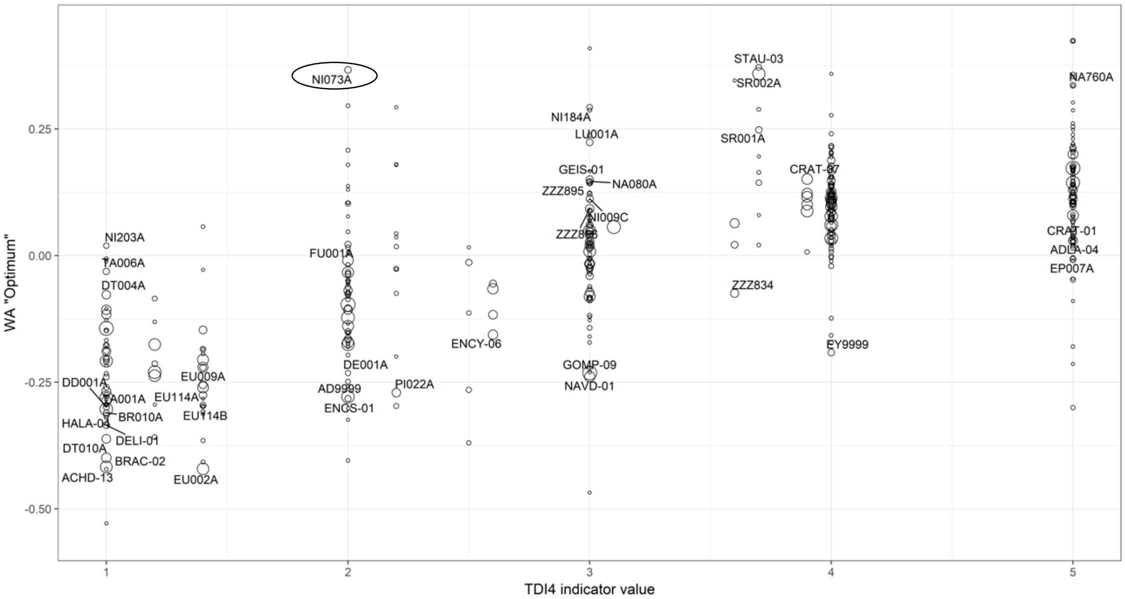

Figure 2.5 TDI4 indicator values plotted against WA optima derived for the PC1 nutrient pressure gradient

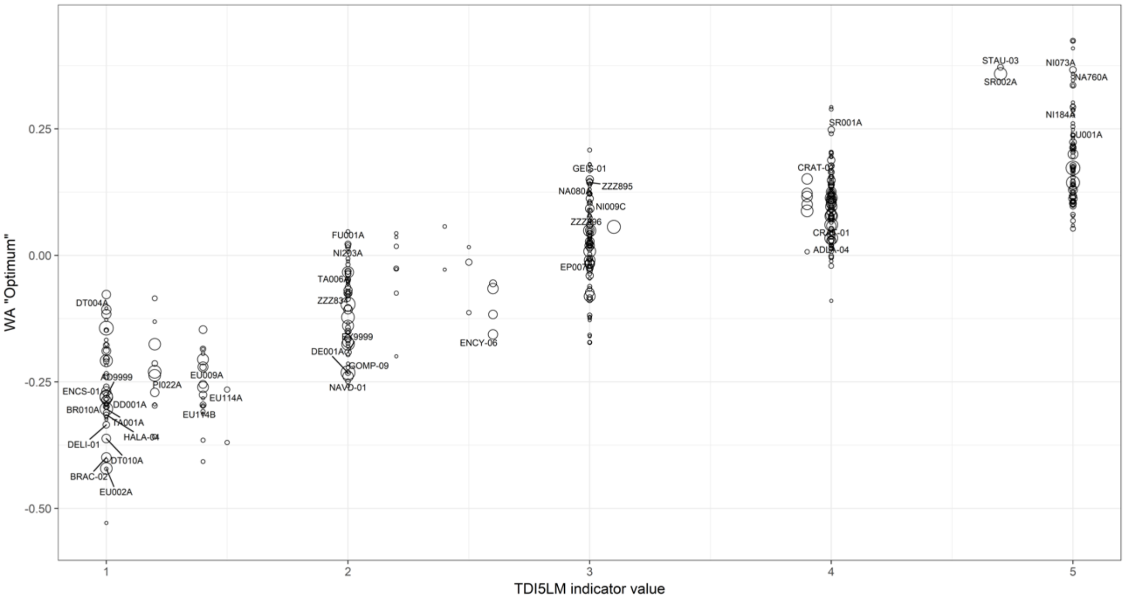

Figure 2.6 Revised ('TDI5LM') indicator values plotted against WA optima derived for the PC1 nutrient pressure gradient

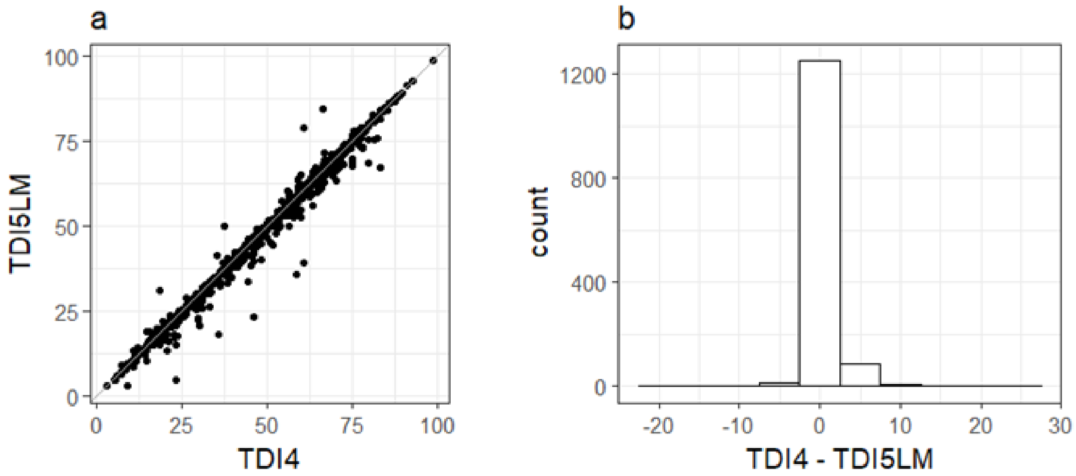

Figure 2.7 Comparison between TDI4 and TDI5LM values computed on the LM dataset (Section 2.2) showing: (a) scatter plot of TDI4 versus TDI5LM scores (diagonal line shows slope = 1); and (b) histogram of differences (that is, TDI4 –TDI5LM)

These changes to species indicator values had no effect on the relationship with the PC1 pressure gradient (Figure 2.8). There are, as a result, only small implications for classification and relationship with pressure when using the present DARLEQ2 reference model, with a small number of sites moved into a higher quality class under TDI5LM.

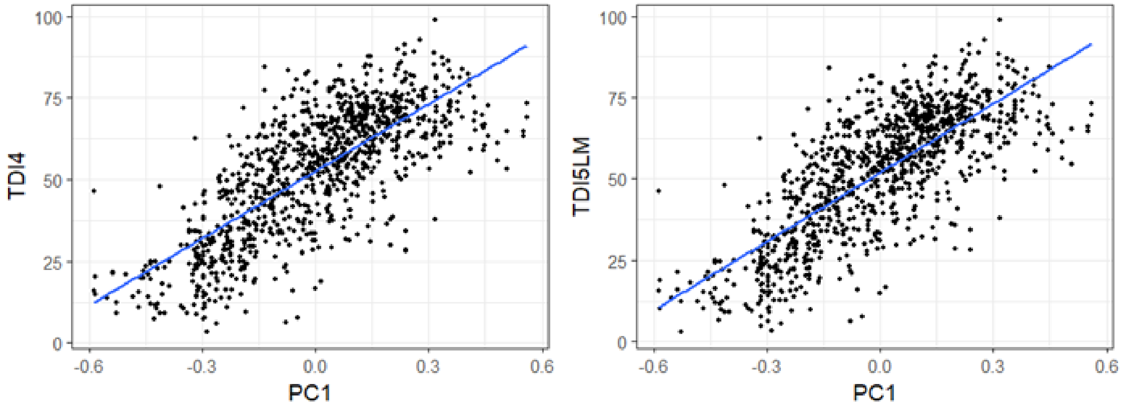

Figure 2.8 Relationship between TDI4 (a) and TDI5LM (b) and PC1

Notes: The correlations for these relationships are 0.75 and 0.76 respectively.

In Table 2.3 and subsequent classification tables in this report, ‘agreement’ refers to the percentage of sites that are classified into the same class using both methods and ‘bias’ refers to the tendency for one method to produce more stringent classifications than the other. This is calculated as the difference between the proportions of sites that are classified more stringently by method A (‘columns’) minus the proportion classified more stringently by method B (‘rows’). In the example shown in Table 2.3, 32 (9 + 12 + 11) sites (4.8%) in the left hand table are classified more stringently by TDI4 while 6 (5 + 1) (0.9%) are classified more stringently by TDI5LM. Therefore the bias is -3.9% for the full dataset. For sites where alkalinity is <120mgL^{-1} $CaCO_3$ (right hand table), the bias is -5.3%.

A separate analysis is given for sites with alkalinity <120mgL^{-1} $CaCO_3$ as the Environment Agency does not currently use diatoms for status assessment for rivers with higher alkalinities due to weaknesses with the DARLEQ2 reference model (see Section 2.7.1).

Table 2.3 Comparison between WFD ecological status classes for sites computed by TDI4 (rows) and TDI5LM (columns) using the current reference model

		TDI5LM									
		All sites					<120mgL^{-1} $CaCO_3$ only				
		Bad	Poor	Mod	Good	High	Bad	Poor	Mod	Good	High
TDI4	Bad	1	0	0	0	0	0	0	0	0	0
	Poor	0	22	9	0	0	0	13	7	0	0
	Mod	0	0	177	12	0	0	0	133	10	0
	Good	0	0	5	218	11	0	0	1	167	11
	High	0	0	0	1	208	0	0	0	1	149

Notes: Green shading: identical classification for both LM and NGS; yellow shading: agreement to within one class between LM and NGS

Left hand table shows all sites (N = 664): agreement: 94%; bias: -3.9%.
Right hand table shows sites <120mgL^{-1} $CaCO_3$ (N = 492) subset: agreement: 94%; bias: -5.3%.

In most cases, 2 samples per site are available; however, there are a few sites (for example, those used in Sections 7 and 8 of Environment Agency 2018) where the number of samples per site is higher and also some where only a single sample is available.

Taxa with revised indicator values include *Adlafia suchlandtii*, which was responsible for some anomalous classifications in the past. The revised model is therefore more robust and site-level predictions should be better.

2.5 Derivation of new NGS metric (TDI5NGS)

Having optimised the LM metric, the next step was to use this to derive a new NGS metric (TDI5NGS). For operational reasons, TDI5NGS was designed to mimic light microscopy TDI scores as closely as possible. As in project SC140024, this was done using a WA algorithm to derive NGS taxon indicator values[1] that best predicted LM TDI values. In the earlier study, TDI4 was used for the LM values; in this study, TDI5LM was used.

A further modification to the procedure used in project SC140024 was introduced to avoid the possibility of negative taxon indicator scores for taxa associated with very low levels of pressure. These were originally set to 1.0 and, similarly, a few taxa that had scores >5 were set to 5.0. The modification to the method uses a narrower range of taxon indicator values along with a non-linear rescaling of the TDI5NGS values against the original TDI5LM values using a monotonic generalised additive model (GAM). This ensures that the range of TDI values lies within the range 0-100 and gives a better fit of TDI5NGS to TDI5LM at lower TDI values than the original method.

A complete description of the stepwise method used to derive NGS indicator values and sample predictions is as follows.

1. WA regression of the NGS species assemblage data and TDI5LM sample values is used to calculate NGS species indicator values that best predict the TDILM data. These species indicator values represent weighted centroids or ‘optima’ of NGS taxa along the TDI5LM gradient.

2. WA regression is known to shrink the range of optima compared with the range of the target gradient (TDI5LM in this case) and so the species indicator values were expanded using a deshrinking regression of TDI5NGS sample scores on TDI5LM sample scores. This is a usual and necessary step in WA regression and calibration (see, for example, Birks et al. 1990). The final TDI5NGS scores are listed in Appendix 2.

3. WA calibration is used to predict TDI5NGS sample scores from the NGS species indicator values and assemblage data. Again, WA tends to

[1] A species or taxon indicator value represents its sensitivity across the nutrient gradient (1 = sensitive; 5 = tolerant).

shrink the range of predictions relative to the range of the target gradient. This shrinkage is more pronounced at the gradient ends and so non-linear deshrinking using a monotonic smoothing spline fitted using a GAM (Birks and Simpson 2013) was used to deshrink the original TDI5NGS sample scores to the range of TDI5LM scores. The monotonic regression removes the edge effects inherent in WA calibration, but tends to overestimate values at the low end and overestimate values at the high end of the TDI gradient. A final linear deshrinking was therefore performed using major axis regression of TDI5NGS scores on TDI5LM scores to optimise TDI5NGS sample scores and avoid under/over prediction at the gradient ends.

R code (R Development Core Team 2017) that implements the above algorithm is available in the R package darleq3 at https://github.com/nsj3/darleq3.

Figure 2.9a shows the relationship between TDI5LM and TDI5NGS for the original method (Environment Agency 2018, Section 6), while Figure 2.9b shows it using the new rescaling procedure. Both plots are based on the relative abundances of taxa in NGS output without transformation or taxon downweighting. The monotonic GAM rescaling procedure results in a neater fit to TDI5LM. There is still a tendency to slightly overestimate at low TDI values and underestimate at high values, however, and the overall improvement in correlation and concordance is small (Table 2.4). The correlation between TDI5LM and TDI5NGS.original (Figure 2.9a) is similar to that between TDI4 and the prototype NGS metric using data collected in 2014 (reported in Environment Agency 2018), despite the larger dataset and barcode database used in the present study.

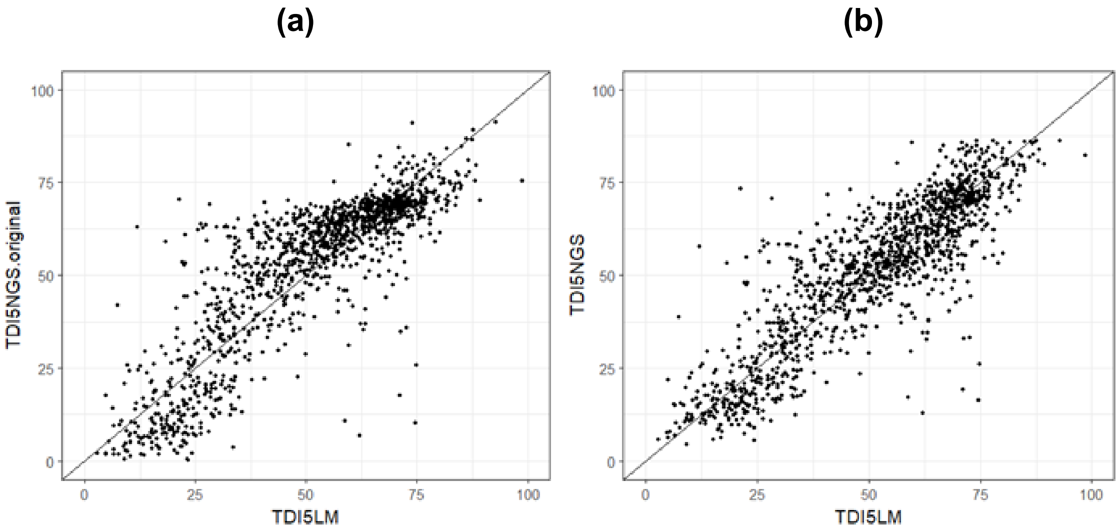

Figure 2.9 Relationship between TDI5LM and TDI5NGS for variants without (a) and with (b) non-linear rescaling

Table 2.4 Correlations between TDI5LM and TDI5NGS variants (with and without non-linear rescaling)

	TDI5NGS without rescaling	TDI5NGS with rescaling
Pearson's correlation	0.854	0.873

Lin's concordance	0.851	0.873

Subsequent analyses use the variant of TDI5NGS with monotonic GAM rescaling. This metric had a slightly weaker relationship to the nutrient pressure gradient than TDI5LM (Figure 2.10 and Table 2.5). TDI5LM, in turn, had a very similar relationship with this gradient to TDI4 – see Table 2.2). This raises a question about whether the weaker relationship observed for TDI5NGS is a function of the way TDI5NGS was derived or is an intrinsic feature of the NGS data.

This was addressed by generating WA and RC models for NGS in the same way as for LM data in Section 3 (Figure 2.11). There was, however, little difference in the correlations with PC1 for the 3 methods (Table 2.6) and all are lower than the corresponding relationships with LM data (Table 2.2). WA has a slightly higher correlation with PC1, but exhibits a lack of sensitivity at the high end of the gradient. RC better differentiates samples when nutrient pressures are high, but there is a 'gap' around 0.25 units (right hand figure in top row of Figure 2.11) which is currently unexplained.

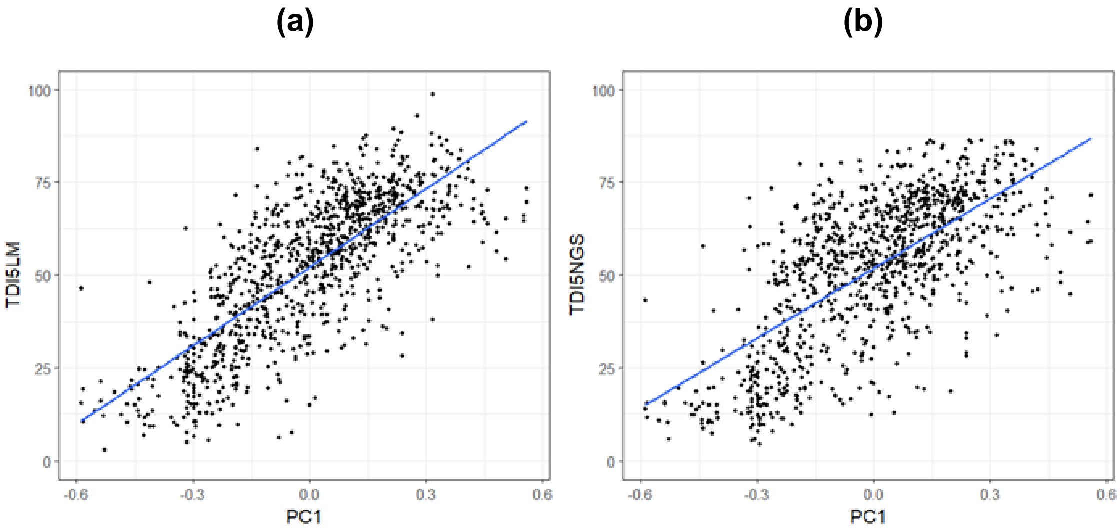

Figure 2.10 Relationship between TDI5LM (a) and TDI5NGS (b) and the nutrient pressure gradient (PC1)

Table 2.5 Pearson correlation coefficients for the relationship between TDI5LM and TDI5NGS and the nutrient pressure gradient, PC1

	TDI5LM	TDI5NGS
PC1	0.76	0.673

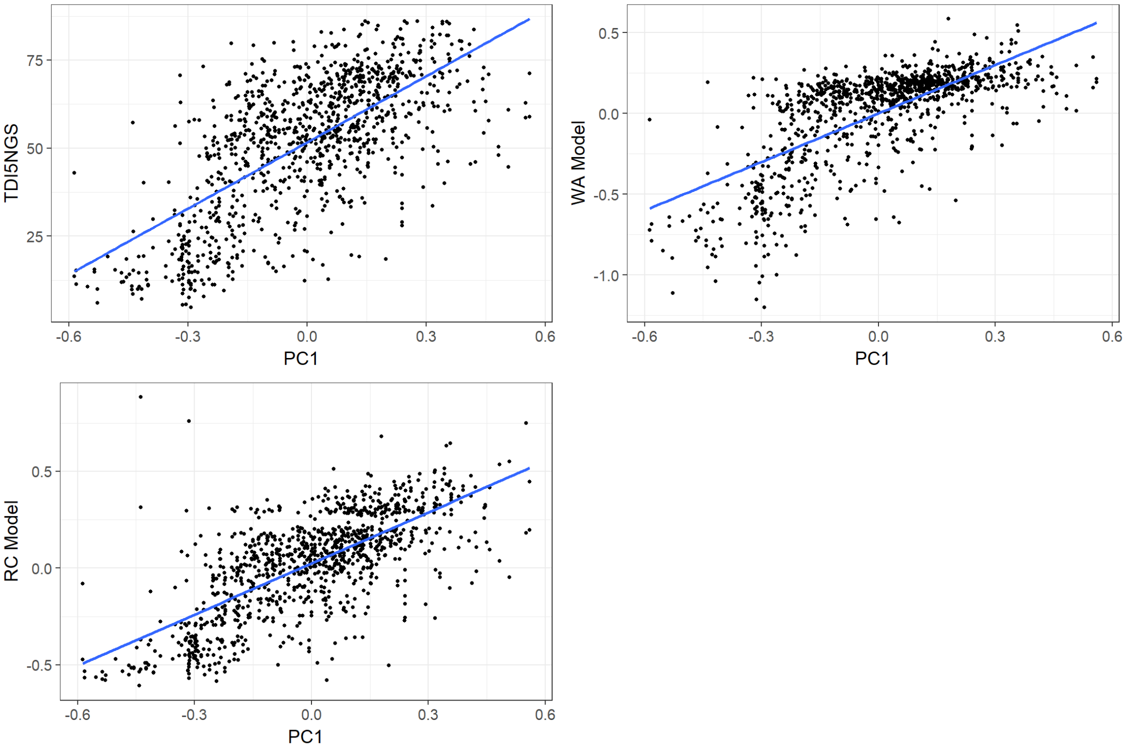

Figure 2.11 Relationships between 3 NGS based models and the PC1 combined pressure gradient with no differential taxon weights. Upper left: TDI5NGS derived as above. Upper right: WA 'optimised' model. Lower left: RC optimised model

Table 2.6 Pearson correlation coefficients for the models shown in Figure 2.11, with and without upweighting and downweighting of key taxa

	TDI5NGS	**TDI5NGS.WA.PC**	**TDI5NGS.RC.PC**
No weighting	0.673	0.692	0.700
With taxon up/down weighting	0.682	0.701	0.713

Despite these differences between methods, the results suggest that the differences between the LM and NGS relationships with nutrient pressure are not due to issues with the methodology of deriving NGS species indicator values from TDI5LM, but instead due to characteristics of the NGS data themselves.

Substantial differences in the quantification of some taxa between LM and NGS were observed in project SC140024. The number of chloroplast per cell was proposed as a partial explanation for this effect, although a range of factors is likely to be involved. This issue was addressed in SC140024 by overriding the taxon indicator values for a few quantitatively important taxa. In the prototype NGS metric, *Navicula lanceolata* and *Melosira varians* were downweighted (×0.5) and *Achnanthidium minutissimum* was upweighted (×1.5). When these adjustments are applied to TDI5NGS, there was a slightly higher correlation with the pressure gradient for all 3 models but, again, the improvement with WA and RC methods was small (Table 2.6).

2.5.1 Improving TDI5NGS

The original motivation to differentially weight taxa was based on the under- and over-representation of some common taxa in the NGS dataset. The new rescaling described in Section 2.4 yields only a very small improvement over the non-downweighted NGS model when using taxon weightings based on expert judgement. However, a further attempt to optimise taxon weightings to improve the fit between LM and NGS TDI scores was made by calculating:

- the mean difference in relative abundance of each taxon in the LM and NGS data (that is, TDILM – TDINGS)
- the variance of the differences in abundance

This offers some insights into how the LM and NGS metrics differ (Table 2.7).

As expected, *Achnanthidium minutissimum* is consistently underestimated and *Melosira varians*, *Navicula lanceolata* and others are overestimated in the NGS dataset. In project SC140024, arbitrary weights were applied to taxa that had large differences in relative abundance between LM and NGS data. This time, a constrained optimisation procedure was used to derive taxon weights that give the best fit to the LM data (Byrd et al. 1995).

First, the optimisation routine was run for individual taxa with high variance and/or a high mean difference. The optimisation was then run on a set of taxa that had weights significantly different from zero in the original run. Six taxa met that criterion (Table 2.8). This included the 3 taxa used in the prototype NGS metric along with *Ulnaria ulna* and *Diatoma vulgare*, both of which had low weights (that is, consistently over-represented in NGS data) and *Tabellaria flocculosa*, which had a high weight (that is, under-represented) in the NGS dataset. The latter is of interest as this species has multiple chloroplasts, a condition hitherto regarded as being associated with over-representation in NGS relative to LM.

Table 2.7 Differences in relative abundances of taxa in LM and NGS datasets

Taxon ID	Taxon	Difference	Variance	LM.N	LM.N2	LM.Max	NGS.N	NGS.N2	NGS.Max
ZZZ835	*Achnanthidium minutissimum* type	14.38	292.55	1,297	628.0	96.9	1,285	210.3	75.3
ME015A	*Melosira varians*	-11.35	305.85	484	107.7	62.8	1,278	399.4	98.8
NA009A	*Navicula lanceolata*	-9.96	189.22	1,026	341.4	74.4	1,324	537.2	99.4
ACHD-02	*Achnanthidium pyrenaicum*	-3.29	79.83	248	60.6	72.9	1,255	269.0	70.5
AMPH-05	*Amphora pediculus* type	3.23	75.26	927	330.1	78.3	1,115	186.6	70.9
ULNA-02	*Ulnaria ulna*	-2.60	51.26	444	68.1	48.5	1,173	162.6	84.0
NA023A	*Navicula gregaria*	2.55	44.85	1,085	445.4	68.7	1,233	377.2	57.7
FIST-01	*Fistulifera saprophila*	-2.11	35.72	130	31.6	45.6	1,110	180.9	70.1
DIAT-01	*Diatoma vulgare* agg.	-2.04	45.50	269	50.0	48.6	812	119.0	86.0
FR009A	*Fragilaria capucina*	1.67	29.05	598	133.0	69.5	475	62.8	12.3
ZZZ896	*Planothidium frequentissimum*	1.48	8.15	774	286.5	35.3	734	77.1	12.0
NI015A	*Nitzschia dissipata*	1.45	9.74	827	256.8	37.3	982	190.6	4.9
CO005A	*Cocconeis pediculus*	-1.26	24.47	448	151.3	25.1	995	147.3	59.4
SU073A	*Surirella brebissonii*	-1.12	17.63	701	249.2	24.1	1,125	206.1	59.7
NI025A	*Nitzschia recta*	-1.04	7.43	220	92.9	7.0	1,139	213.9	44.4
HN001A	*Hannaea arcus*	-1.03	27.81	231	49.8	71.8	1,001	92.4	66.1
GO052A	*Gomphonema olivaceoides*	1.01	13.44	338	93.8	48.0	0	0	0.0

FR007A	*Fragilaria vaucheriae*	0.97	5.45	656	225.2	26.4	909	145.4	6.5
EOLI-01	*Eolimna minima*	0.90	17.06	777	166.6	66.8	1,024	93.0	61.1
RC002A	*Rhoicosphenia abbreviata*	0.90	19.03	736	254.7	50.0	904	155.4	49.8

Notes: ‘Difference’ is the mean difference in relative abundance between NGS and LM datasets. Positive differences indicate greater abundance in LM and vice versa.

Table 2.8 Taxa with high mean difference and/or variance selected by optimisation routine, along with assigned weights

Taxon ID	Taxon	Assigned weight
ME015A	*Melosira varians*	0.1
NA009A	*Navicula lanceolata*	0.2
TA001A	*Tabellaria flocculosa*	1.6
ZZZ835	*Achnanthidium minutissimum* type	2.3
ULNA-02	*Ulnaria ulna*	0.1
DIAT-01	*Diatoma vulgare* agg.	0.1

These weights were then applied to the relative abundance of the relevant taxa and used to derive a (second version of the) downweighted TDI5NGS metric (Figure 2.12).

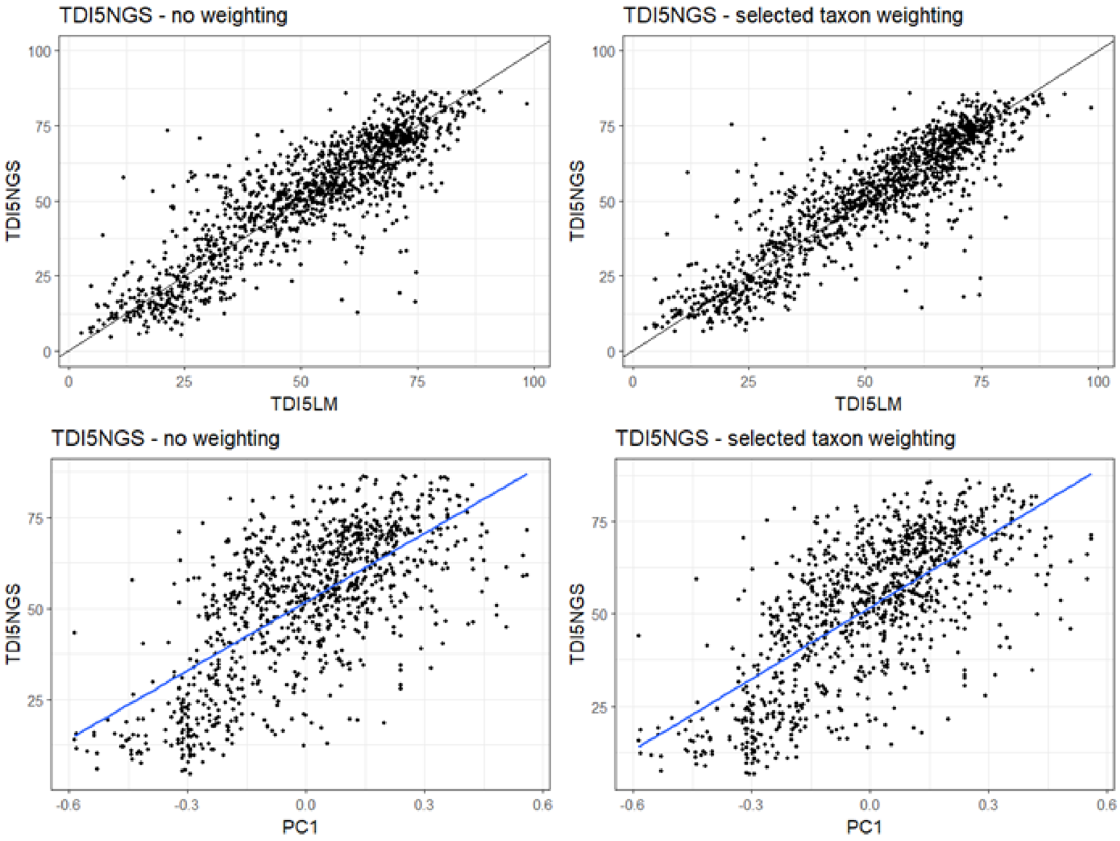

Figure 2.12 Effect of downweighting on TDI5NGS. Left hand images show TDI5NGS without weights. Right hand images show TDI5NGS with weights

Notes: TDI5NGS is plotted on the y axis against TDI5LM scores (top row) and PC1 pressure gradient (bottom row).

Correlation coefficients obtained using the revised downweighted TDI5NGS metric against TDI5LM and the PC1 pressure gradient are slightly higher than the non-downweighted version (Table 2.9). However, the improvements are small and there is no marked change in the overall fit.

Table 2.9 Pearson correlation coefficients for relationships in Figure 2.12

	TDI5LM	PC1
No taxon weightings	0.873	0.673
With taxon up/downweighting	0.889	0.696

The issue of weighting needs further work. It is possible that further optimisation of the weights may yield a higher correlation with TDI5LM, but a major improvement is unlikely. There are arguments for (for example, to take account of over-representation in NGS data because of multi-chloroplasts), but also arguments against (for example, *a* certain taxa may be under-represented in NGS because coverage of different barcode sequences (genotypes) is low – this could be addressed in future making upweighting of the taxon unnecessary).

It is also possible that rbcL reads representing a major photosynthetic enzyme is a more direct measure of the potential productivity of each taxon than simply the number of cells. In general, the downweighted taxa are larger than those that are upweighted. It is important to emphasise that NGS data and LM data are fundamentally different in nature and, unless there are good empirical reasons for weighting, the NGS data should not be massaged simply to fit preconceptions formed from long experience of LM data.

As the weights add an extra step to the TDINGS calculations and the improvement in the model is offset by the loss of parsimony, the use of the unweighted TDI5NGS is recommended. This is because use of weighting needs more work to understand the reasons and consequences.

2.5.2 Model performance

The TDI5NGS metric (unweighted) described above uses the full dataset to:

- derive NGS taxon indicator values and the non-linear deshrinking model
- assess model performance via correlation with TDI5LM

As such, the model will be optimised to the combined 2014 and 2016 dataset. A five-fold cross-validation was therefore used to test the robustness of this relationship and its likely performance when confronted with new NGS data. In this, the dataset was split at random into 5 equal-sized fractions and the model developed, or trained, on four-fifths of the data. The left-out one-fifth was used to test the model. This process was repeated 5 times for each left-out group, and the TDI5NGS scores aggregated across the 5 test groups.

Under five-fold cross-validation, the correlation between TDI5LM and TDI5NGS is only marginally lower that for non-cross-validated model (Table 2.10). This indicates that:

- the model is robust
- the correlation cited above between LM and NGS methods is a good guide to the expected agreement between the 2 methods when applied to new data

Table 2.10 Correlation between TDI5LM and TDI5NGS using all data and a cross-validated model

	All data	Cross-validated model
Pearson's correlation	0.87	0.86

2.6 Implications for classification

Tables 2.11 and 2.12 show the effect on classifications of adopting TDI5LM and TDI5NGS respectively.

About two-thirds of the sites are classified as high or good status, probably reflecting issues with the reference model (see below). Overall, however, the bias between ecological status calculated with the current metric (TDI4) and that with the recalibrated TDI5NGS metric fell from 21% for the prototype NGS metric (Environment Agency 2018) to <3% (Table 2.11). Similar levels of bias are observed when the TDI5NGS is compared to the optimised LM metric ('TDI5LM'; Table 2.12) with 64% of samples classifying in the same status class, 31% within one status class and 4% greater than one class. TDI5NGS has a tendency to be slightly more stringent than TDI5LM.

Table 2.11 WFD ecological status classes for sites classified using TDI4 (rows) and TDI5NGS (columns)

		TDI5NGS									
		All sites					**<120mgL^{-1} CaCO$_3$ only**				
		Bad	**Poor**	**Mod**	**Good**	**High**	**Bad**	**Poor**	**Mod**	**Good**	**High**
TDI4	**Bad**	0	1	0	0	0	0	0	0	0	0
	Poor	0	9	16	5	1	0	5	12	2	1
	Mod	0	11	126	49	5	0	8	101	34	2
	Good	0	0	47	135	53	0	0	37	106	38
	High	0	0	6	51	151	0	0	3	36	109

Notes: Green shading: identical classification for both LM and NGS; yellow shading: agreement to within one class between LM and NGS; red shading: greater than one class difference between methods.

Left hand table shows all sites (N = 666): agreement: 63%; bias: -2.3%.
Right hand table shows sites <120mgL^{-1} $CaCO_3$ subset (N = 494): agreement: 65%; bias: -1.0%

Table 2.12 WFD ecological status classes for sites classified using TDI5LM (rows) and TDI5NGS (columns)

		TDI5NGS									
		All sites					<120mgL^{-1} CaCO$_3$ only				
		Bad	Poor	Mod	Good	High	Bad	Poor	Mod	Good	High
TDI5LM	Bad	0	1	0	0	0	0	0	0	0	0
	Poor	0	7	11	2	1	0	3	8	1	1
	Mod	0	12	129	48	4	0	9	103	29	2
	Good	0	1	47	137	49	0	1	37	109	33
	High	0	0	8	53	156	0	0	5	39	114

Notes: Green shading: identical classification for both LM and NGS; yellow shading: agreement to within one class between LM and NGS; red shading: greater than one class difference between methods.
Left hand table shows all sites (N = 666): agreement: 64%; bias: 0.8%.
Right hand table shows sites <120mgL^{-1} CaCO$_3$ subset (N = 494): agreement: 67%; bias: 3.2%

2.7 Ecological Quality Ratios and use of an alternative reference model

As explained in Box 1 of the SC140024 report (Environment Agency 2018, p. 1), the WFD requires that the condition of water bodies is expressed as a ratio – the Ecological Quality Ratio (EQR) – using the value of the biological parameter 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 DARLEQ, which calculates the EQR as the observed TDI divided by the expected TDI for any lake or river; the current version of the tool is DARLEQ2.

All the steps up to this point focused on optimising the raw metric whether for use with LM or NGS data. However, the final calculations that show the effect on classifications assumed that the reference model that is applied is the same as that used at present. In light of concerns about the efficacy of this reference model and some parallel work to consider alternatives, this section examines the consequences for classification if the switch to either the LM or NGS variants of TDI5 is accompanied by a change in the reference model.

2.7.1 TDI5LM reference model

The alternative reference model was derived by quantile regression analysis of a separate dataset by Geoff Phillips[2] and is referred to subsequently in this report as the 'GP model' (Equation 2.1).

$$\text{Expected TDI} = 9.3502 + (-3.2504 \times \text{Alkalinity}^3) + (12.877 \times \text{Alkalinity}^2) + (3.3573 \times \text{Alkalinity}) \qquad (2.1)$$

Note that this reference model has not been adopted by UK agencies. It is included here simply as a demonstration of the scale of effect that may be expected if the reference model were to be changed.

Equation 2.1 predicts expected TDI values that are lower than those used at present at high alkalinity but which are higher than those used in the original version of DARLEQ (Kelly et al. 2008) (Figure 2.13). Moreover, the current (DARLEQ2) reference model approximates to the median of the full dataset; this suggests, counter-intuitively, that a very large number of high alkalinity sites are exceeding expectations. This issue led to the decision not to use DARLEQ2 for classification purposes when alkalinity was >120mgL^{-1} $CaCO_3$. In contrast, the alternative reference model, GP model, does appear to follow a lower quantile of the data cloud, meaning that relatively few sites will exceed expected values.

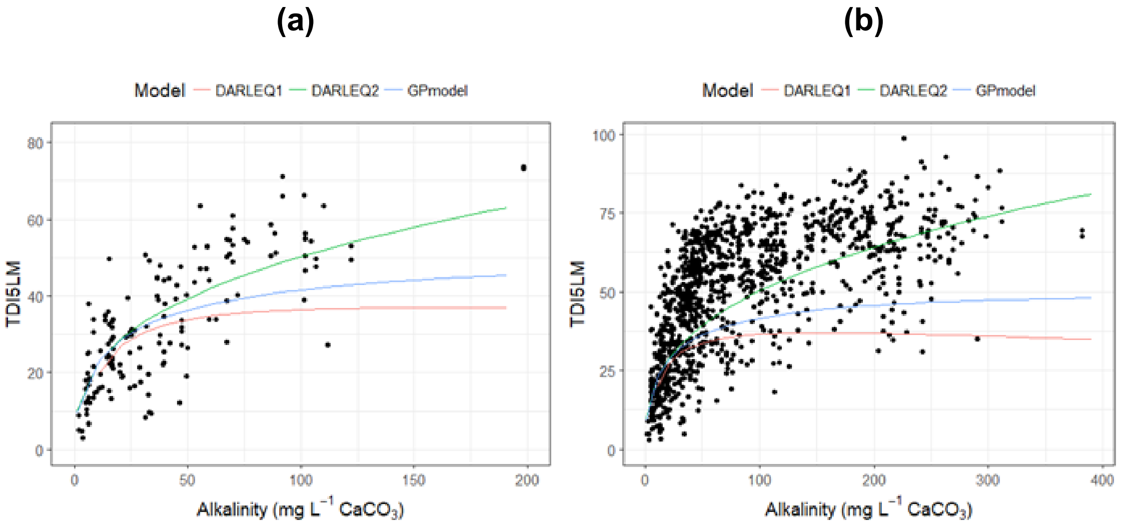

Figure 2.13 Reference dataset from SC140024[1] (a) and the full 2014 and 2016 dataset (b) with the original DARLEQ and current DARLEQ2 reference models and the new reference model (GP model)

Notes: [1] See Environment Agency (2018, Figure 6.11)

As efforts until now have been made to obtain close correspondence between LM and NGS variants of the TDI, it has been possible to assume that status

[2] Retired Environment Agency research scientist and now an honorary professor at the Department of Biological and Environmental Sciencies, Univeristiy of Stirling. The reference model was derived from analyses performed as part of another Environment Agency-funded project and has, subsequently, been further refined.

class boundaries will remain unchanged. If a new reference model is adopted, however, this assumption would not hold and new status class boundaries will need to be derived.

Figure 2.14 shows the distribution of sensitive taxa (Group 1: indicator values ≤2) and tolerant taxa (Group 2: indicator values ≥4) along an EQR gradient calculated using TDI5LM and the GP model. The crossover of these 2 groups (shown in the figure as monotonic RCs fitted using logistic regression) was originally used to set the boundary between 'good' to 'moderate' ecological status under the WFD (Kelly et al. 2008).

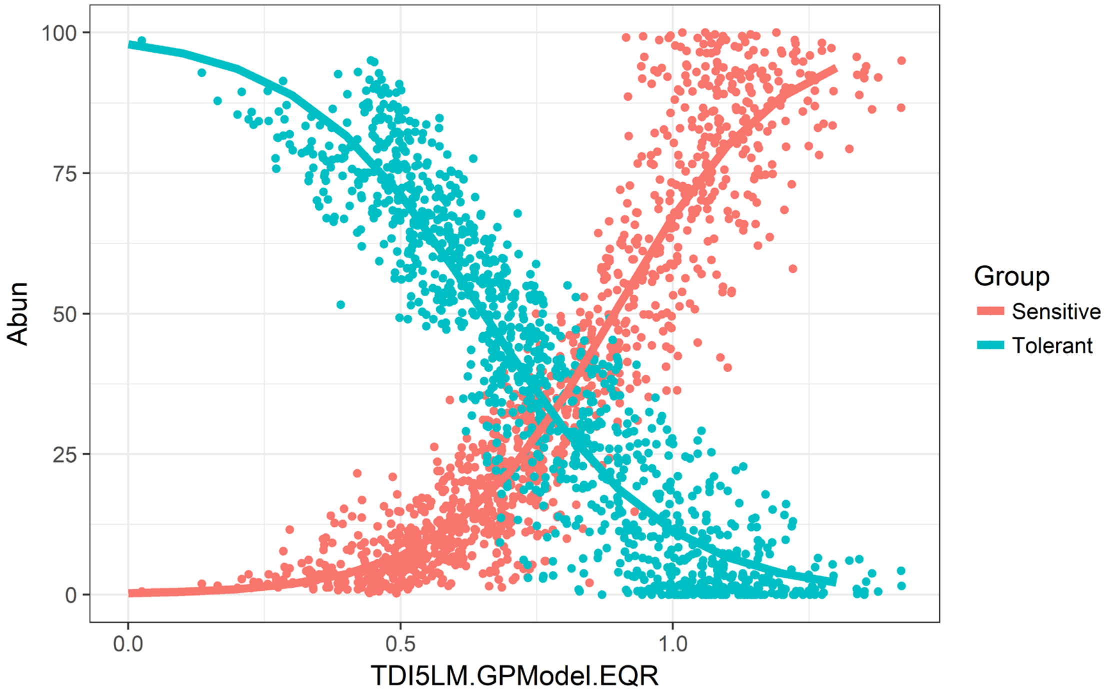

Figure 2.14 Distribution of sensitive (Group 1) and tolerant (Group 2) taxa along the EQR gradient using the GP model for TDI5LM data

Moderate to poor and poor to bad ecological status classes were derived by dividing the remaining scale into 3 portions. This yielded the following boundaries:

- high to good = 0.9
- good to moderate = 0.8
- moderate to poor = 0.5
- poor to bad = 0.25

Tables 2.13 and 2.14 show the implications for classification. Overall, there is a high level (~30%) of bias between ecological status class using the current LM DARLEQ2 metric and that based on the new GP reference model, the latter being much more stringent. Figure 2.15 shows this graphically, with a smaller proportion of sites classified at 'high' ecological status using the GP model compared with the current DARLEQ2 model.

Table 2.13 WFD ecological status classes for sites classified using TDI4 based on DARLEQ2 reference model (rows) and TDI4 based on GP model (columns)

		TDI4 / GP model									
		All sites					<120mgL^{-1} CaCO$_3$ only				
		Bad	**Poor**	**Mod**	**Good**	**High**	**Bad**	**Poor**	**Mod**	**Good**	**High**
TDI4 / DARLEQ2	**Bad**	1	0	0	0	0	0	0	0	0	0
	Poor	3	29	0	0	0	0	21	0	0	0
	Mod	0	64	154	0	0	0	24	147	0	0
	Good	0	9	123	54	50	0	0	77	54	50
	High	0	0	31	15	131	0	0	0	3	116

Notes: Green shading: identical classification for both LM and NGS; yellow shading: agreement to within one class between LM and NGS; red shading: greater than one class difference between methods.
Left hand table shows all data (N = 666): agreement: 56%; bias: 29.3%.
Right hand table shows sites <120mgL^{-1} CaCO$_3$ subset (N = 494): agreement: 67%; bias: 11%

Table 2.14 WFD ecological status classes for sites classified using TDI5LM based on DARLEQ2 reference model (rows) and GP model (columns)

		TDI5LM / GP model									
		All sites					<120mgL^{-1} CaCO$_3$ only				
		Bad	**Poor**	**Mod**	**Good**	**High**	**Bad**	**Poor**	**Mod**	**Good**	**High**
TDI5LM / DARLEQ2	**Bad**	1	0	0	0	0	0	0	0	0	0
	Poor	2	19	0	0	0	0	13	0	0	0
	Mod	0	64	129	0	0	0	21	122	0	0
	Good	0	4	132	57	41	0	0	82	57	41
	High	0	0	27	15	175	0	0	0	2	156

Notes: Green shading: identical classification for both LM and NGS; yellow shading: agreement to within one class between LM and NGS; red shading: greater than one class difference between methods.
Left hand table shows all data (N = 666): agreement: 57%; bias: 31%.
Right hand table shows sites <120mgL^{-1} CaCO$_3$ subset (N = 494): agreement: 70%; bias: 13%

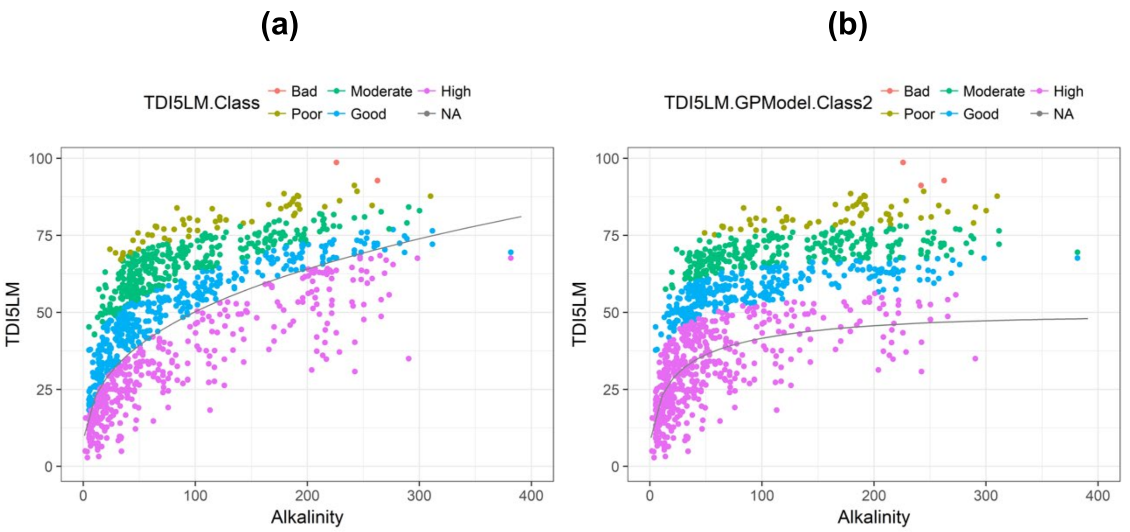

Figure 2.15 TDI5LM samples versus alkalinity with: (a) DARLEQ2 reference model and status classes; and (b) GP model and status classes

2.7.2 TDI5NGS reference model

The exercise was repeated using TDI5NGS rather than TDI5LM as the base metric. The crossover is taken as 0.9 (Figure 2.16), and other ecological status class boundaries set at 0.7, 0.5 and 0.25. A similar level of bias is seen using the NGS metric as with the LM metric (Table 2.15).

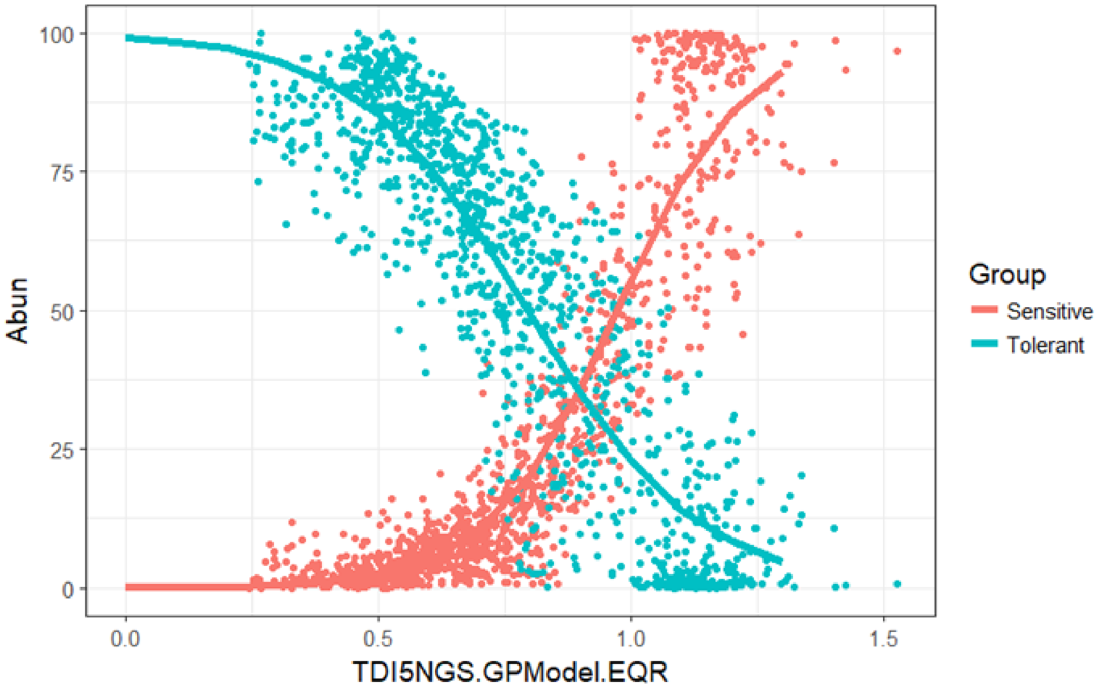

Figure 2.16 Distribution of sensitive (Group 1) and tolerant (Group 2) taxa along the EQR gradient using the GP reference model for TDI5NGS data

Table 2.15 WFD ecological status classes for sites classified using TDI5NGS based on DARLEQ2 reference model (rows) and GP model (columns)

	TDI5NGS / GP model									
TDI5NGS / DARLEQ2	Bad	Poor	Mod	Good	High	Bad	Poor	Mod	Good	High
Bad	0	0	0	0	0	0	0	0	0	0
Poor	1	20	0	0	0	0	13	0	0	0
Mod	0	40	155	0	0	0	13	140	0	0
Good	0	4	144	64	28	0	0	86	64	28
High	0	0	24	21	165	0	0	0	7	143

Notes: Green shading: identical classification for both LM and NGS; yellow shading: agreement to within one class between LM and NGS; red shading: greater than one class difference between methods.
Left hand table shows all data (N = 666): agreement: 61%; bias: 30.9%.
Right hand table shows sites <$120mgL^{-1}$ $CaCO_3$ subset (N = 494): agreement: 73%; bias: 15.8%

These scenarios help to put the consequences of changing from the LM metric to the NGS metric into perspective. Although mismatches are likely to be encountered with any change of method, the effect of changing from LM to NGS is much smaller at <5% (see Tables 2.3, 2.11 and 2.12) than the potential consequences of changing the reference model (see Tables 2.13, 2.14 and 2.15).

Two further points to make at this stage are that:

- it was not possible to obtain an estimate of background ‘noise’ associated with classifications (that is, how many sites/water bodies change WFD ecological status class between classification periods without a plausible explanation)
- the discussion of ‘bias’ takes no account of whether the new method offers a more accurate insight into WFD ecological status

2.8 Conclusions

This section describes further development of the prototype NGS method described in Environment Agency (2018). The NGS metric has been recalibrated using a larger dataset, a bigger barcode database and an optimised LM metric. Options such as upweighting or downweighting taxa were evaluated but not adopted at this stage in favour of a parsimonious model that accepts that NGS data have fundamentally different properties to LM data. The new model has a similar statistical strength to the original TDI5NGS, but some of the changes incorporated will mean that individual sites are less likely to be misclassified. Method bias is lower than was reported for the prototype NGS method and is lower than that likely to accompany a change to the reference model.

3 Potential for cross-contamination of diatom DNA samples from toothbrushes

3.1 Introduction

The standard method for sampling diatoms for ecological status and water quality assessment in Europe is to brush or scrape the upper surface of a hard substratum (Kelly et al. 1998, CEN 2014). Many workers use a toothbrush for this purpose, in many cases reusing the same toothbrush at several sites and using stream water to rinse the biofilm off the stones and into containers.

Kelly and Zgrundo (2013) showed that the scale of contamination with this method was low and was unlikely to have a significant effect on ecological status assessments when diatoms were analysed by LM. However, the same procedure has been adopted for sampling diatoms for NGS analyses, a much more sensitive procedure. Sampling using disposable, sterile equipment is more common in molecular ecology studies (see, for example, Bista et al. 2017). However, such approaches would generate large quantities of non-biodegradable waste if adopted for a nationwide sampling campaign, as well as requiring samplers to carry pure water in the field.

A study of the scale of contamination from toothbrushes that have already been used at other sites was therefore made to provide insights into the most practicable approach for sampling diatoms for NGS analysis.

3.2 Materials and methods

3.2.1 Study design

The study design is similar to that used by Kelly and Zgrundo (2013), with 2 sites with very different characteristics selected to ensure that the diatom assemblages encountered had very little similarity with each other.

Details of the sites are given in Table 3.1. The River Nadder is a chalk stream in Wiltshire which is classified as having moderate ecological status, with macrophytes and phosphorus driving the classification (fish are at good status, invertebrates are high status and other chemical parameters are all high status). Ober Water, in contrast, is a stream in the New Forest with softer (but still around neutral) water and which is classified as being at good ecological status, with macrophytes and phytobenthos and all chemical parameters at high status.

Five samples were collected at each site for each of 3 treatments:

- samples collected using brand new toothbrushes and using distilled water
- samples collected using brand new toothbrushes and using river water from site

- samples collected using a toothbrush previously used at the other site, along with river water from the sampling site

In addition, 3 control samples were collected:

- one using just distilled water
- one each using river water from the 2 sites

3.2.2 Sampling and analysis of benthic diatoms

Sampling involved brushing the upper surface of 5 cobble-sized stones and collecting the suspension. 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 consisting of 3.5M ammonium sulphate, 17mM sodium citrate and 13mM ethylenediaminetetraacetic acid (EDTA). These samples were transferred to the laboratory in a cool box and frozen at -30°C prior to extraction of the deoxyribonucleic acid (DNA). The methods used for DNA extraction, amplification and analysis follow those described in Section 3.3 of Environment Agency (2018).

Table 3.1 Background information on the sites used in the study

	Nadder, Tisbury Station	**Ober Water, upstream A35**
Location	ST 94616 29152	SU 24964 03815
Altitude (m)	90	30
Alkalinity (mgL^{-1} $CaCO_3$)	177	13.6
pH	8.0	7.9
Conductivity (µScm^{-1})	476	138
Ammonia-N (mgL^{-1})	0.0008	0.0006
Nitrate-N (mgL^{-1})	4.02	0.22
Reactive P (mgL^{-1})	0.169	0.005
Current ecological status:		
Overall	Moderate	Good
Macrophytes and phytobenthos	Moderate	High
Phosphorus	Moderate	High

Notes: Values for chemical variables are averages for the 12 months before March 2017.

3.3 Results

The distilled water control sample contained just 23 reads compared with an average of 27,181 reads for all other samples. This control sample is not considered further.

The diatom assemblage from the River Nadder (based on LM) was dominated by *Navicula lanceolata* (average 41% in ‘pure’ samples) along with *Amphora pediculus* (8%), *Melosira varians* (8%), *Nitzschia recta* (7%) and *Navicula gregaria* (5%).

In contrast, the diatoms from Ober Water were dominated by *Achnanthidium minutissimum* (38%) along with *Achnanthes oblongella* (15%) and *Gomphonema truncatum* (11%).

Non-metric multidimensional scaling (NMDS) of the dataset yielded an ordination with very low stress (0.0684), with a clear separation between the 2 sites along axis 1 (Figure 3.1). However, some samples from Ober Water which were scrubbed using toothbrushes previously used at the Nadder site had lower scores on axis 1 than those scrubbed with clean toothbrushes, suggesting some contamination from site 1. The river water control samples (brown and blue dots in Figure 3.1) are apart from each other and from the biofilm samples along axis 2.

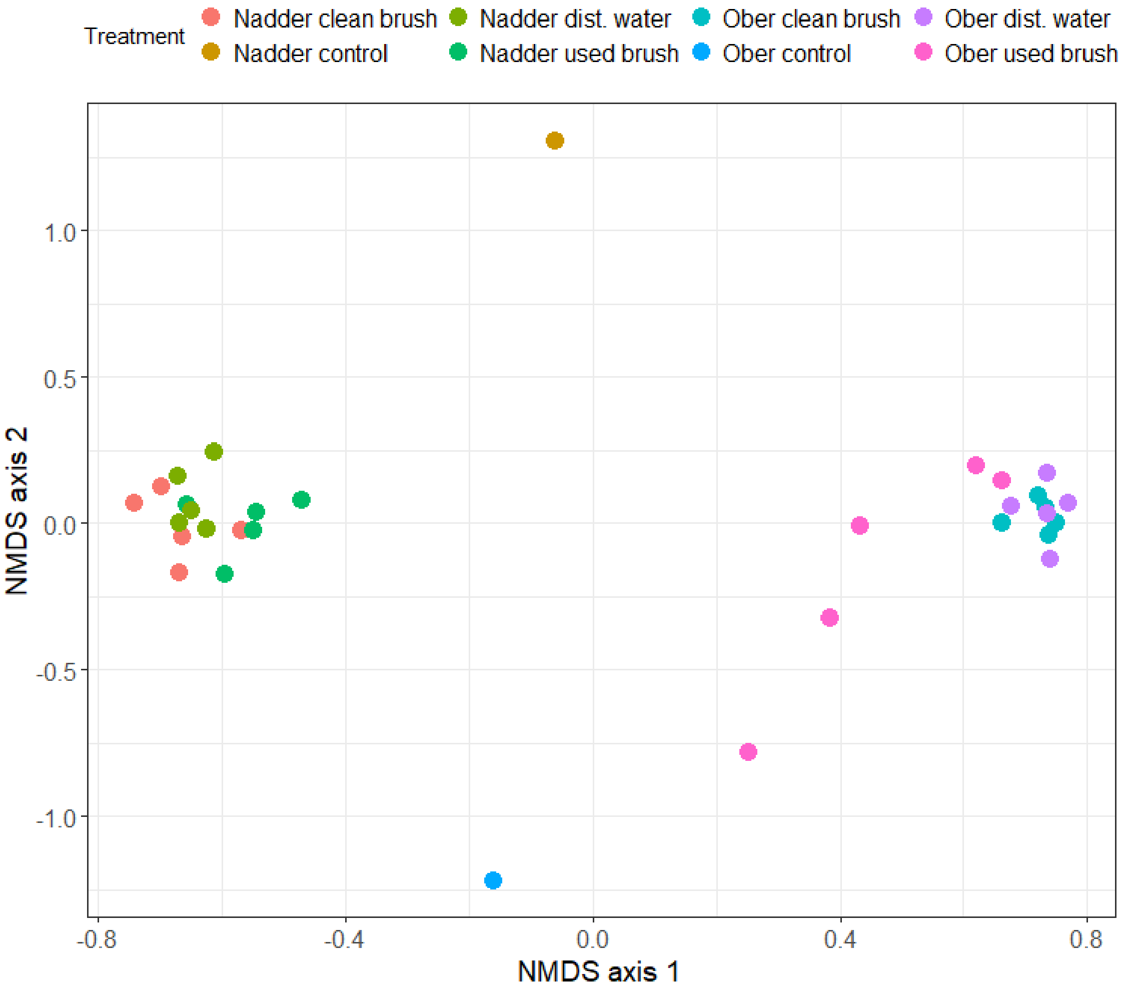

Figure 3.1 Plot showing position of samples from River Nadder and Ober Water relative to the first 2 axes of an NMDS ordination

If there is a significant amount of contamination, then taxa that are abundant at one site should be present in raised numbers in samples collected using dirty equipment at the other site but rare in the others. Although significant effects were observed for several taxa, the scale of the effect was generally small, particularly for samples from the River Nadder where the increase in samples collected with contaminated toothbrushes exceeded 1% only for *Achnanthidium minutissimum* (Figure 3.2). The scale of the increase was greater in Ober Water samples, with a median increase for *Melosira varians* of about 2%, but with one replicate having an increase >10% relative to the sample collected with clean equipment.

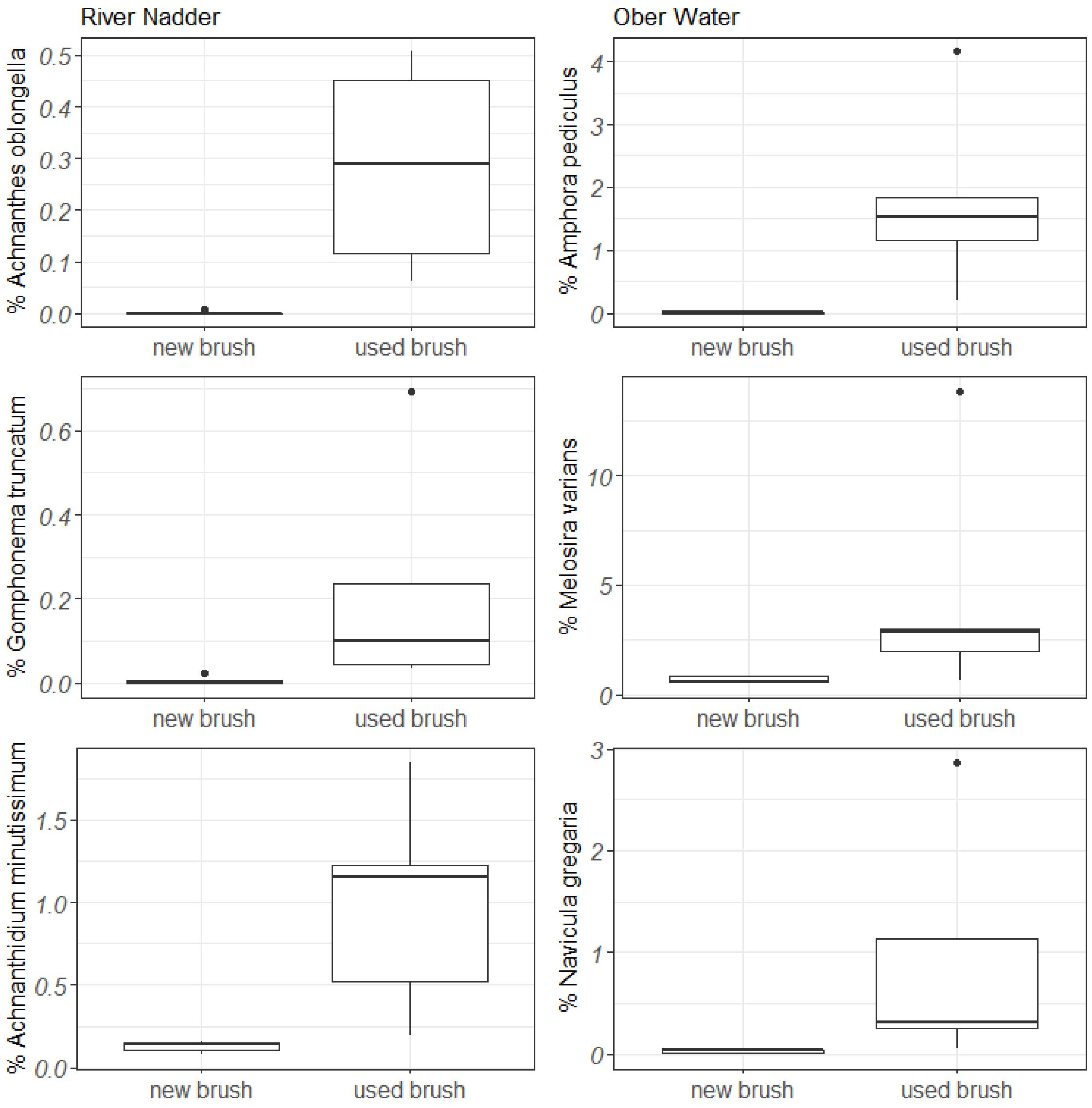

Figure 3.2 Variation in proportions of taxa that formed a major part of the assemblage at one site in samples collected with clean and contaminated equipment at the other site

A similar approach was adopted to look at possible contamination from stream water. The relative abundance of the most abundant taxa in the stream water sample from each stream was compared with the samples washed with stream and distilled water from that location.

In the case of the River Nadder, the stream water was dominated by planktonic diatoms (65% of total reads), probably originating from ponds and fish farms upstream. Three of these – *Stephanodiscus hantzschii, Cyclostephanos invisitatus* and *Discotella* sp. – were all elevated with respect to distilled water sample (Figure 3.3), but only in relatively small numbers (that is, still <1% in the worst case). Differences for *S. hantzschii* and *C. invisitatus* were both significant (Kruskal test).

There were almost no planktonic diatoms in the Ober Water stream water; however, the composition of the sample was quite different to that of biofilm samples, with a greater proportion of nutrient-rich taxa. There was, however, no significant increase in proportions of these taxa in the biofilm when stream water was used to wash the stones (Figure 3.4).

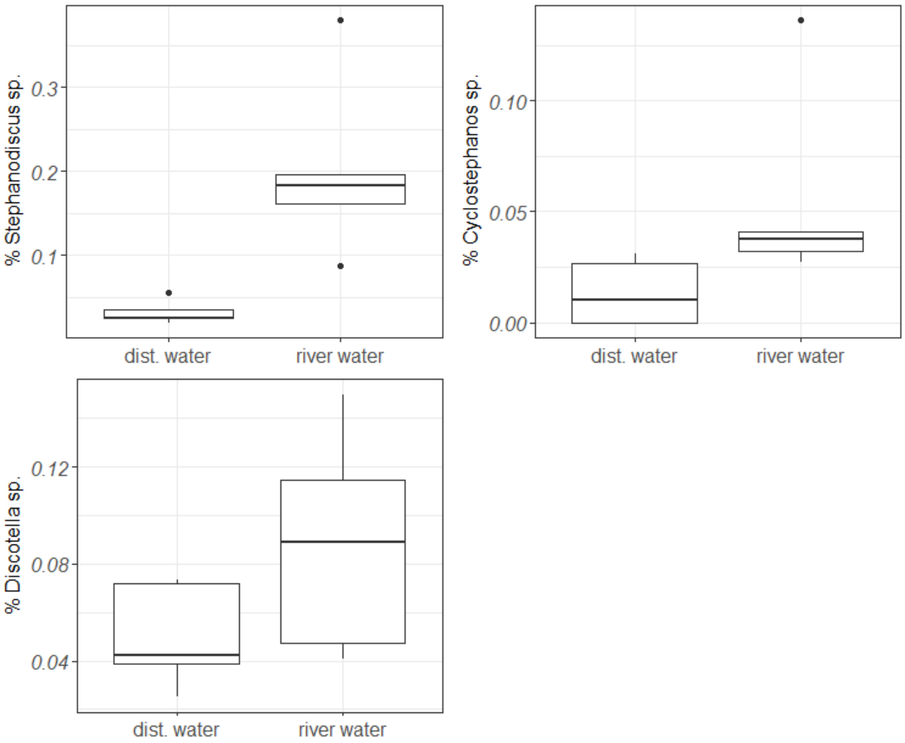

Figure 3.3 Variation in proportions of taxa that were abundant in stream water from the River Nadder at the time of sampling in biofilm samples collected with stream and distilled water respectively

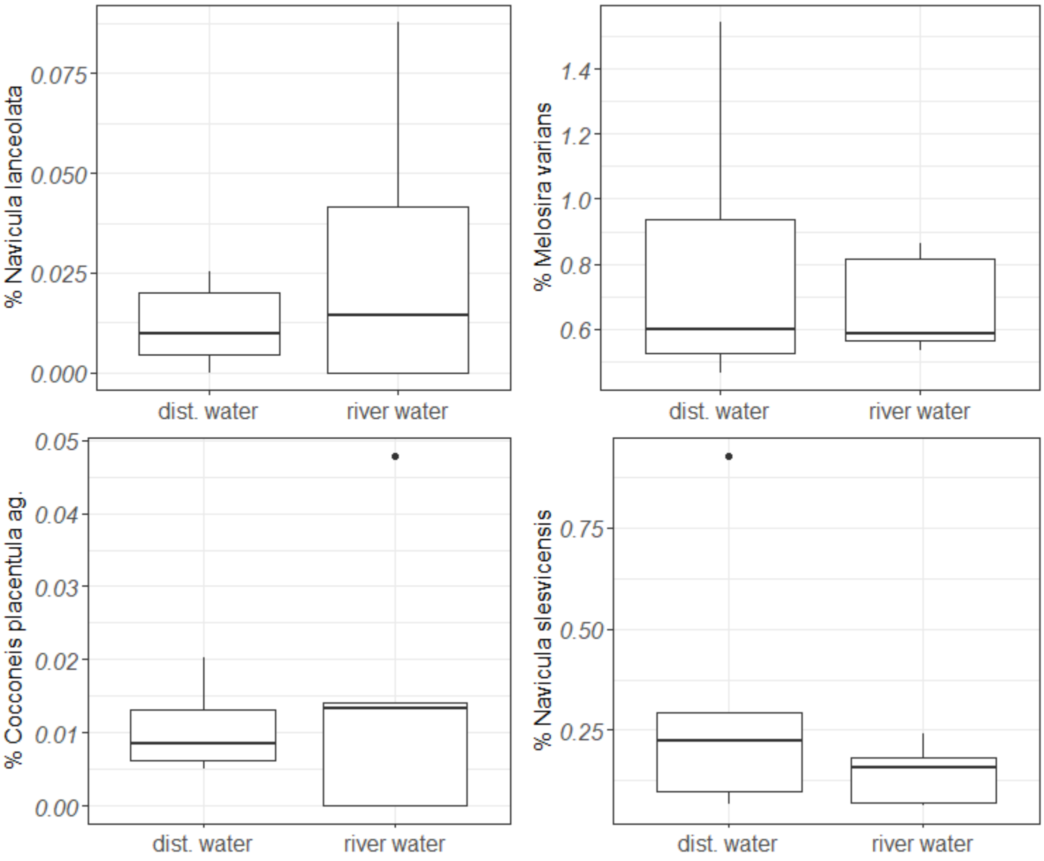

Figure 3.4 Variation in proportions of taxa that were abundant in stream water from Ober Water at the time of sampling in biofilm samples collected with distilled and river water respectively

The effect on the TDI5 score is counterintuitive in both the River Nadder and Ober Water. There are differences between treatments in both cases, though these are only significant for Ober Water. However, the TDI5 score is lower in the treatments where Ober Water biofilms were removed with toothbrushes formerly used in the more enriched River Nadder (Figure 3.5).

In Section 7 of the SC140024, report (Environment Agency 2018), the scale of variation was shown to vary between rivers. However, if the average level of variation measured at a site on a single day is assumed then the variation between treatments in both cases falls within the expected range.

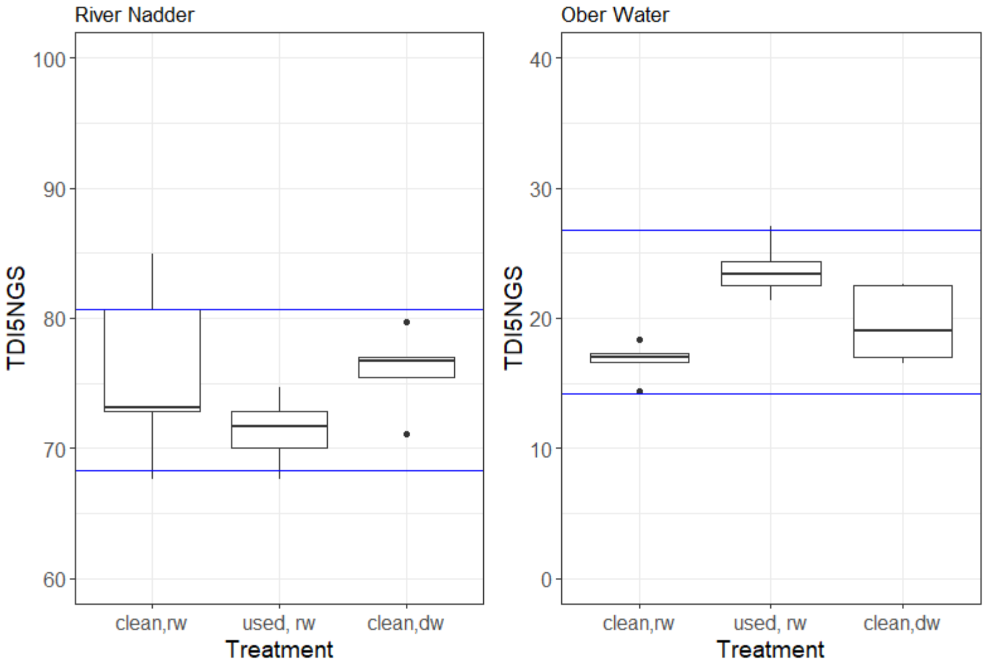

Figure 3.5 Variation in TDI5 between treatments for River Nadder and Ober Water

Notes: Horizontal lines show the upper and lower limits of variation expected for replicate samples from a site on a single day (6.2; twice the average standard deviation observed – see Section 3 of the SC140024 report), using the samples collected using clean toothbrushes and distilled water as the benchmark.

3.4 Discussion

The results of this study highlight a potential for toothbrushes to retain traces of the diatom assemblage even after the routine cleaning procedure (washing bristles in the stream and rubbing against waders). The scale of this contamination is relatively low but is, nonetheless, present. In particular, sampling a thick biofilm where there are entangling filamentous algae and then using the same toothbrush at a site with a very thin biofilm is more likely to lead to problems than the reverse situation. Similarly, given how WA equations work, sampling a ‘clean’ site after a visit to a ‘polluted’ one is more likely to result in problems than the other way around.

Contamination from the stream water used to wash the samples appears to be less of a problem. In the case of the River Nadder, although planktonic taxa dominated the suspended diatom assemblage, these do not contribute to the TDI5 score and so should have no effect on the final index value. Many of the planktonic diatoms have multiple chloroplasts, however, and there may be issues when sampling coincides with a plankton bloom. The possibility that this may introduce competition within the polymerase chain reaction (PCR) for amplification of rbcL sequences from benthic taxa is small, but cannot be ruled out entirely. However, with an average of over 27,000 rbcL reads per sample, even a 90% reduction in the amplification of benthic diatoms would still yield 7 times more data per sample than the current approach using LM.

In other words, there is a case for taking sensible precautions to reduce the scale of contamination from both stream water and sampling equipment. However, there is no evidence of a significant increase in precision through an approach based on equipment cleanliness and purity of reagents alone. The focus should instead be on the production of data that are an accurate reflection of conditions prevailing at the site at the time of sampling and it is clear that low levels of contamination are present, even if this does not have an effect on index calculations.

As bleach has been shown to degrade DNA, it is recommended that toothbrushes are used once and then washed in a bleach-containing solution at the end of each day, ensuring that the bristles are rubbed vigorously to remove algal traces. If this is not possible, single use toothbrushes offer an alternative – albeit creating non-biodegradable waste in the process. Similarly, while stream water is unlikely to be a major source of error, substitution by distilled water should be encouraged.

4 Potential for metabarcoding to evaluate the ecological status of lakes

4.1 Introduction

The focus of the work so far has been to develop an analogue for the present LM based method for assessing WFD ecological status using phytobenthos in rivers. At present, phytobenthos are also used to assess WFD ecological status in some lakes using a LM based method developed alongside that for rivers (Bennion et al. 2014). This section explores the possibility of extending the NGS method from rivers to lakes using samples collected during 2014 and 2016.

Phytobenthos communities in the littoral zones of lakes are similar in many ways to those in rivers, as the turbulent hydrological regime in the littoral exerts similar physical pressures to those experienced in running waters (Cantonati and Lowe 2014). There are some differences in composition, with some species showing clear preferences for one habitat over the other. Examples include *Diatoma mesodon, Hannaea arcus* and *Platessa oblongella*, which are more common in rivers, and *Cymbella* and *Epithemia* species, which are more common in lakes (although exceptions do occur). In general, however, it is possible to use metrics developed for rivers to assess lakes (see, for example, Kahlert and Gottschalk, 2014), while metrics that have been optimised for use in lakes often give similar results to those developed for rivers (Bennion et al. 2014). It is therefore possible that a set of methods and a barcode database developed for rivers will also give acceptable results when applied to lakes.

The questions addressed in this section are as follows.

- How effective for lakes is the current LM based method for assessing ecological status when calculated using NGS analyses using the existing barcode database?
- Would there be any benefit to recalibrating the Lake Trophic Diatom Index (LTDI) to produce a version optimised for NGS analyses?
- Is there any need to expand the barcode database to incorporate taxa that are more likely to be encountered in lakes?

4.2 Methods

Matched LM and NGS analyses were available for 162 samples from 42 lakes in England. Corresponding total phosphorus (TP) and total nitrogen (TN) values were available for 125 of these samples. The lakes included in these analyses are listed in Appendix 3.

For this initial look at the data, no changes to the sample processing or bioinformatics were applied. The LTDI2 was calculated in an identical manner using both LM and NGS data (Bennion et al. 2014). The current LTDI2 reference values, established using LM data, were used (low alkalinity: 22; moderate alkalinity: 35; high alkalinity: 42) and the present WFD ecological status boundaries were employed without any modification.

4.3 Results

The LTDI calculated by the 2 methods showed good agreement (Figure 4.1a; Pearson's correlation coefficient, r = 0.82, p < 0.001), with only slight deviation from slope = 1 (Lin's concordance correlation coefficient = 0.79). Whereas NGS data gave distinctly non-linear fits to LM variants of river metrics (Figure 6.7 in Environment Agency 2018), the relationship in Figure 4.1a is relatively straight albeit with a tendency to overestimate LTDI2 values in the middle and upper part of the range (Figure 4.1b). LTDI2 values obtained using NGS data were on average 5.4 units greater than those generated using LM data.

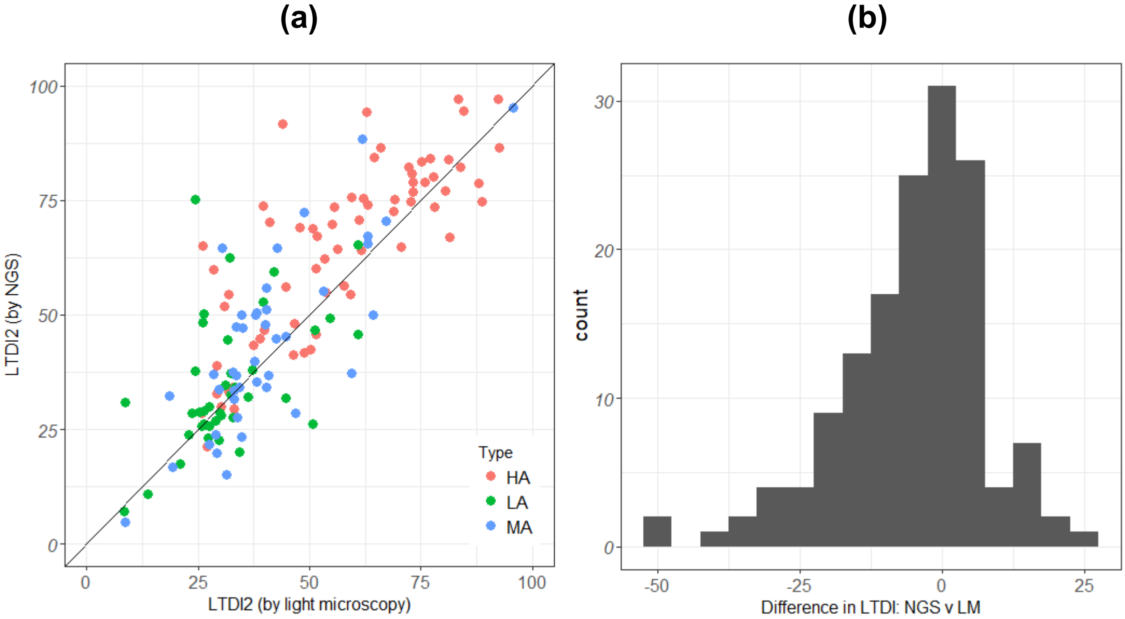

Figure 4.1 (a) Relationship between LTDI2 calculated using data generated by LM and NGS. (b) Difference between LTDI2 calculated by LM and NGS

Notes: Diagonal line in (a) indicates slope = 1; HA = high alkalinity (red dots); MA = moderate alkalinity (blue dots); LA = low alkalinity (green dots)

On average, 44% of total reads from each NGS sample (standard deviation 17.5%) matched with taxa used to calculate LTDI2 (Figure 4.2a), but there was no systematic trend of samples with high or low percentages of assigned reads having higher or lower LTDI2 values (Figure 4.2b). The proportion of reads not assigned to taxa is similar to that observed in Section 2 during development of the LTDI5NGS for rivers (see Figure 2.1). It is possible that a proportion of these unassigned taxa will relate to planktic taxa that have settled from the benthos. Most of the planktic genera have multiple chloroplasts and planktic populations are generally larger in lakes than in rivers, so this could be a source of noise in NGS analyses of lake samples. Generally, however, this was not the case (Figure 4.2c), with instances of samples with both very high and very low observed percentages of assigned reads being associated with low numbers of planktic taxa, although there were a few samples that had high percentages of both planktic taxa and unassigned reads. It should also be noted that, while the standard method for LM analysis states that planktic taxa should be recorded, practice does vary as these taxa are not used as part of the WFD ecological status assessment.

The proportion of planktic taxa also varies between LM and NGS samples (Figure 4.2d). Most littoral samples have low proportions, regardless of the mode of analysis; however, samples with high proportions recorded in LM analyses tend not to have high proportions in NGS and vice versa. Whether this relates to gaps in the barcode database, LM counting procedures or is an inherent difference between the 2 types of data cannot be ascertained with certainty. Although this does not influence final ecological status assessments, it does offer some useful insights into the performance of the NGS data.

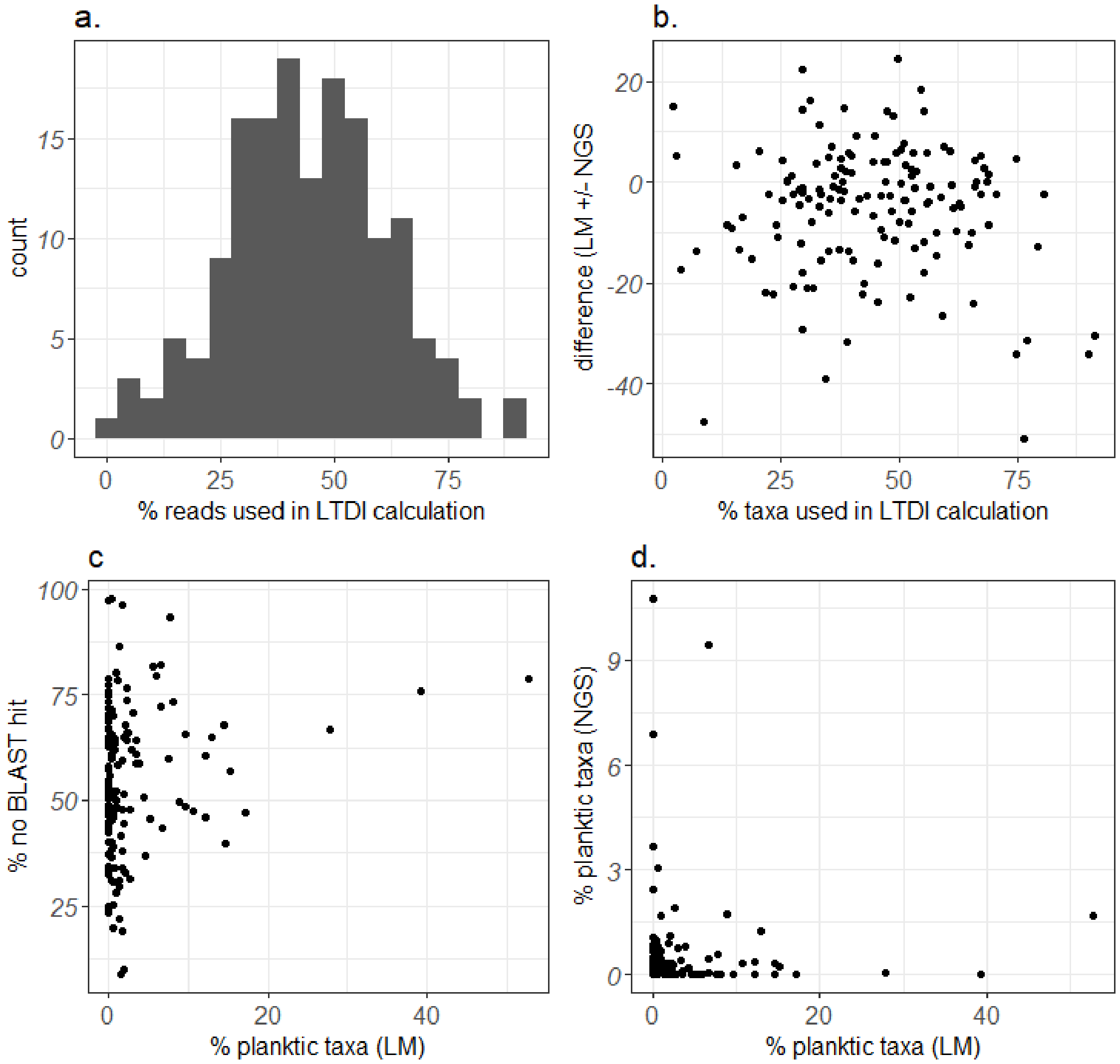

Figure 4.2 Properties of NGS lake data: (a) percentage of total reads used in LTDI2 calculation; (b) relationship between percentage of reads used in calculation and difference between LTDI2 calculated with LM and NGS data; (c) relationship from proportion of total reads not assigned to any taxon in database ('no BLAST hit') and the percent of reads used in LTDI calculation; and (d) comparison of proportion of count used to calculate LTDI2 using LM and NGS data

Notes: BLAST = Basic Local Assignment Search Tool

LTDI2 calculated with both LM and NGS data had significant relationships with both TP and TN (Figure 4.3 and Table 4.1). Regressions produced with LM data were slightly stronger than those produced with NGS data for both variables, but the difference in both cases was small.

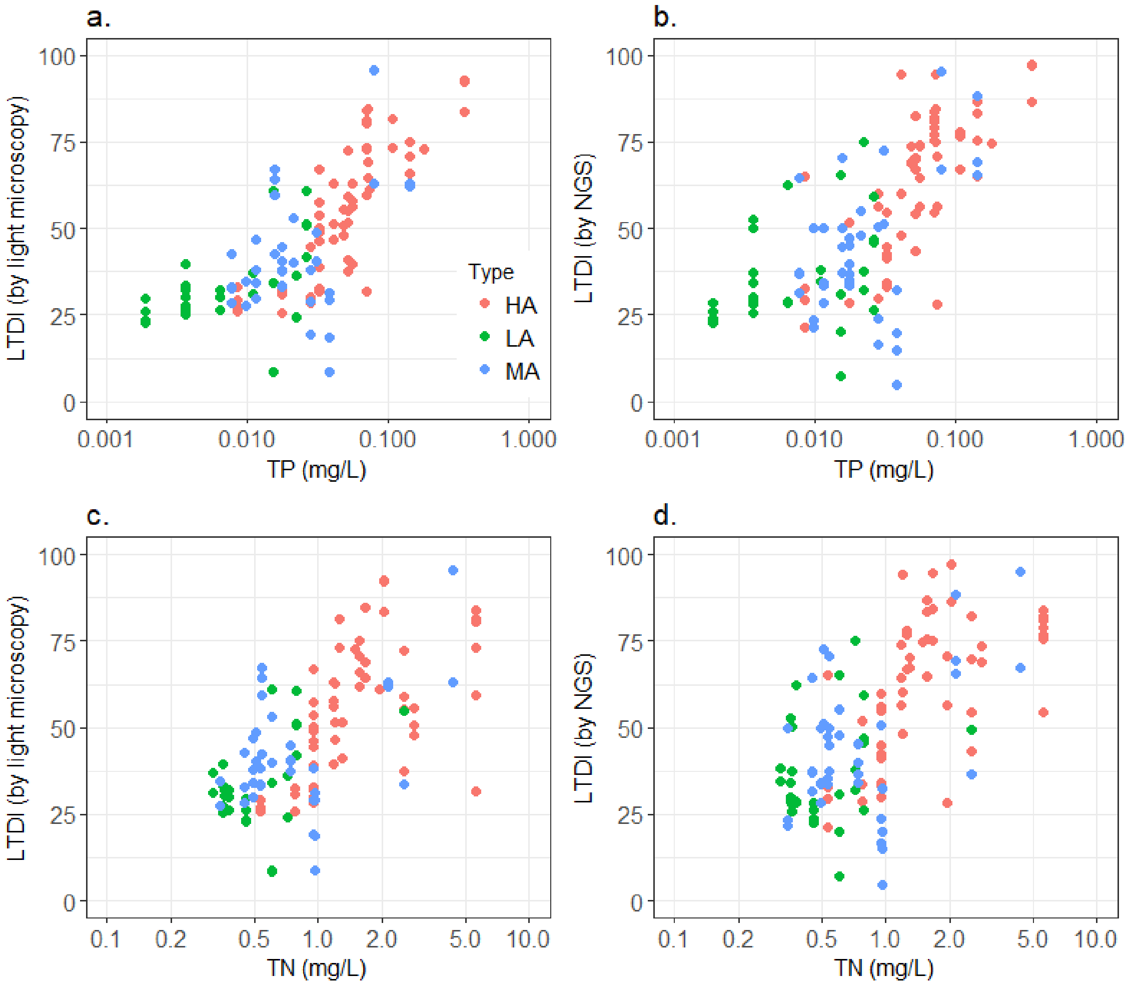

Figure 4.3 Relationship between LTDI2 computed with LM and NGS data and nutrients: (a) LTDI2 calculated with LM data versus TP; (b) NGS LTDI2 calculated with NGS data versus TP; (c) LTDI2 calculated with LM data versus TN; and (d) LTDI2 calculated with NGS data versus TN.

Notes: HA = high alkalinity (red dots); MA = moderate alkalinity (blue dots); LA = low alkalinity (green dots).

Table 4.1 Regression parameters for relationships between LTDI2 calculated with LM and NGS data and nutrients

Relationship	F (degrees of freedom)	Significance	r^2
TP versus LTDI2			
LM data	126 (1/119)	<0.001	0.51
NGS data	103 (1/121)	<0.001	0.46
TN versus LTDI2			
LM data	89 (1/121)	<0.001	0.42
NGS data	80.5 (1/123)	<0.001	0.39

The final stage of this preliminary analysis of the potential for the NGS method to assess WFD ecological status in lakes involved computing EQRs for all samples from a water body, then aggregating these to give a mean EQR for that water body from which ecological status could be determined.

Overall, about half of the lakes (52%) were assigned to the same ecological status class using both methods (Table 4.2), with just 4 lakes being classified more than one class apart. The tendency for LTDI2 calculated with NGS data to return higher values than when calculated with LM data translates into a tendency for NGS data to produce more stringent classifications, with 38% of samples being downgraded compared with just 10% being moved to a higher status class. The overall rate of bias of the NGS method is 29%, with the effect being particularly marked in moderate and high alkalinity lakes. Using the relationship between LM and NGS variants to calculate approximate conversion factors for reference value and boundaries reduces this bias to 10%.

Table 4.2 Ecological status classification of 52 lakes in England using LM and NGS variants of LTDI2

		Status assessed using NGS data				
		High	**Good**	**Moderate**	**Poor**	**Bad**
Status assessed using LM data	**High**	13	6	2		
	Good	3	9	7	1	
	Moderate	1	1	1	4	
	Poor				3	
	Bad					1

Notes: Green shading: identical classification for both LM and NGS; yellow shading: agreement to within one class between LM and NGS; red shading: greater than one class difference between methods

4.4 Conclusions

The results from this brief study are encouraging and suggest that it should be possible to produce a version of the LTDI that is optimised for NGS data relatively easily.

Although the overall agreement between LTDI2 values calculated using LM and NGS data is good, there is a tendency for NGS data to yield higher values of the LTDI, particularly at the middle and upper part of the scale. This translates into more stringent ecological status classifications than those generated by LM data when the current reference values are employed. In practice, the tendency of the NGS metric to yield higher values will mean that the 'expected' value for moderate and high alkalinity lakes are too low for use with NGS data and adjusting these will, in turn, reduce the bias. No attempt has been made to examine reference values at this stage, largely because several of the reference sites, particularly at low and moderate alkalinity, are in parts of the UK not covered by this study.

However, a sensible first step would be to follow the procedures described in Section 2 and optimise the LTDI species sensitivity scores for use with NGS data. This should improve agreement between LM and NGS versions, and also reduce any differences between reference values employed by the 2 methods (though a recalibration of reference values is still recommended).

Although a large proportion of the NGS reads from lake samples cannot be assigned to taxa, this is a problem shared with data from rivers; the evidence from Figures 4.1 and 4.3 is that this is having little effect on index performance. As for rivers, the issue of unmatched reads should not be dismissed, as it is a sign that a large part of the ecological signal is not being incorporated into the barcode database. In the case of lakes, it is possible that a large number of reads belong to planktic centric diatoms (each of which with several plastids per cell) that play no role in the LTDI. However, as for rivers, it is also likely that the full range of genetic variation of some key indicator taxa is not yet represented in the barcode database. Better coverage of planktonic diatoms will have added benefits were a NGS tool for evaluating lake phytoplanktons to be considered.

It should therefore be possible to build on this preliminary study and establish a lake counterpart to the TDI5NGS as described in Section 2. The next step should include sampling from lakes throughout the UK, particularly to fill out the lower end of the trophic and alkalinity gradients. There would need to be a particular focus on reference lakes as it will be necessary, at the very least, to validate the reference values used to compute EQRs and perhaps refine them.

The possibility of using the reference dataset to establish a predictive reference equation to replace the type-specific denominators should also be considered, as this will bring the phytobenthos assessment system in line with other lake assessment tools.

Methods for developing a NGS based analogue of the current LM based system are described in Section 2 and this is a sensible first step in order to gain understanding of how the properties of molecular data differ from those of conventional biological assessment data. The barcode database, as established for WFD ecological assessment of rivers is adequate for the assessment of ecological status in lakes; however, it is possible that enlargement of this to include a wider range of both species and haplotypes will ultimately allow the full potential of the method to be exploited.

5 Case study: changes in diatom assemblages along Polly Brook, Devon

5.1 Introduction

Development of a metric requires the collection of data from a wide range of ecological conditions in order to establish the behaviour of biological communities in response to major pressure gradients, while also taking account of confounding factors. In practice, the datasets developed for this project provide broad spatial coverage of England, along with some coverage of other parts of the UK. In most cases, this involved collecting 2 samples per site (in the spring and autumn) following standard practice in the Environment Agency – although initial studies (Kelly et al. 2009) recommended 4 or 6 samples per classification period. In most cases, data from only a single site per water body are available even though the latter is the standard unit on which classifications are based. It was therefore necessary to extend the metric development process beyond statistical analysis of spatial datasets to consider patterns of variation at different scales in light of the uncertainties that these bring to the ecological status classification of water bodies.

Studies of variation at different spatial and temporal scales are included in Section 7 of the SC140024 report, with Section 8 extending this approach to look at an operational investigation within 2 (connected) water bodies (Environment Agency 2018). Understanding how ecological status assessment tools work at this scale is important as Environment Agency biology teams will use them for operational investigations designed to inform regulatory decisions. In addition, it is useful to understand how variation within a water body is influenced by the pressures that will vary in intensity along the length of that water body. Some insights into those processes were obtained during the earlier study and this section continues this learning process by focusing on a small catchment draining into the estuary of the River Exe in Devon.

Polly Brook is located in south Devon, about 6km south-east of Exeter. It rises at altitude of about 140m on the side of Woodbury Common in woodland underlain by Triassic epoch mudstones, and flows westwards through farmland to join the estuary of the River Exe at Exton, approximately 12km from the source. The stream skirts the village of Woodbury and receives effluent from a small sewage treatment works (STW) (population equivalent of 1,600, rising to 2,400 during the tourist season) as well as from storm sewers. In the past, there have been pollution incidents associated with this storm sewer overflow (SSO) which resulted in a fish kill. However, investigations at the time revealed dead invertebrates upstream of the SSO while *Cladophora glomerata* is also conspicuous on the bed of the stream above the village suggesting that there may be other impacts on the stream (for example, from agriculture). Maize was grown on land adjacent to the stream upstream of the village during 2016 and there are also 2 properties with consented discharges upstream of the village itself, both of which may contribute nutrients to the river.

5.2 Materials and methods

Four locations were identified by local Environment Agency staff with the aim of:

- separating the effects of the village and the sewage works
- coinciding with the sole routine monitoring site in the catchment (BIOSYS site 9569) (Table 5.1)

Each site was sampled in September, October and December 2016. The December sample replaced a planned November visit that was postponed due to high flows. Water chemistry samples were collected from each site at the same times as the diatom sample.

Sampling and analysis protocols follow those used in Environment Agency (2018). Thresholds for supporting element standards were obtained from the WFD UK Technical Advisory Group website (www.wfduk.org). A similar threshold for nitrate-N is not available, and so a guide value, derived using the same principles as the supporting element standard for phosphorus, was obtained from Kelly (2016). As only 3 chemical samples were available for each site, their relationship to supporting element standards (developed using annual mean chemistry) should be interpreted with caution.

Table 5.1 Site locations and summary information

Site	BIOSYS site ID	Location	NGR	Distance (km)	Altitude (m)	Alkalinity (mgL^{-1} CaCO$_3$)
1	174389	u/s Woodbury	SY0113 8667	5	100	208
2	165868	u/s bridge, u/s Woodbury STW	SY0016 8667	8	60	212
3	166923	30 d/s discharge Woodbury STW	SX 9977 8682	10	30	207
4	9569	200m d/s A376 bridge at Exton	SX9836 8627	11.5	10	210

Notes: d/s = downstream; u/s = upstream

5.3 Results

5.3.1 Key environmental variables

Ammonia-N and phosphorus as orthophosphate show a clear trend with increased concentrations downstream of Woodbury STW. Concentrations of ammonia-N downstream of the STW are still low, not exceeding the current threshold to support 'high' status (Figure 5.1a). However, phosphorus concentrations immediately downstream exceed the threshold to support 'bad' status and can only support 'poor' status at the routine site at Exton (Figure 5.1b). Phosphorus concentrations at both sites upstream of the STW exceed the current threshold to support 'good' status, suggesting some enrichment from agriculture and/or septic tanks (Figure 5.1b). Routine monitoring data (not shown) show phosphorus concentrations at Exton averaging ~0.4mgL^{-1}, again only able to support poor status, while the limited data from upstream of the STW suggest intermittent pulses, some exceeding 0.2mgL^{-1}.

There is a sharp increase in total oxidised nitrogen between the first and second sites, after which concentrations remain high but decline slightly (Figure 5.1c). There no ecological thresholds for inorganic nitrogen in the UK. Using approximate values derived using similar principles to the phosphorus standards, however, site 1 is below the threshold to support good status (3.9mgL^{-1}) while the other sites can only support poor status.

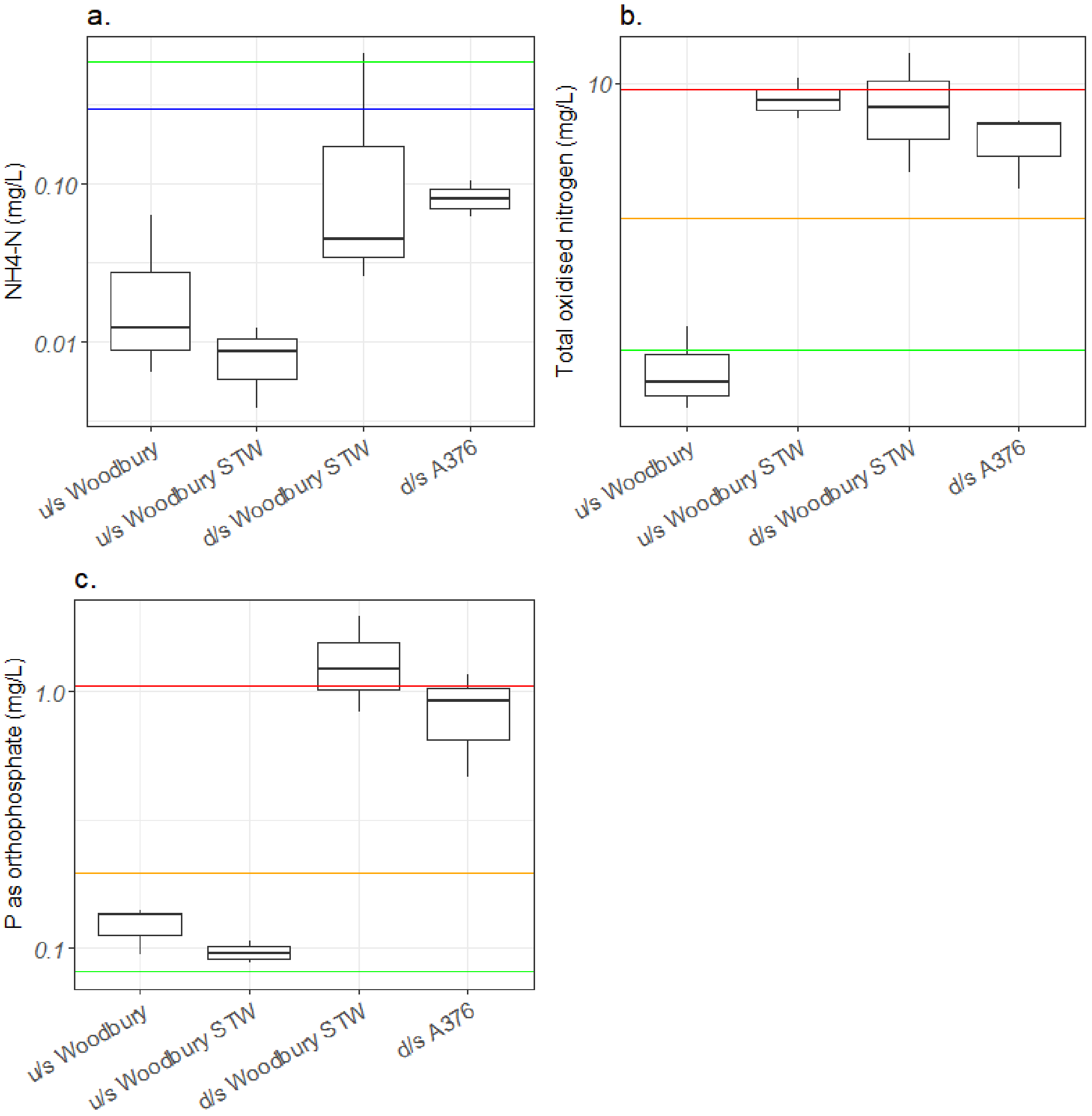

Figure 5.1 Variation in ammonia-N (a), total oxidised nitrogen (b) and phosphorus as orthophosphate (c) at 4 sites on Polly Brook in south Devon, September to December 2016

Notes: Horizontal lines show UK standards for supporting elements (blue = high/good, green = good/moderate; orange = moderate/poor; red = poor/bad ecological status boundaries) for ammonia-N and P as orthophosphate.
Horizontal lines on the total oxidised nitrogen graph (b) are guide values with no regulatory significance calculated using similar principles to the phosphorus standards (see text for details).

5.3.2 TDI scores

Values for the TDI score, whether calculated using LM or NGS data, only partially reflect the patterns shown by the water chemistry. The TDI score calculated using LM and NGS differ by ≤7 TDI units (the threshold for acceptable replication within LM audits) in 6 instances. Variation among the other samples ranges from 9 to 31 units, with results from December 2016 showing the greatest variability (Figure 5.2). LM detects a step change between upstream and downstream of the Woodbury STW (sites 2 and 3), with the lowermost site, downstream of the A376 bridge (site 4), having similar TDI values to the site immediately downstream of the STW. However, the uppermost site in the catchment (site 1) also has relatively high TDI values, for reasons that are hard to explain. In contrast, there are no step changes between sites in the NGS data, with no obvious effect from the Woodbury STW.

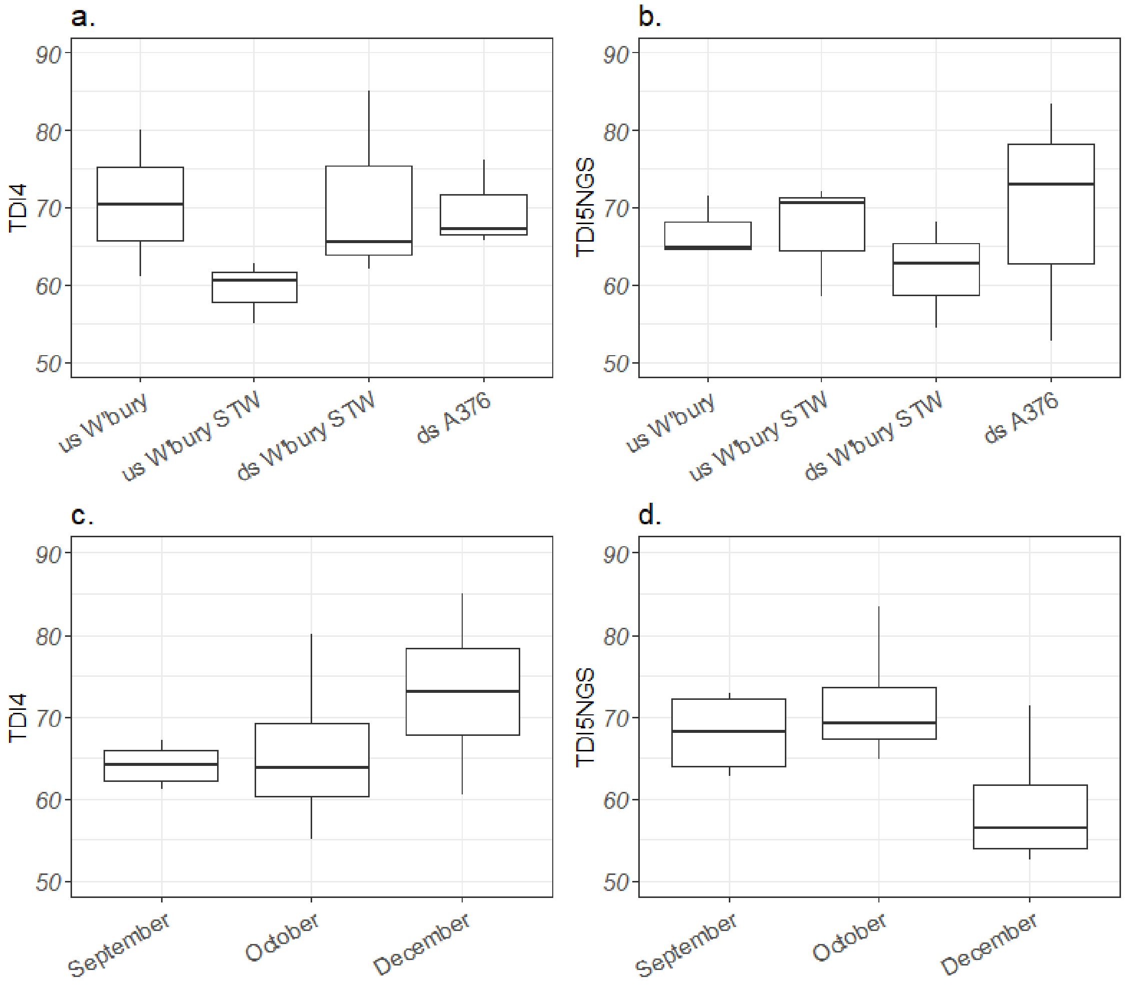

Figure 5.2 Variation in TDI by reach (top row) and month (bottom row) for analysis by LM (TDI4, left) and NGS (TDI5, right)

All the samples fall within the range of variation seen in project SC140024 (Figure 5.3). Two of the 4 samples that are below the 1:1 line were collected in December 2016 and are linked to the taxa changes associated with that month (see Figure 5.7).

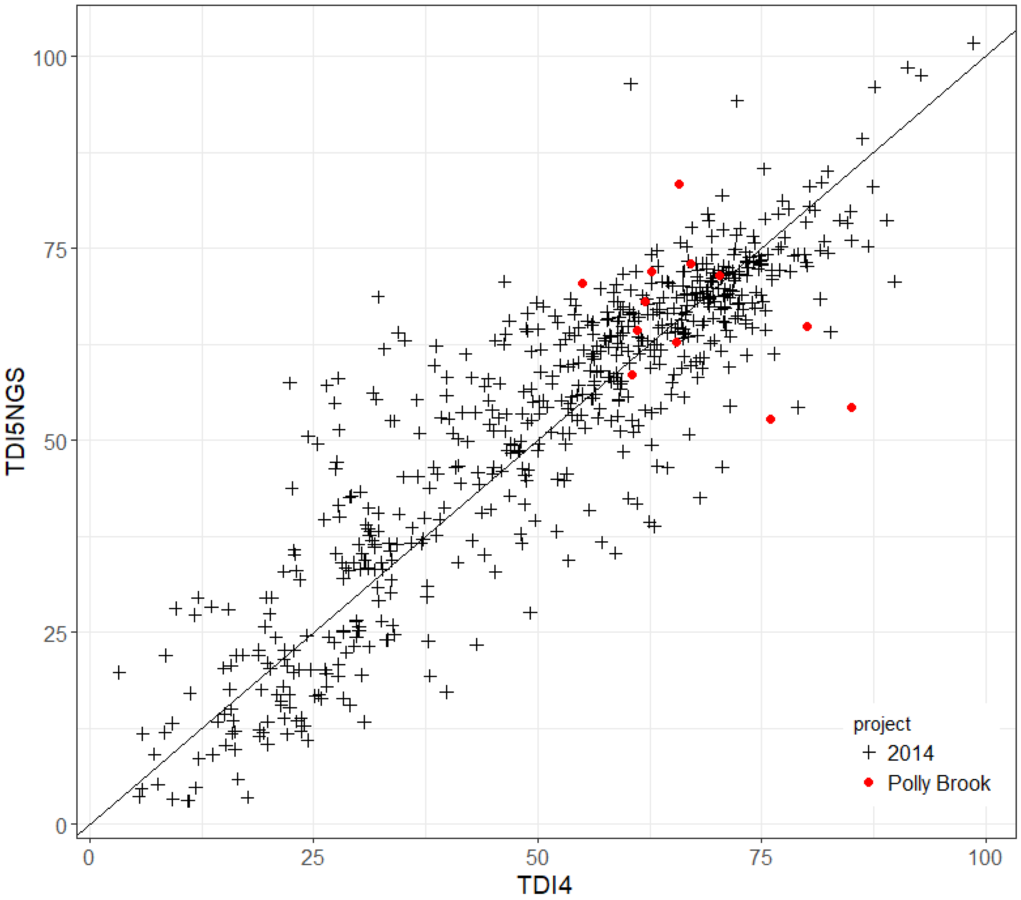

Figure 5.3 TDI4 and TDI5 values for Polly Brook diatom samples overlain onto the 2014 dataset used in project SC140024

Notes: Diagonal line indicates slope = 1.

5.3.3 NMDS analysis of LM and NGS data

NMDS of both LM and NGS data yields an ordination with low stress (0.137 and 0.096 respectively) (Figures 5.4 and 5.5).

Season is a significant factor determining the LM distribution but not that for NGS (Table 5.2).

Reach is not significant for either ordination and, for both LM and NGS, there is no significant correlation between the TDI score and either of the first 2 axes (Table 5.2).

The seasonal effect is largely the result of different behaviour in September. There is no significant difference between the LM or NGS ordinations

(Procrustes test), but this may be simply a consequence of comparing 2 small and rather variable datasets.

Table 5.2 Results of analysis of similarity (ANOSIM) for diatom assemblages measured by LM and NGS using reach and month as factors

	LM	**NGS**
Reach	-0.142 N.S.	-0.145 N.S.
Month	0.526 **	-0.026 N.S.

Notes: ** $p \geq 0.01$; N.S. = not significant

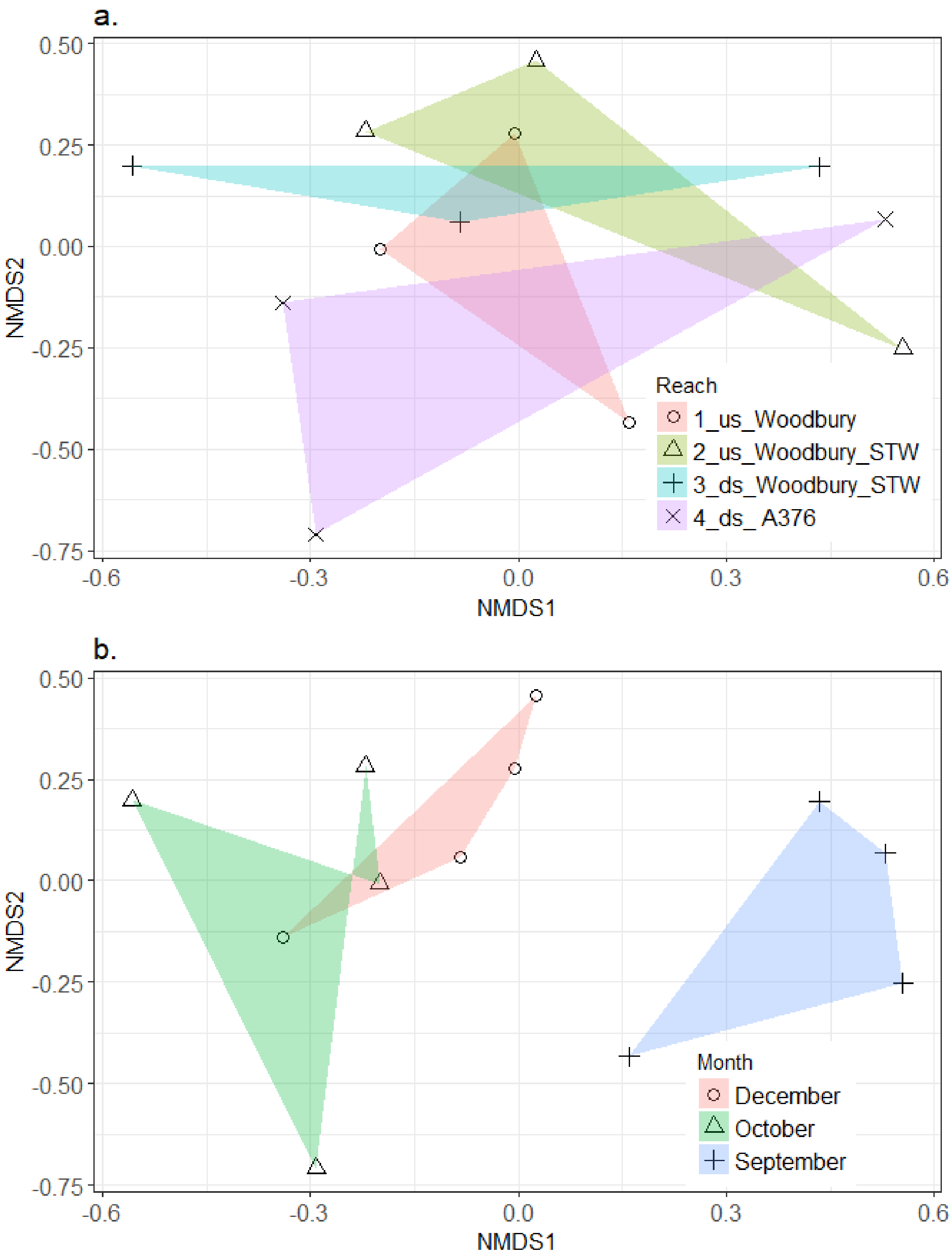

Figure 5.4 Ordination plots showing the first 2 axes of an NMDS analysis for diatoms from Polly Brook analysed by LM with samples grouped by (a) site and (b) month

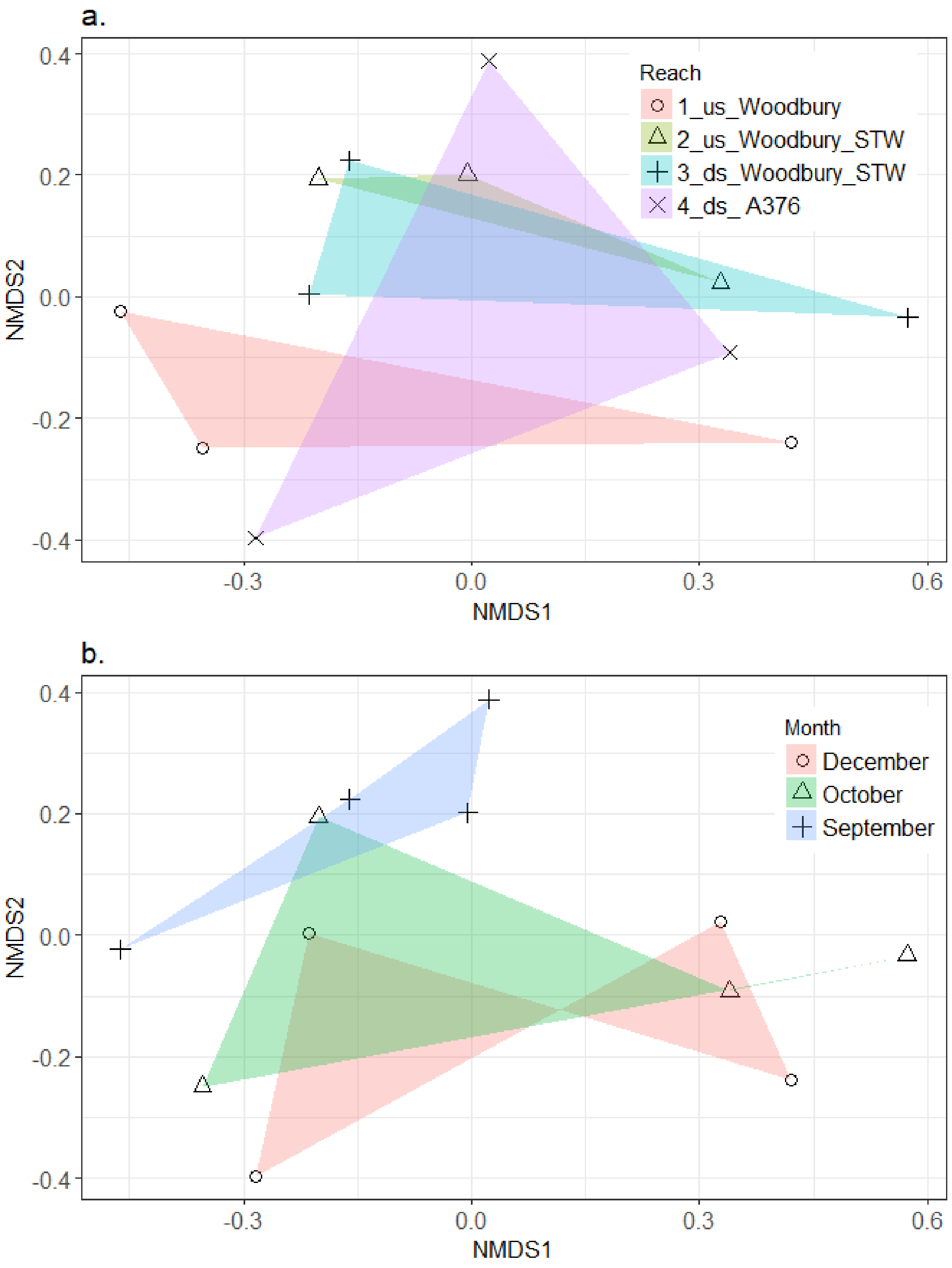

Figure 5.5 Ordination plots showing the first 2 axes of an NMDS analysis for diatoms from Polly Brook analysed by NGS with samples grouped by (a) site and (b) month

5.3.4 Species composition

In general, the species composition was similar between the 2 methods. However, several differences were apparent in their relative abundance along the stream according to the 2 methods (Figures 5.6 and 5.7).

(a)

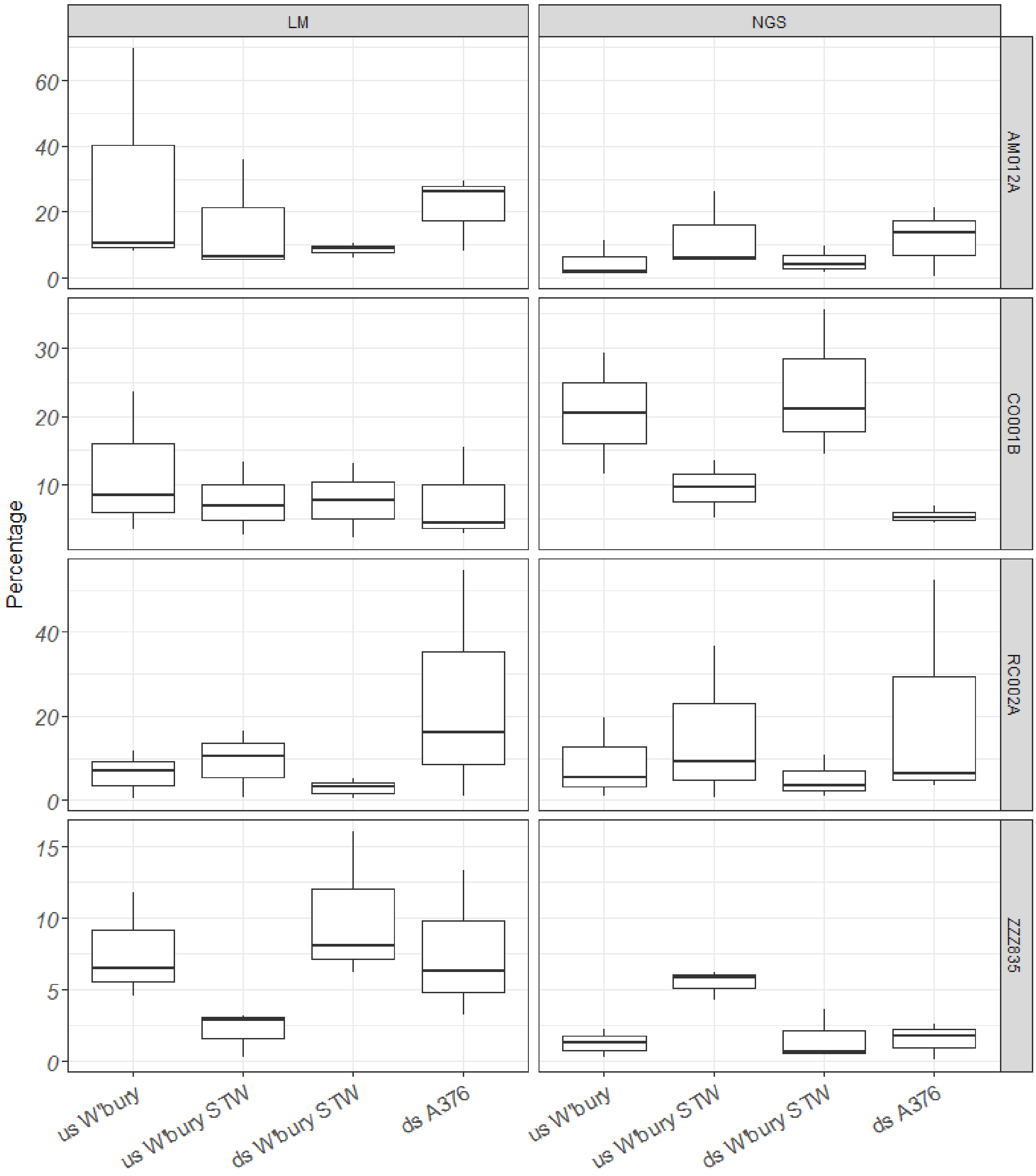

Figure 5.6a Distribution of common taxa in Polly Brook by reach: *Amphora pediculus* (AM012A); *Cocconeis placentula* group (CO001B); *Rhoicosphenia abbreviate* (RC002A); and *Achnanthidium minutissimum* complex (ZZZ835)

Notes: LM = proportion recorded by light microscopy; NGS = proportion recorded by next generation sequencing

(b)

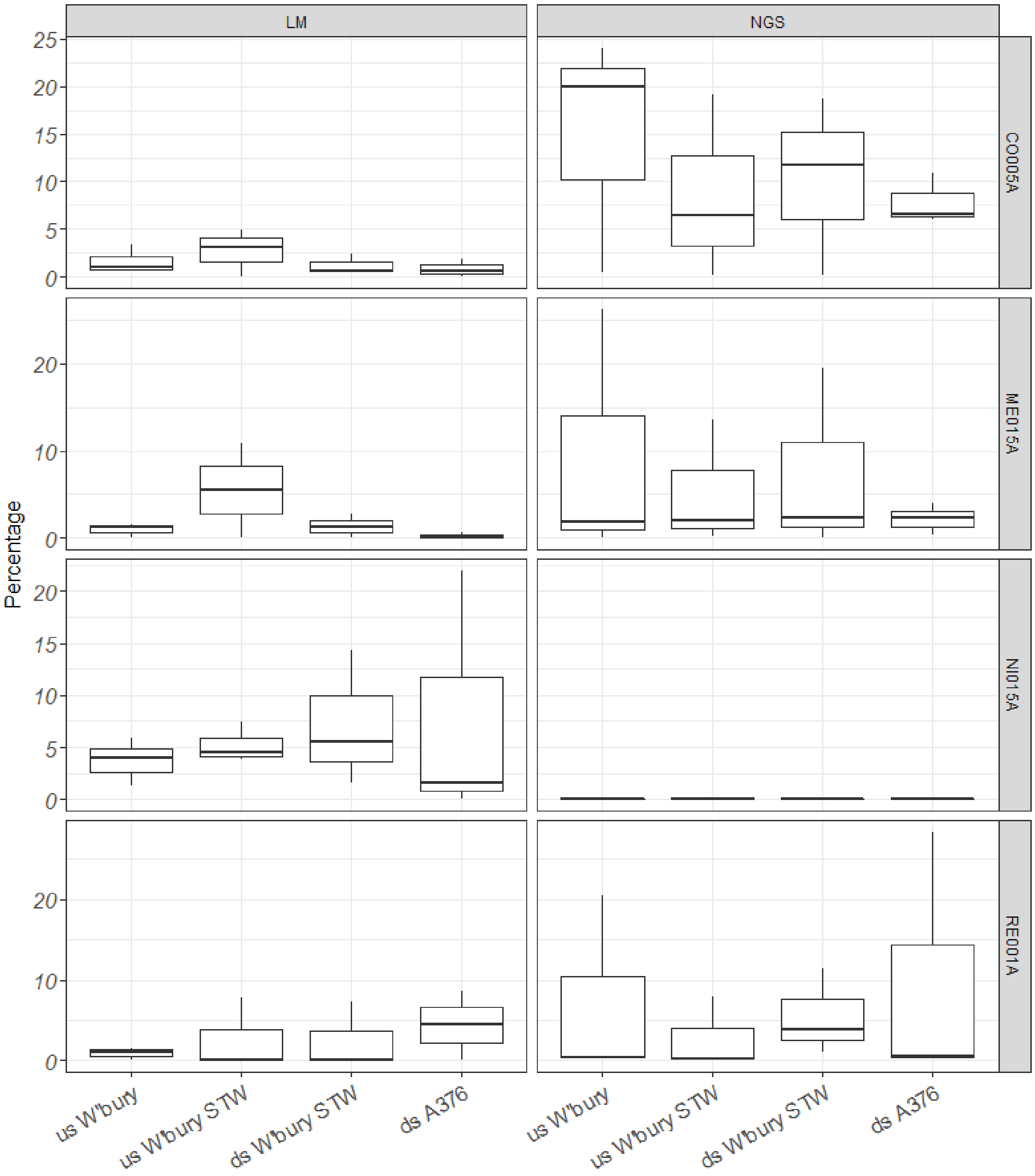

Figure 5.6b Distribution of common taxa in Polly Brook by reach: *Cocconeis pediculus* (CO005A); *Melosira varians* (ME015A); *Nitzschia dissipata* (NI015A); and *Reimeria sinuate* (RE001A)

Notes: LM = proportion recorded by light microscopy; NGS = proportion recorded by next generation sequencing

(c)

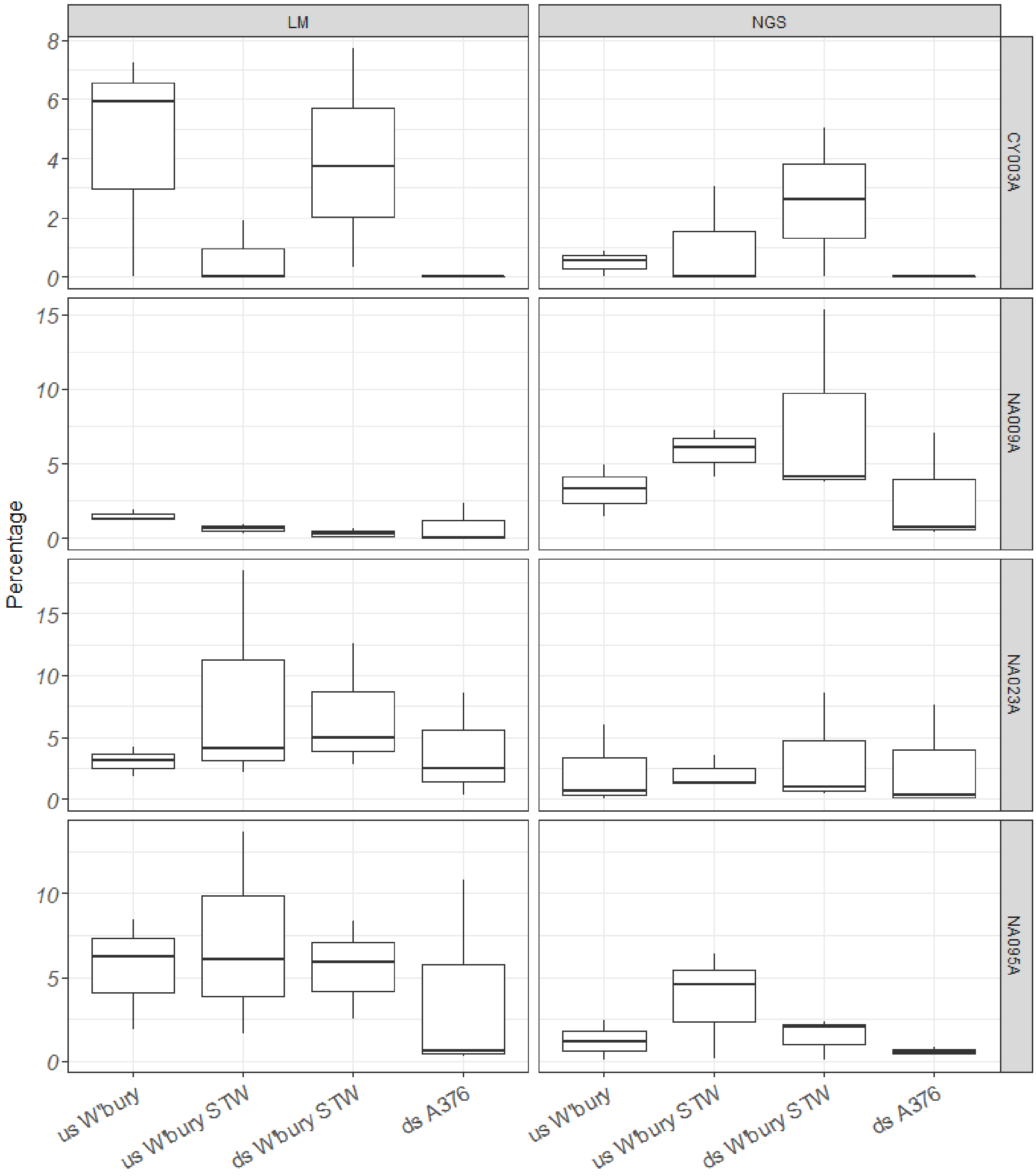

Figure 5.6c Distribution of common taxa in Polly Brook by reach: *Cyclotella meneghiniana* (CY003A); *Navicula lanceolata* (NA009A); *Navicula gregaria* (NA023A); and *Navicula tripunctata* (NA095A)

Notes: LM = proportion recorded by light microscopy; NGS = proportion recorded by next generation sequencing

(a)

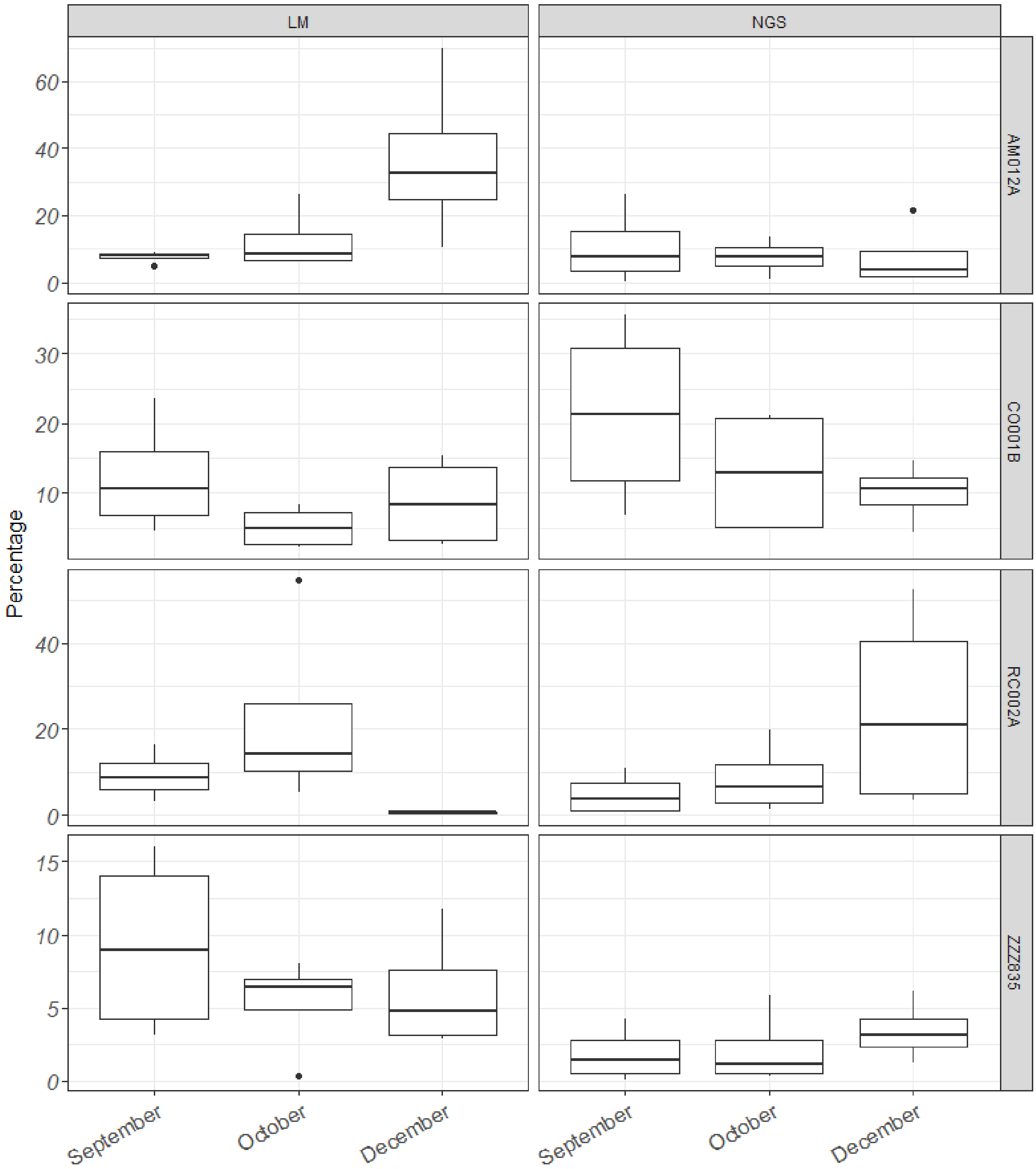

Figure 5.7a Distribution of common taxa in Polly Brook by month: *Amphora pediculus* (AM012A); *Cocconeis placentula* group (CO001B); *Rhoicosphenia abbreviate* (RC002A); and *Achnanthidium minutissimum* (ZZZ835)

Notes: LM = proportion recorded by light microscopy; NGS = proportion recorded by next generation sequencing

(b)

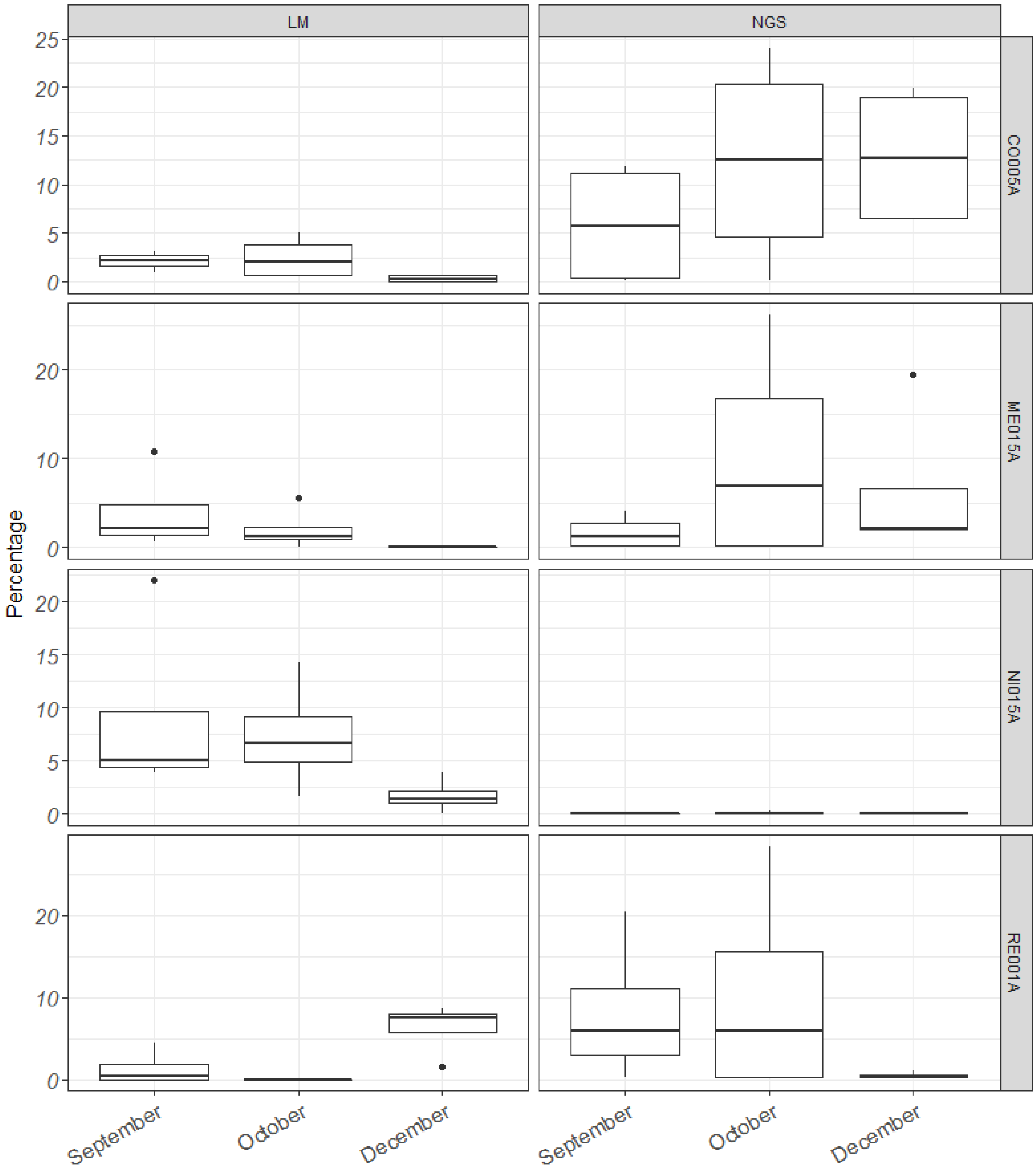

Figure 5.7b Distribution of common taxa in Polly Brook by month: *Cocconeis pediculus* (CO005A); *Melosira varians* (ME015A); *Nitzschia dissipata* (NI015A); and *Reimeria sinuate* (RE001A)

Notes: LM = proportion recorded by light microscopy; NGS = proportion recorded by next generation sequencing

(c)

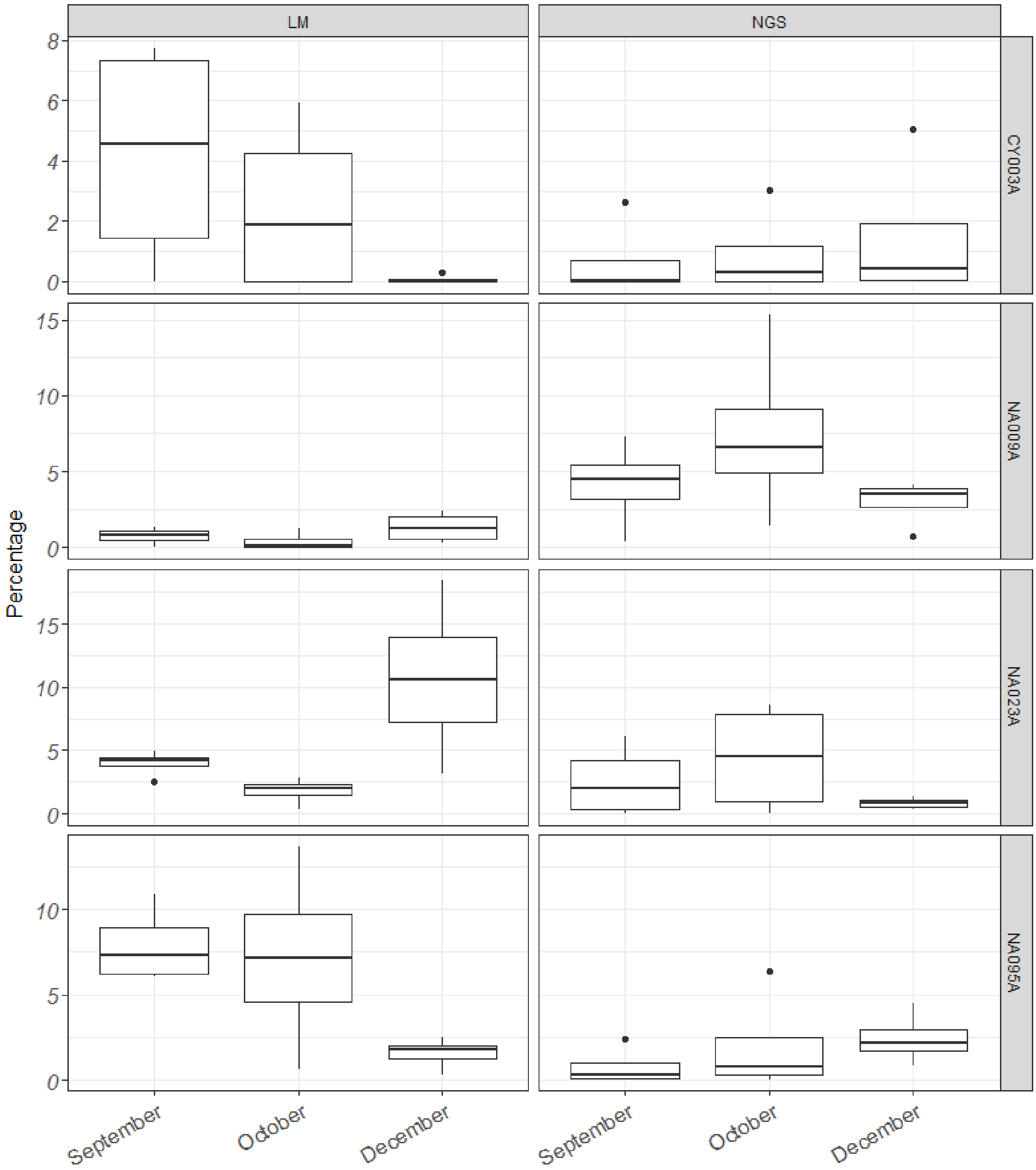

Figure 5.7c Distribution of common taxa in Polly Brook by month: *Cyclotella meneghiniana* (CY003A); *Navicula lanceolate* (NA009A); *Navicula gregaria* (NA023A); and *Navicula tripunctata* (NA095A)

Notes: LM = proportion recorded by light microscopy; NGS = proportion recorded by next generation sequencing

The bioinformatics routines will assign an operational taxonomic unit (OTU) that does not match a sequence in the database to the closest match in the database. Provided agreement exceeds 95%, some of the differences between LM and NGS may represent these ‘near misses’. Figure 5.8 shows the scale of genus level differences for the 5 samples with the greatest variability. Some general tendencies are apparent.

Taxa that are more abundant in LM than in NGS

- Case 1 – for example, *Amphora pediculus* and *Achnanthidium minutissimum* complex. These patterns were recognised in earlier phases of the work (see Section 6 of Environment Agency 2018) and are thought to be due to a combination of small cells and single chloroplasts. In both cases (but more likely for *Amphora pediculus*), it is possible that the full genetic breadth has not yet been captured in the barcode database and this may also be a contributory factor. However, the barcode database should contain enough breadth to assign unknowns to the correct genus, even if rarer species are absent from the database.

- Case 2 – for example, *Nitzschia dissipata, Navicula tripunctata, Navicula gregaria* and *Cyclotella meneghiniana.* Reasons for differences are harder to explain for some other taxa. *Nitzschia dissipata, Navicula tripunctata* and *Navicula gregaria* all have moderate size cells with 2 chloroplasts, while *Cyclotella meneghiniana* has a moderate sized cell with multiple chloroplasts. Once again, it is possible that the barcode database does not capture all of the diversity within these species.

Amphora, Achnanthidium and *Nitzschia* were all better represented in LM than NGS in the examination of genus level differences (Figure 5.8).

Taxa that are more abundant in NGS than in LM

- Case 3 – for example, *Cocconeis placentula* complex, *Cocconeis pediculus* and *Navicula lanceolata Melosira varians.* This is the opposite situation to Case 2 above. The behaviour of *Melosira varians* (moderate to large cell size, many chloroplasts) is already understood as is, to a lesser extent, that of *Navicula lanceolata* (see Section 6). The *Cocconeis* spp. both have moderate sized cells and single chloroplasts.

- Case 4 – *Reimeria sinuata.* This was unexpected as *Reimeria* cells are relatively small (similar in size to *Achnanthidium minutissimum* and *Amphora pediculus*) and have a single chloroplast. However, LM records did reveal a second species of *Reimeria uniseriata* in some samples at a similar level of abundance to *R. sinuata*, so it is possible that this was being assigned to *R. sinuata* by the bioinformatics. Other genera which showed this trend (Figure 5.8) were *Gomphonema, Planothidium*, *Fistulifera, Mayamaea* and *Surirella*. It is possible that *Fistulifera* does not survive the LM preparation process and that NGS is, in this case, giving a more accurate indication of the relative abundance than LM. *Melosira varians* was present in samples, but generally at low levels in both LM and NGS.

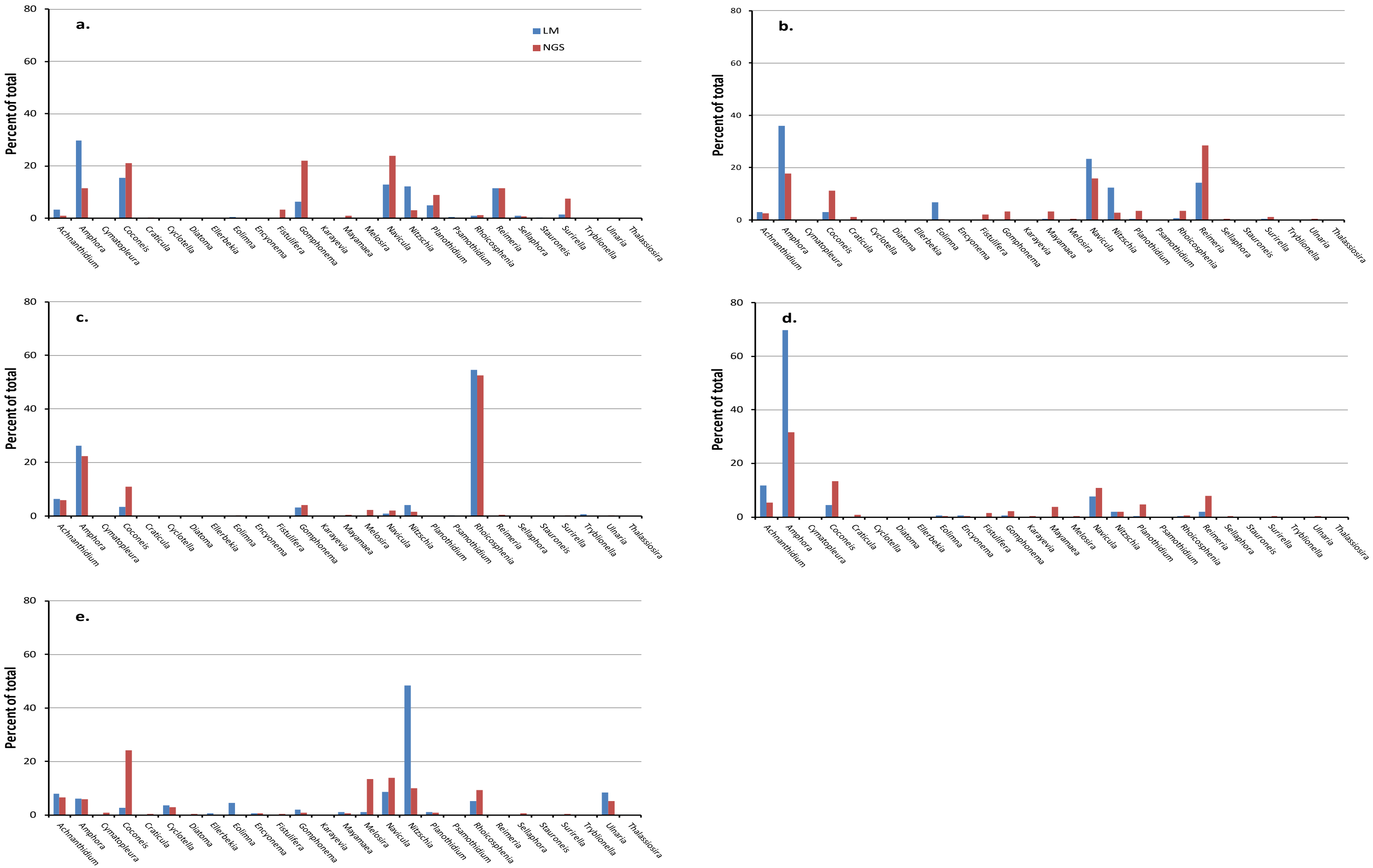

Figure 5.8 Variation between NGS and LM results for genera in the 5 samples with variation >7 TDI units: (a) u/s Woodbury, December 2016 (difference: 12 units); (b) d/s A376 bridge, December 2016 (13 units); (c) u/s Woodbury, October 2016 (15 units); (d) d/s Woodbury STW, December 2016 (17 units); and (e) u/s Woodbury STW, October 2016 (24 units)

5.4 Discussion

The study was initiated to investigate the effect of the Woodbury STW on diatom assemblages as assessed by LM and NGS. As was the case for the River Browney (see Section 8 of Environment Agency 2018), this small study involved a trade-off between insights gained from examining a small catchment in detail and limitations created by spatial and temporal autocorrelation within the dataset.

Nutrient concentrations in the river upstream of the STW may exceed the thresholds that support good ecological status for phosphorus; nitrogen levels are also very high. Diatom assemblages are therefore already compromised, largely as a result of the rich agricultural land and septic tanks upstream of the village, as well as by storm overflows in the village itself. The hoped-for step change from 'good' to 'moderate or lower' ecological status was, as a result, not observed in the data. A further confounding factor was the shift in the composition of the diatom assemblage between the October and December sampling visits, with the December samples, irrespective of site, having an assemblage characteristic of winter and early spring conditions.

However, these factors do not explain the poor match between LM and NGS data, whether viewed as relative abundance of particular taxa (Figures 5.6 and 5.7) or TDI scores (Figures 5.4 and 5.5). Some of the differences observed (for example, *Achnanthidium minutissimum* and *Amphora pediculus*) follow patterns already understood from earlier phases of the project. Some of the patterns shown by other taxa (for example, *Cocconeis* spp.) may be manifestations of similar systematic trends in read numbers that need to be corroborated from other data sources. Other differences, particularly the occasional contradictory behaviour of *Rhoicosphenia abbreviata* with the 2 methods, are less easy to explain. The general decline in relative abundance of this species (observed in both LM and NGS) from autumn to winter is to be expected, as this taxon is a common epiphyte on *Cladophora* (a similar trend is also shown for *Cocconeis pediculus*), but the occasional outliers where one method gave much higher values than the other are problematic.

The possibility of occasional 'misfires' at the DNA extraction and amplification stages cannot be ruled out. However, ongoing work at the Food and Environment Research Agency (FERA) has shown that amplification bias is unlikely to be a significant reason for the observed differences between NGS and LM, with newer sequencing technologies and different diatom rbcL primer pairs giving consistent results for samples and LM being the major difference. The case of *Reimeria sinuata/uniseriata* shows that there are situations, even within a genus, where sequences may not be picked up by the barcode database. Ongoing attention to possible reasons for major discrepancies between relative abundance of particular taxa does therefore seem to be a sensible precaution.

The interpretation of trends between LM and NGS data disguise local variation in the composition of the assemblage that probably reflects a variety of ecological and physiological processes working together. It is suspected that the LM data largely reflect ecological changes (that is, in numbers of individuals) in relation to a number of natural and pressure-related factors, but that NGS data include an extra layer of complexity, reflecting adaptation of the photosynthetic apparatus of cells to local circumstances (see Sections 6.3 and 6.4, and Appendix 4).

This, in turn, emphasises the importance of regarding LM and NGS data as 2 alternative means of looking at stream phytobenthos. Indeed, it is possible that focusing on a gene controlling expression of an important component of the photosynthetic pathway is a more sensitive means of looking at primary productivity in streams than simply recording presence or relative abundance of species – the largest and smallest of whom may have cells that differ in volume by up to 2 orders of magnitude. The problem is that biologists are still learning how to understand this new form of data.

Overall, this case study sounds a warning note, particularly when the method is used as part of operational investigations within a limited area. Whereas the consequences of variations in individual taxa are damped down when considering large-scale spatial datasets, such variation in space and time may be difficult to explain at a local level. That such variation occurs in both LM and NGS data suggests that it may, ultimately, be a consequence of using a group of fast-growing organisms. Neither method suggested unimpacted conditions, with typical good ecological status indicators such as *Achnanthidium minutissimum* being relatively uncommon throughout the study. A likely explanation for most subsequent changes is that the proportions of a selection of taxa adapted to thriving in nutrient-rich conditions were shuffled by a range of (largely) non-pressure related factors. That ‘signal’ was, to a large extent, overwhelmed by this ‘noise’ is further complicated by the choice of 2 different ‘lenses’ through which this was observed.

6 Discussion

The SC140024 project report (Environment Agency 2018) described a large-scale proof-of-concept study that demonstrated that it was possible to assess the ecological status of rivers using data generated by NGS rather than LM analysis of benthic diatoms.

The subject of this report, project SC160014, developed the prototype NGS metric using a larger database and larger barcode database (Section 2) and the possibility of extending the approach to lakes as well as rivers is considered (Section 4). Work on adapting sampling protocols to make them suitable for NGS is also reported (Section 3), along with a case study that examines similarities and differences in LM and NGS data and metrics within a small catchment (Section 5). This provides a useful opportunity to understand the challenges that biologists will face as they move from interpreting LM to NGS data.

6.1 Development of an NGS-compatible metric

The prototype NGS metric, reported in Environment Agency (2018), mostly gave very similar results to those obtained using the current LM based TDI4 approach (Pearson's correlation coefficient, r = 0.90; Lin's concordance correlation coefficient, r = 0.89; root mean square error = 9.3). This was despite having a barcode database that contained less than 10% of all freshwater diatoms recorded from Britain and Ireland. The possibility of improving this fit by adding extra taxa to the barcode database was one motivation for the present project.

But despite having twice as many taxa represented in the barcode database, as well as more paired LM and NGS samples, optimisation of the relationship between the LM metric and pressure, and modified statistical procedures, there is little change in the relationship between the LM and NGS metrics (Table 2.9). However, the bias between a classification obtained using the new metric (TDI5NGS) compared with that obtained using TDI4 reduced from 21% (Environment Agency 2018) to <3% (Table 2.11).

When classifications obtained using the optimised LM metric (TDI5LM) and TDI5NGS are compared, the bias is similar to that between TDI4 and TDI5NGS (Table 2.12). Optimisation of the relationship between the LM metric and pressure provides some important context for understanding the bias values. Identifying and removing anomalies (such as the high indicator value for *Adlafia suchlandtii* in TDI4) would provide more confidence in the classifications produced by TDI5LM and hence any 'bias' actually represents weaknesses in the existing metric. The effect of any remaining bias is placed in context by the analyses presented in Section 2.7 in which the reference model (the denominator in EQR calculations) is adjusted. Moving from the current reference model to a plausible alternative model results in much greater bias than observed simply by switching between LM and NGS.

It is important, however, to recognise that there are still differences between LM and NGS outputs even after optimisation of both models (Figure 2.9, Table 2.4).

The barcode database is likely to be significant in this, even though the number of missing taxa that are both frequent and sometimes abundant is relatively small. In particular, the *Gomphonema olivaceum / olivaceoides / calcifugum* complex remains a key 'gap' in coverage and addition of more representatives of these is likely to help.

A second issue is whether the barcode sequences that are in the database provide full coverage of the genetic variability of the species that they represent. Figure 6.10 in Environment Agency (2018) shows a very strong relationship between TDI4 calculated with all taxa and with just those represented in the barcode database (r = 0.99), though this assumes that representation in the barcode database equates to detection of the species in NGS analyses. In theory, the bioinformatics routines should assign barcodes that are not exact matches to anything in the database to a nearest neighbour so long as the sequence similarity is >95%. Agreement at genus level even if not at species level should therefore be assumed. But when LM and NGS data to genus for samples from Polly Brook were aggregated (Figure 5.8), there were still some striking differences. *Nitzschia*, for example, is often over-represented in LM despite relatively good representation in the barcode database and *Reimeria* was often more abundant in NGS than LM, possibly because *R. uniseriata*, a less common species not represented in the barcode database, is being allocated to *Reimeria sinuata.*

These observations also need to be interpreted in light of the large number of NGS reads that could not be assigned at all (Figure 2.1). The proportions (average 40%) are such that it is unlikely that these all constitute chimeras and other low grade DNA. It may therefore be informative to analyse the distribution of OTUs from which the Polly Brook data are composed in relation to the phylogenetic structure of the barcode database so as to further understand this.

6.2 Ongoing development of the barcode database

An important lesson learned from this project is that it is important not to judge the effectiveness of a barcode database solely by the number of species represented. This will be particularly true for any methods based on statistical procedures such as weighted averaging, where abundant taxa make the greatest contribution to the final result. The absence of such taxa is likely to be detrimental to the sensitivity of the method. However, simply adding barcode sequences of rare taxa or multiple different barcodes of common taxa is not necessarily going to lead to improvements. Several factors may contribute to this include:

- the procedures by which non-exact matches are assigned to species
- the breadth of genetic variation within a morphologically defined species (bearing in mind that this does not necessarily represent a biological species)
- the scale of divergence between species and genera (some genera are much larger and older and more internally divergent groupings than others)

Moreover, some genera are not monophyletic. For example, though *Nitzschia* and *Sellaphora* are relatively well-represented, a relaxation of the threshold that allows more sequences to be assigned to *Nitzschia* would quite possibly lead to misidentifications of *Fallacia* and *Sellaphora*, these having diverged from their common ancestor much more recently. Since *Nitzschia* is not a monophyletic group and representation in the barcode database is not spread evenly through this group, relaxation here might lead to misidentifcations in relation to, for example, *Bacillaria, Cylindrotheca,* pseudo *Nitzschia*, and *Fragilariopsis*.

An alternative would be to match via phylogenetic position rather than phenetic similarity, that is, to construct phylogenetic trees of the sample and reference sequences, using model based maximum likelihood or Bayesian approaches (see, for example, Yang 2006). However, phylogenetic assignment is currently impractical. Overall, it is probably better to keep the matching criteria strict and live with data loss in the short term, and to work to increase the evenness and depth of sampling represented in the barcode database. In the absence of funding to develop the barcode database in such a structured manner, an intention to upgrade it with published sequences might well lead to incremental improvement (it will certainly not reduce the efficacy of the method). While the need for strong and stable environmental regulation is recognised, it is important to differentiate between changes that improve the precision of the method from any that influence the position of ecological status class boundaries. The latter changes may have consequences for regulation, but it should be possible to have straightforward routines that improve the sensitivity of the method while not compromising stability within a river basin management planning cycle.

An alternative approach would be to bypass traditional taxonomy altogether. The barcode database used in this study contains just over 10% of all freshwater diatoms recorded from Britain and Ireland, and it is likely that the true number of diatom species is much greater than this (see Mann and Vanormeligan 2013). The implication is that many diatom 'species' likely to be encountered in the UK are yet to be discovered and so it will be impossible to match OTUs representing these species to traditional Linnaean binomials. It may therefore be worth considering a hybrid approach, exploiting the potential for a taxonomy-free approach like that proposed by Apothéloz-Perret-Gentil et al. (2017) while retaining traditional taxonomic assignments wherever possible so as to utilise existing autecological understanding.

6.3 Reasons for differences between LM and NGS metrics

Although increasing the coverage of the barcode database should improve the match between LM and NGS metrics, it is important to recognise that these are 2 fundamentally different types of data, possibly offering 2 different viewpoints of the state of the diatom assemblage. It would be wrong to assume that LM is automatically 'right', and that all error and uncertainty lie within the NGS data. On the other hand, some of the differences between the 2 approaches (particularly those found in Section 5), involve taxa that are easily recognised under the light microscope and are unlikely to be mistaken with other taxa present in the sample. As this is a long-established method, there is a case to answer when NGS produces very different results. Overall, however, LM and NGS offer complementary insights into the biological species concept as applied to diatoms.

The primary use of benthic diatoms for ecological status assessment in Europe at present is to characterise biological responses to the nutrient pressure gradient (Kelly 2013). LM does this by counting cell walls while NGS counts copy numbers of a key photosynthetic gene. LM is an established technique whose practitioners are experienced in interpreting differences in the relative abundance of species in terms of nutrient pressure. However, it is possible that NGS is actually providing a more direct measure of the relative contribution of each taxon to primary productivity. There are, however, 2 problems as outlined below.

Both methods measure relative, rather than absolute, abundance and thus have a limited capacity to predict secondary effects (‘undesirable disturbances’) of eutrophication, the absence of which forms part of the normative definition of ‘good ecological status’ (Poikane et al. 2016). Schneider et al. (2016) refer to composition based approaches as measuring ‘eutrophication potential’ rather than eutrophication per se. In theory, there is no reason why either method could not be made fully quantitative; in practice, spatial and temporal variability of benthic biofilms complicates the issue.

The relationship between rbcL reads and cell numbers appears to be far from straightforward. There is some evidence (Vasselon et al. 2017) of a relationship between cell size and read number, and Environment Agency (2018) suggested that the number of chloroplasts also plays a role. Little is known about the relationship between the number of rbcL reads and the number of cells, and this may vary even with a single species. However, the number of reads per cell is unlikely to bear a simple relationship to chloroplast or cell size, and by analogy with what occurs in green plants (see, for example, Rauwolf et al. 2010, Liere and Börner 2013, Kabeya and Miyagishima 2013), it is likely that there will also be variation within a single species according to growth conditions and the life cycle stage.

In other words, even if there are difficulties in relating rbcL reads to cell numbers, it is by no means clear that rbcL reads are an intrinsically poorer measure of diatom abundance and activity than cell counts, especially given that, in WFD monitoring, cell counts are based on material in which it is impossible to determine which cells were alive and which dead at the time of sampling.

6.4 Sources of bias in metabarcoding analyses of diatoms

There are many potential areas of bias introduction within the metabarcoding workflow that could explain some of the differences between NGS and LM.

DNA extraction techniques have been shown to influence species detection in aquatic systems (Deiner et al. 2015). It is thought that the lysis method in particular affects the DNA extraction efficiency (Deiner et al. 2015, Vasselon et al. 2017). The Qiagen DNeasy® Blood & Tissue kit, which uses an enzymatic lysis step and which was used in the workflow for this project and project SC140024, is designed for eukaryotic cell lysis and causes less damage to DNA compared with mechanical lysis methods (Deiner et al. 2015). Tests of the Qiagen DNeasy® Blood & Tissue kit showed that it gave high average and consistent amounts of purified DNA (Environment Agency 2018).

Vasselon et al. (2017) tested 5 different extraction approaches which included the Qiagen DNeasy® Blood & Tissue kit and recommended a combined lysis (sonication, enzymatic and temperature) method for diatom assessment using metabarcoding. However, they did find that:

- all the extraction methods tested provided DNA of sufficient quality and quantity
- the composition of diatom assemblages was not affected by the choice of extraction method when compared with the composition assessed using microscopy

The relative abundance of some taxa did vary, but the variability did not affect the pollution indices calculated for water quality assessment. Vasselon et al. (2017) hypothesised that diatom species with long and thin skeletons may be more easily broken by mechanical lysis than small species with thick skeletons, thus affecting their relative representation in the metabarcoding outputs. This, however, should not be a factor in this study as mechanical lysis was not used. Vasselon et al. (2017) also observed, although data were not shown, that small species (<20µm length) were proportionally less represented in the metabarcoding data than in the microscopy data, whereas species >50µm long appeared to be proportionally more abundant in the metabarcoding data.

PCR inhibition by both organic and inorganic substances (for example, humic and fulmic acids and metal ions) within an environmental sample can decrease sensitivity and cause false-negative results (Schrader et al. 2012). However, this can be controlled by the addition of substances such as bovine serum albumin (BSA) to the PCR protocol.

The efficiency of PCR is sensitive to the primers used. Primers should be able to amplify their target DNA barcode efficiently in the presence of non-target DNA and potential inhibitors. Differences in primer efficiency and robustness can result in strong bias in favour of more easily amplified sequences during PCR, which can potentially give a skewed representative of the community composition (Op De Beeck et al. 2014). However, the project team believe that amplification bias due to the primers developed and tested in Environment Agency (2018) is unlikely to be a significant reason for the observed differences between NGS and LM. This is because work using newer sequencing technologies and different full length diatom rbcL primer pairs gave consistent results for the samples tested (unpublished data).

Sequencing errors can also arise from artefacts in the PCR such as the production of chimeric sequences, which can occur during PCR amplification and result in sequences that may be partly one species and partly another, or from Taq DNA polymerase (Acinas et al. 2005). Chimeric sequences and singletons (OTUs present only once across the entire dataset) may represent artefacts of the PCR and are usually removed prior to bioinformatic analysis. The project team decided not to remove chimeras as it did not have a ‘chimerafree' barcode database to identify them. Taxonomic-free chimera checking has quite a high false positive rate (as noted in Environment Agency 2018) and removal of false positive 'chimeric' OTUs may well have affected the TDI. In order to perform chimera checking on the OTUs prior to downstream analysis, all DNA barcodes in the reference database would need to be checked to ensure they themselves are not chimeric PCR products before using a taxonomy based (that is, BLAST based) chimera detection method. The removal of singletons is controversial as some may not be errors and may constitute true OTU occurrence, and so it was decided not to remove them as they are likely to have no impact on the TDI metric.

Primer efficiency, PCR amplification and sequencing error bias are not easily estimated and corrected for unless control material is introduced as an internal standard (Schrader et al. 2012, Vasselon et al. 2017). This procedure is generally accepted as standard practice and it could be argued that it should have been followed in this project and PCR replication built into the workflow to reduce bias. To remove PCR bias, most studies perform triplicate amplifications and pool each sample prior to sequencing. However, Smith and Peay (2014) questioned the benefit of doing this. They found that PCR replication and pooling has no detectable effect on biodiversity measures and may be a poor investment in resources. In their study, the number of sequences observed between low replicate PCRs (for example, 1) and high (for example, 16) was highly correlated in a 1:1 relationship. This suggested that:

- pooling prior to sequencing does not affect the relative abundance of taxa found in each sample
- the sequencing abundance in one PCR can accurately predict sequence abundance per taxon in a pool of 16 replicates

There may be areas of the current workflow that could be improved to further minimise bias. However, what was developed is a pragmatic approach to diatom assessment that balances the need to produce a cost-effective method while providing good resolution to diatom assessment compared with the traditional LM approach.

6.5 Interpreting diatom composition data generated by NGS

This project sought to develop a molecular analogue of the current LM based approach. This provides stability for environmental regulation but, more importantly, creates an opportunity to learn about how outputs from the 2 methods differ. The approach encapsulated in Section 2 of this report is to reconcile the metrics generated from the 2 forms of data while respecting the integrity of the underlying data. This is preferred to applying 'correction factors' to convert relative abundances of individual taxa in NGS to an equivalent LM value. As discussed above, there is no good reason for believing that such an approach will generate better insights into the condition of the stream algal flora.

Diatoms play an important role in generating national classifications of ecological status. For these, diatom data are processed using standard algorithms and species level information plays no particular role other than in providing the feedstock for a largely automated process. For this, a high level of agreement and low level of bias between LM and NGS means that concepts of ecology status should be transferable. When developing programmes of measures for rivers that do not achieve good ecological status, however, Environment Agency biology staff may need to interpret species level data in order to advise on the suitability of particular measures. Particularly during the changeover from LM to NGS, biologists will need to understand how individual taxa are expressed between the 2 methods when interpreting samples; this is considered in more detail in Appendix 4.

6.6 Conclusions

This project has built on the foundations provided by project SC140024 (Environment Agency 2018). It has developed a tool for the assessment of ecological status in rivers that provides classifications that are compatible with those generated by the present LM based tool.

Although the project team is satisfied that TDI5NGS is a good analogue for the existing LM based approach, there are still concerns about the large number of reads that are not yet assigned to species in the barcode database. The project team believes that the most likely way to improve the method in the short or medium term is a better understanding of:

- the nature of these unassigned reads
- the factors affecting the rbcL read numbers for a species under any set of conditions

As the project is the first of its kind in Europe to deliver an operational tool, there is much to learn from the outputs. An important message to take forward is that NGS data are fundamentally different to those generated by traditional means and that this, in turn, requires adaptation at several levels throughout an organisation. At a practical level, this may necessitate modifying procedures (see Section 3), modifying databases and writing new software to generate classifications. However, there are also broader implications in terms of retraining staff to understanding the new types of data, and particularly in how they differ in form from that which they are used to interpreting.

NGS is a field that is developing quickly and one challenge that regulators need to face is how to accommodate refinements and improvements into a tool while, at the same time, ensuring a stable regulatory environment. These aspects are not incompatible, but NGS will require a more flexible attitude than has been the case until now, both to improve the precision of regulatory decisions and to ensure that an adopted method does not depend on outdated technology.

References

ACINAS, S.G., SARMA-RUPAVTARM, R., KLEPAC-CERAJ, V. AND POLZ, M.F., 2005. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, 71 (12), 8966-8969.

APOTHÉLOZ-PERRET-GENTIL, L., CORDONIER, A., STRAUB, F., ISELI, J., ESLING, P. AND PAWLOWSKI, J., 2017. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular Ecology Resources*, 17 (6), 1231-1242.

BENNION, H., KELLY, M.G., JUGGINS, S., YALLOP, M.L., BURGESS, A., JAMIESON, 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. AND SIMPSON, G.L., 2013. 'Diatoms and pH reconstruction' (1990) revisited. *Journal of Paleolimnology*, 49 (3), 363-371.

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*, 327 (1240), 263-278.

BISTA, I., CARVALHO, G.R., WALSH, K., SEYMOUR, M., HAJIBABAEI, M., LALLIAS, D., CHRISTMAS, M. AND CREER, S., 2017. Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nature Communications,* 8, article number 14087.

BYRD, R.H., LU, P., NOCEDAL, J. AND ZHU, C., 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing*, 16 (5), 1190-1208.

CANTONATI, M. AND LOWE, R.L., 2014. Lake benthic algae: towards an understanding of their ecology. *Freshwater Science*, 33 (2), 475-486.

CEN, 2014. *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.

DEINER, K., WALSER, J.-C., MÄCHLER, E. AND ALTERMATT, F., 2015. Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, 183, 53-63.

ENVIRONMENT AGENCY, 2012. *A streamlined taxonomy for the Trophic Diatom Index.* Report SC070034/TR1. Bristol: Environment Agency.

ENVIRONMENT AGENCY, 2018. *A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers*. Report SC140024/R. Bristol: Environment Agency.

KABEYA, Y. AND MIYAGISHIMA, S., 2013. Chloroplast DNA replication is regulated by the redox state independently of chloroplast division in *Chlamydomonas reinhardtii*. *Plant Physiology*, 161 (4), 2102-2112.

KAHLERT, M. AND GOTTSCHALK, S., 2014. Differences in benthic algal assemblages between streams and lakes in Sweden and implications for ecological assessment. *Freshwater Science*, 33, 655-669.

KELLY, M.G., 2013. Data rich, information poor? Phytobenthos assessment and the Water Framework Directive. *European Journal of Phycology*, 48 94), 437-450.

KELLY, M.G., 2016. *This is not a nitrate standard …* [online]. Available from: https://microscopesandmonsters.wordpress.com/2016/12/18/this-is-not-a-nitrate-standard/ [Accessed 11 June 2018].

KELLY, M.G. AND ZGRUNDO, A., 2013. Potential for cross-contamination of benthic diatom samples when using toothbrushes. *Diatom Research,* 28 (4)*,* 359-363.

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.

LIERE, K. AND BÖRNER, T., 2013. Development-dependent changes in the amount and structure of plastid DNA. In *Chloroplast Development during Leaf Growth and Senescence*, Advances in Photosynthesis and Respiration, Vol. 36 (edited by B. Biswal, K. Krupinska, and U.C. Biswal), pp. 215-237. Dordrecht, the Netherlands: Springer.

MANN, D.G. AND VANORMELINGEN, P., 2013. An inordinate fondness? The number, distributions and origins of diatom species. *Journal of Eukaryotic Microbiology*, 60 (4), 437-495.

MYUNG, I.J., 2003. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47 (1), 90-100.

OP DE BEECK, M., LIEVENS, B., BUSSCHAERT, P., DECLERCK, S., VANGRONSVELD, J. AND COLPAERT, J.V., 2014. Comparison and validation of some ITS primer pairs useful for fungal metabarcoding studies. *PLoS ONE*, 9 (6), e97629.

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, 603-613.

RAUWOLF, U., GOLCZYK, H., GREINER, S. AND HERRMANN, 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].

SCHNEIDER, S., HILT, S., VERMAAT, J.E. AND KELLY, M., 2016. The ‘forgotten ecology’ behind ecological status evaluation: re-assessing the roles of aquatic plants and benthic algae in ecosystem functioning. *Progress in Botany*, 78, 285-304.

SCHRADER, C., SCHIELKE, A., ELLERBROEK, L. AND JOHNE, R., 2012. PCR inhibitors – occurrence, properties and removal. *Journal of Applied Microbiology*, 113 (5), 1014-1026.

SMITH, D.P. AND PEAY, K.G., 2014. Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS ONE*. 9 (2), e90234.

TER BRAAK, C.J.F., 1986. Canonical Correspondence Analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67 (5), 1167-1179.

VASSELON, V., DOMAIZON, I., RIMET, F., KAHLERT, M. AND BOUCHEZ, A., 2017. Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? *Freshwater Science*, 36 (1), 162-177.

YANG, Z., 2006. *Computational Molecular Evolution.* Oxford: Oxford University Press.

List of abbreviations

BLAST	Basic Local Assignment Search Tool
BOD	biochemical oxygen demand
DARLEQ	Diatoms for Assessing River and Lake Ecological Quality
DAM	Diatom Acidification metric
DNA	deoxyribonucleic acid
EQR	Ecological Quality Ratio
GAM	generalised additive model
LDTI	Lake Trophic Diatom Index
LM	light microscopy
NGS	next generation sequencing
NMDS	non-metric multidimensional scaling
OTU	operational taxonomic unit
PCR	polymerase chain reaction
rbcL	ribulose bisphosphate carboxylase large chain [gene]
RC	response curve
SSO	storm sewage overflow
STW	sewage treatment works
TDI	Trophic Diatom Index
TN	total nitrogen
TP	total phosphorus
WA	weighted average
WFD	Water Framework Directive

Glossary

Prototype NGS	The first NGS based variant of the TDI, as described in Environment Agency (2018) as ‘TDI5’
TDI4	Current LM based variant of the TDI
TDI5LM	A recalibrated variant of TDI4, for use with LM data
TDI5NGS	The NGS based metric derived from TDI5LM using WA in combination with a monotonic GAM rescaling procedure
TDI5NGS original	A version of the NGS based metric resulting from initial analyses but dropped in favour of TDI5NGS
Species/taxon indicator value	A value that represents the sensitivity of a taxon across the nutrient gradient (1 = sensitive; 5 = tolerant)
TDI	Unless otherwise qualified, this refers to the value computed for an individual sample

Appendix 1: Diatom taxa represented in the barcode database

Taxon	**Authority**	**Number of barcodes** [1]
Achnanthes_coarctata	(Brébisson in W. Smith) Grunow in Cleve and Grunow 1880	1
Achnanthidium_caledonicum	(Lange-Bertalot) Lange-Bertalot 1999	2
Achnanthidium_kranzii	(Lange-Bertalot) Round and L.Bukhtiyarova 1996	1
Achnanthidium_lineare	W. Smith 1855	1
Achnanthidium_minutissimum	(Kützing) Czarnecki 1994	85
Achnanthidium pyrenaicum	(Hustedt) Kobayasi 1997	10
Achnanthidium_reimeri	(Camburn) Ponader and Potapova 2007	1
Achnanthidium_rivulare	Potapova and Ponader 2004	1
Achnanthidium_saprophilum	(Kobayasi and Mayama) Round and Bukhtiyarova 1996	1
*Achnanthidium*_sp.	Kützing 1844	1
*Actinocyclus*_sp.	Ehrenberg 1837	1
Adlafia_bryophila	(Petersen) Lange-Bertalot In Moser et al. 1997	1
Adlafia_minuscula	(Grunow) Lange-Bertalot in Lange-Bertalot and Genkal 1999	2
Amphipleura_pellucida	(Kützing) Kützing 1844	1
Amphora_aff._atomoides	Levkov 2009	1
Amphora_berolinensis	N. Abarca and R. Jahn in Zimmermann et al. 2014	1
Amphora_ovalis	(Kützing) Kützing 1844	5
Amphora_pediculus	(Kützing) Grunow in Schmid et al. 1875	9
Asterionella_formosa	Hassall 1850	3
Aulacoseira_granulata	(Ehrenberg) Simonson 1979	2
Bacillaria_paxillifer	(O.F. Müller) Hendey 1951	2

Taxon	**Authority**	**Number of barcodes** [1]
Brachysira_microcephala	(Grunow) Compère 1986	1
Brachysira_neoexilis	Lange-Bertalot in Lange-Bertalot and Moser 1994	1
Brachysira_vitrea	(Grunow) R. Ross in B. Hartley 1986	1
Brebissonia_lanceolata	(C. Agardh) Mahoney and Reimer 1986	1
Caloneis_amphisbaena	(Bory) Cleve 1894	3
Caloneis_lewisii	Patrick 1945	1
Caloneis_silicula	(Ehrenberg) Cleve 1894	2
Campylodiscus_clypeus	(Ehrenberg) Ehrenberg ex Kützing 1844	2
Campylodiscus_hibernicus	Ehrenberg 1845	12
Campylodiscus_levanderi	Hustedt 1925	2
Campylodiscus_marginatus	Jurilj 1954	4
Campylodiscus_striatus	Ehrenberg ex Kützing 1844	1
Centronella_reicheltii	Max Voigt 1901	1
Cocconeis_euglypta	Ehrenberg 1854	1
Cocconeis_pediculus	Ehrenberg 1838	4
Cocconeis_placentula	Ehrenberg 1838	4
Cocconeis_stauroneiformis	(W. Smith) H. Okuno 1957	1
Conticribra_weissfloggii	(Grunow) Fryxell and Hasle 1977	3
Coscinodiscus_wailesii	Gran and Angst 1931	1
Craticula_accomoda	(Hustedt) D.G. Mann in Round et al. 1980	1
Craticula_buderi	(Hustedt) Lange-Bertalot in Rumrich et al. 2000	1
Craticula_cuspidata	(Kütxing) D.G. Mann in Round et al. 1990	2
Ctenophora_pulchella	(Ralfs ex Kutz.) Williams and Round	1
Cyclophora_tenuis	Castracane 1878	1

Taxon	Authority	Number of barcodes [1]
Cyclostephanos_dubius	(Fricke in A. Schmidt) Round	1
Cyclotella_atomus	Hustedt 1937	1
Cyclotella_bodanica	Grunow in Schneider 1878	1
Cyclotella_distinguenda	Hustedt in Gams 1928	2
Cyclotella_gamma	Sovereign 1963	1
Cyclotella_meneghiniana	Kützing 1844	15
Cyclotella_ocellata	Pantocsek 1901	1
Cyclotella_pseudostelligera	Hustedt 1939	1
Cyclotella_striata	(Kützing) Grunow in Cleve and Grunow 1880	1
Cyclotella_stylorum	Brightwell 1860	1
Cymatopleura_elliptica	(Brébisson) W. Smith 1851	2
Cymatopleura_solea	(Brébisson) W. Smith 1851	2
Cymbella_affinis	Kützing 1844	1
Cymbella_aspera	(Ehrenberg) Cleve 1894	3
Cymbella_baicalensis	Skvortzow and Meyer 1928	1
Cymbella_cf	C. Agardh 1830	1
Cymbella_cistula	(Ehrenberg) Kirchner 1878	2
Cymbella_cymbiformis	C. Agardh 1830	1
Cymbella_helvetica	Kützing 1844	1
Cymbella_heterogibbosa	H. Kobayasi and Mayama in Mayama et al. 2002	1
Cymbella_janischii	(A. Schmidt) De Toni 1891	1
Cymbella_mexicana	(Ehrenberg) Cleve 1894	1
Cymbella_proxima	Reimer in Patrick and Reimer 1975	1
Cymbella_stuxbergii	(Cleve) Cleve 1894	2
Cymbella_tumida	(Brébisson ex Kützing) Grunow in Van Heurck 1880	2
Cymbopleura_naviculiformis	(Auerswald) K. Krammer 2003	1
Denticula_kuetzingii	Grunow 1862	1

Taxon	Authority	Number of barcodes [1]
*Diatoma*_*moniliformis*	Kützing 1833	7
*Diatoma*_sp.	Kützing 1844	1
*Diatoma*_*tenuis*	Agardh 1812	4
*Diatoma*_*vulgaris*	Agardh 1812	3
*Didymosphenia*_*dentata*	(Dorogostaisky) Skvortzow and K.I. Meyer 1928	1
*Diploneis*_*subovalis*	Cleve 1894	1
*Ellerbeckia*_sp.	Crawford 1988	1
*Encyonema*_*macedonicum*	Z. Levkov, Metzeltin and S. Krstic 2006	1
*Encyonema*_*minutum*	(Hilse in Rabenhorst) D.G. Mann in Round et al. 1990	6
*Encyonema*_*minutum*_var._ *pseudogracilis*	(Cholnoky) D.B. Czarnecki 1994	2
*Encyonema*_*muelleri*	(Hustedt) D.G. Mann in Round, R.M. Crawford and D.G. Mann 1990	1
*Encyonema*_*norvegica*	(Grunow) Mayer 1947	1
*Encyonema*_*prostratum*	(Berkeley) Kützing 1844	1
*Encyonema*_*silesiacum*	(Bleisch in Rabenhorst) D.G.Mann in Round et al. 1990	4
*Encyonema*_sp.	Kützing 1833	6
*Encyonema*_*triangulum*	(Ehrenberg) Kützing 1849	1
*Encyonopsis*_*falaisensis*	(Grunow) Krammer 1997	2
*Encyonopsis*_*microcephala*	(Grunow) Krammer 1997	1
*Encyonopsis*_sp._*TN-2014*	Krammer 1997	1
*Entomoneis*_*ornata*	(J.W. Bailey) Reimer in Patrick and Reimer 1975	1
*Eolimna*_*minima*	(Gronow) Lange-Bertalot 1998	2
*Eolimna*_sp.	Lange-Bertalot and W. Schiller in W. Schiller and Lange-Bertalot 1997	2
*Epithemia*_*argus*	(Ehrenberg) Kützing 1844	2

Taxon	**Authority**	**Number of barcodes** [1]
Epithemia_sorex	Kützing 1844	1
Epithemia_turgida	(Ehrenberg) Kützing 1844	1
Eucocconeis_laevis	(Øestrup) Lange-Bertalot 1999	1
Eunotia_arcus	Ehrenberg 1837	1
Eunotia_bilunaris	(Ehrenberg) Mills 1934	8
Eunotia_cf._latitaenia	H. Kobayasi, K. Ando and T. Nagumo 1981	6
Eunotia_exigua	(Brébisson) Rabenhorst 1864	4
Eunotia_formica	Ehrenberg 1843	2
Eunotia_glacialis	Meister 1912	1
Eunotia_implicata	Norpel, Lange-Bertalot et al. 1991	1
Eunotia_minor	(Kützing) Grunow in Van Heurck 1881	4
Eunotia_naegelii	Migula 1905	1
Eunotia_pectinalis	(Kützing) Rabenhorst 1864	1
*Eunotia*_sp.	Ehrenberg 1837	1
Fallacia_cf._forcipata	(Greville) Stickle and D.G. Mann in Round, Crawford and Mann 1990	1
Fallacia_monoculata	(Hustedt) D.G. Mann in Round et al. 1980	1
Fallacia_pygmaea	(Kützing) Stickle and D.G. Mann in Round, Crawford and Mann 1990	1
*Fallacia*_sp.	A.J. Stickle and D.G. Mann in Round, Crawford and Mann 1990	1
Fistulifera_pelliculosa	(Brébisson) Lange-Bertalot 1997	2
Fistulifera_saprophila	(Lange-Bertalot and Bonik) Lange-Bertalot 1997	1
Fistulifera_solaris	S. Mayama, M. Matsumoto, K. Nemoto and T. Tanaka in Matsumoto et al. 2014	1
Fragilaria_bidens	Heiberg 1863	1
Fragilaria_capucina	Desmazières 1925	3
Fragilaria_crotonensis	Kitton 1869	2

Taxon	Authority	Number of barcodes [1]
Fragilaria_famelica	(Kützing) Lange-Bertalot 1980	1
Fragilaria_gracilis	Øestrup 1910	67
Fragilaria_mesolepta	Rabenhorst 1861	1
Fragilaria_pararumpens	Lange-Bertalot, G. Hofmann and Werum 2011	19
Fragilaria_perminuta	(Grunow) Lange-Bertalot 2000	3
Fragilaria_radians	(Kützing) Lange-Bertalot in Hofmann et al. 2011	1
Fragilaria_rumpens	(Kützing) Carlson 1913	3
Fragilaria_sp	Lyngbye 1819	7
Fragilaria_tenera	(W. Smith) Lange-Bertalot 1980	1
Fragilaria_vaucheriae	(Kützing) Petersen 1938	5
Fragilariforma_virescens	(Ralfs) D.M. Williams and Round 1988	2
Frustulia_crassinervia	(Brébisson) Lange-Bertalot and Krammer in Lange-Bertalot and Metzeltin 1996	2
Geissleria_decussis	(Hustedt) Lange-Bertalot and Metzeltin 1996	1
Gomphoneis_minuta	(Stone) Kociolek and Stoermer 1988	1
Gomphonema_acuminatum	Ehrenberg 1836	5
Gomphonema_affine	Kützing 1844	4
Gomphonema_angustatum	(Kützing) Rabenhorst 1864	2
Gomphonema_bourbonense	E. Reichardt 1997	2
Gomphonema_brebissonii	Kützing 1849	1
Gomphonema_capitatum	Ehrenberg 1838	1
Gomphonema_carolinense	Hagelstein 1939	1
Gomphonema_clavatum	Ehrenberg 1832	3
Gomphonema_clevei	Fricke in A. Schmidt 1902	4
Gomphonema_cymbelliclinum	E. Reichardt and Lange-Bertalot 1999	1

Taxon	Authority	Number of barcodes [1]
Gomphonema_dichotomum	Kützing 1834	1
Gomphonema_exilissimum	(Grunow) Lange-Bertalot and E. Reichardt 1996	5
Gomphonema_hebridense	Gregory 1854	8
Gomphonema_lagenula	Kützing 1844	5
Gomphonema_micropus	Kützing 1844	2
Gomphonema_minutum	(C. Agardh) C. Agardh 1831	2
Gomphonema_narodoense	R. Jahn, Abarca, J. Zimmermann and Enke	2
Gomphonema_parvulum	(Kützing) Kützing 1849	22
Gomphonema_pseudobohemicum	Lange-Bertalot and E. Reichardt 1993	1
Gomphonema_pumilum	(Grunow) E. Reichardt and Lange-Bertalot 1991	2
Gomphonema_rosenstockianum	Lange-Bertalot and Reichardt in Lange-Bertalot 1993	1
Gomphonema_saprophilum	(Lange-Bertalot and E. Reichardt) Abraca, R. Jahn, J. Zimmermann and Enke 2014	6
Gomphonema_sp	Ehrenberg 1832	2
Gomphonema_subclavatum	(Grunow) Grunow 1884	2
Gomphonema_truncatum	Ehrenberg 1832	2
Gomphonema_vibrio	Ehrenberg 1843	1
Gyrosigma_acuminatum	Ehrenberg 1836	1
Halamphora_coffeiformis	(C. Agardh) Levkov 2009	1
Halamphora_montana	(Krasske) Levkov 2009	1
Hannaea_arcus	R.M. Patrick in R.M. Patrick et Reimer 1966	2
*Hantzschia_amphioxys*_var._*major*	Grunow in Van Heurck 1881	1
Hippodonta_capitata	Ehrenberg 1838	1
Karayevia_oblongella	(Østrup) Aboal in Aboal et al. 2003	1

Taxon	Authority	Number of barcodes [1]
Kolbesia_ploenensis	(Hustedt) Round and L. Bukhtiyarova ex Round 1998	2
Lemnicola_hungarica	(Grunow) Round and Basson 1997	3
*Luticola*_sp.	D.G. Mann in Round, Crawford and Mann 1990	1
Luticola_sparsipunctata	Levkov, Metzeltin and A. Pavlov 2013	2
Luticola_ventricosa	(Kutz.) D.G. Mann in Round, Crawford and Mann 1990	1
*Mastogloia*_sp.	Thwaites in W. Smith 1856	1
Mayamaea_atomus	(Kützing) Lange-Bertalot 1997	2
Mayamaea_atomus_var_permitis	(Hustedt) Lange-Bertalot 1997	1
Mayamaea_terrestris	N. Abarca and R. Jahn in Zimmermann et al. 2014	1
Melosira_moniliformis	(O.F. Müller) Agardh 1824	1
Melosira_varians	C. Agardh 1827	9
Meridion_circulare	(Greville) C. Agardh 1831	1
Navicula_angusta	Grunow 1860	1
Navicula_cf._duerrenbergiana	Hustedt in Schmidt et al. 1934	1
Navicula_cryptocephala	Kützing 1844	8
Navicula_cryptotenella	Lange-Bertalot 1985	3
Navicula_gregaria	Donkin 1861	13
Navicula_lanceolata	(Agardh) Ehrenberg 1838	45
Navicula_radiosa	Kützing 1844	8
Navicula_rhynchotella	Lange-Bertalot 1993	4
Navicula_slesvicensis	Grunow in Van Heurck 1880	1
*Navicula*_sp.	Bory 1822	2
*Navicula*_sp.	Bory 1822	1
Navicula_tripunctata	(O.F. Müller) Bory 1822	4
Navicula_trivialis	Lange-Bertalot 1980	1
Navicula_upsaliensis	(Grunow) Peragallo 1903	1

Taxon	Authority	Number of barcodes [1]
Navicula_veneta	Kützing 1844	1
Neidium_affine	(Ehrenberg) Pfitzer 1871	2
Neidium_bisulcatum	(Lagerstedt) Cleve 1894	1
Neidium_dubium	(Ehrenberg) Cleve 1894	1
Neidium_productum	(W.Smith) 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	5
Nitzschia_capitellata	Hustedt in A. Schmidt et al. 1922	4
Nitzschia_cf._aequorea	Hustedt 1939	1
Nitzschia_cf._ardua	Cholnoky 1961	1
Nitzschia_cf._bulnheimiana	(Rabenhorst) H.L. Smith 1888	1
Nitzschia_cf._fonticola	Grunow in Cleve and Moeller 1879	2
Nitzschia_cf._microcephala	Grunow in Cleve and Grunow 1881	1
Nitzschia_cf._pusilla	Grunow 1862	2
Nitzschia_dissipata	(Kützing) Grunow 1862	4
Nitzschia_dubiiformis	Hustedt 1939	1
Nitzschia_filiformis	(W. Smith) Van Heurck 1896	2
Nitzschia_fonticola	Grunow in Van Heurck 1881	5
Nitzschia_frustulum	(Kützing) Grunow in Cleve et Grunow 1880	3
Nitzschia_hantzschiana	Rabenhorst 1860	2
Nitzschia_heufleuriana	Grunow 1862	1
Nitzschia_inconspicua	Grunow 1862	66
Nitzschia_linearis	(Agardh) W. Smith 1853	7
Nitzschia_palea	(Kützing) W. Smith 1856	47
Nitzschia_paleacea	Grunow in Van Heurck 1881	2
Nitzschia_perminuta	(Grunow) M. Peragallo 1903	1

Taxon	Authority	Number of barcodes[1]
Nitzschia_pusilla	(Kützing) Grunow em. Lange-Bertalot 1976	4
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	2
Nitzschia_sociabilis	Hustedt 1957	2
Nitzschia_soratensis	E. Morales and Vis 2007	10
*Nitzschia*_sp.	Hassall 1845	3
*Nitzschia*_sp._*s0819*	Hassall 1845	1
Nitzschia_sublinearis	Hustedt 1930	1
Nitzschia_vermicularoides	Lange-Bertalot	
Odontella_sinensis	(Greville) Grunow 1884	2
Paralia_sulcata	(Ehrenberg) Cleve 1873	1
Parlibellus_hamulifer	(Grunow) E.J. Cox 1988	1
Parlibellus_protracta	(Grunow) Witkowski, Lange-Bertalot and Metzeltin 2000	1
Pauliella_taeniata	(Grunow) F.E. Round and Basson 1997	1
Peronia_fibula	(Brébisson ex.Kützing) R. Ross 1956	1
Pinnularia_acrosphaeria	W. Smith 1853	3
Pinnularia_brebissonii	(Kützing) Rabenhorst 1864	1
Pinnularia_cf._gibba	(Ehrenberg) Ehrenberg 1843	3
*Pinnularia_cf._subgibba_* var._*sublinearis*	Krammer 2000	1
Pinnularia_divergens	W. Smith 1853	1
Pinnularia_grunowii	Krammer 2000	2
Pinnularia_isselana	Krammer 2000	1
Pinnularia_karelica	Cleve 1891	2
Pinnularia_microstauron	(Ehrenberg) Cleve 1891	3

Taxon	Authority	Number of barcodes 1
Pinnularia_neomajor	Krammer 1992	8
Pinnularia_parvulissima	Krammer 2000	4
Pinnularia_septentrionalis	Krammer 2000	2
Pinnularia_shivae	Cejudo-Figueiras, S. Blanco and Álvarez-Blanco 2012	1
*Pinnularia*_sp.	Ehrenberg 1843	3
Pinnularia_stomatophora	(Grunow) Cleve 1891	2
Pinnularia_subcapitata	Gregory 1856	4
*Pinnularia_subcommutata_* var._*nonfasciata*	Krammer 2000	4
Pinnularia_subgibba	Krammer 2000	2
Pinnularia_termitina	(Ehrenberg) R.M. Patrick 1966	1
Pinnularia_valida	Hustedt	2
Pinnularia_viridiformis	Krammer 1992	4
Pinnularia_viridis	(Nitzsch) Ehrenberg 1843	2
Placoneis_abiskoensis	(Hustedt) Lange-Bertalot and Metzeltin in Metzeltin and Witkowski 1996	1
Placoneis_clementis	(Grunow) E.J. Cox 1987	1
Placoneis_elginensis	(Gregory) E.J. Cox 1987	2
Planothidium_caputium	J. Zimmermann and R. Jahn in Zimmermann et al. 2014	1
Planothidium_frequentissimum	(Lange-Bertalot) Round and L. Bukhtiyarova 1996	4
Planothidium_lanceolatum	(Brébisson) Lange-Bertalot 1999	7
Platessa_conspicua	(A. Meyer) Lange-Bertalot 2004	1
Pleurosira_laevis	(Ehrenberg) Compère 1982	1
Psammothidium_bioretii	(Germain) L. Bukhtiyarova and Round 1996	1
Psammothidium_chlidanos	(Hohn and Hellerman) Lange-Bertalot 1999	1
Psammothidium_daonense	(Lange-Bertalot) Lange-Bertalot 1999	1

Taxon	**Authority**	**Number of barcodes** [1]
Psammothidium_pseudoswazi	(J.R. Carter) L. Bukhtiyarova and Round 1986	1
Pseudostaurosira_brevistriata	(Grunow in Van Heurck) D.M. Williams et Round 1987	3
Reimeria_sinuata	(Gregory) Kociolek et Stoermer 1987	2
Rhoicosphenia_abbreviata	(C.Agardh) Lange-Bertalot 1980	3
Rhopalodia_gibba	(Ehrenberg) O. Müller 1895	1
*Rossia*_sp.	M. Voigt 1960	1
Rossithidium_anastasiae	(Kaczmarska) Potapova 2012	1
Sellaphora_auldreekie	D.G. Mann and S.M. McDonald 2004	12
Sellaphora_bacillum	(Ehrenberg) D.G. Mann 1989	7
Sellaphora_blackfordensis	D.G. Mann and S. Droop 2004	12
Sellaphora_capitata	D.G. Mann and S.M. McDonald in D.G. Mann et al. 2004	9
Sellaphora_caput	K.M. Evans and D.G. Mann 2009	2
Sellaphora_cf._seminulum	(Grunow) D.G. Mann 1989	2
Sellaphora_cf_atomoides	(Grunow) Wetzel and Van de Vijver 2015	1
Sellaphora_cf_nigri	(De Not.) C.E. Wetzel and Ector in Wetzel et al. 2015	5
Sellaphora_joubaudii	(H.Germain) Aboal in Aboal et al. 2003	1
Sellaphora_laevissima	(Kützing) D.G. Mann 1989	9
Sellaphora_lanceolata	D.G. Mann and S. Droop 2004	3
Sellaphora_minima		4
Sellaphora_obesa	D.G. Mann and M.M. Bayer 2004	1
Sellaphora_pupula	(Kützing) Mereschkowsky 1902	93
Sellaphora_seminulum	(Grunow) D.G. Mann 1989	2
*Sellaphora*_sp.	Mereschkowsky 1902	2
Seminavis_cf._robusta	D.B. Danielidis and D.G. Mann 2002	1

Taxon	**Authority**	**Number of barcodes** [1]
Stauroneis_acuta	W. Smith 1853	1
Stauroneis_phoenicenteron	(Nitzsch) Ehrenberg 1843	2
Staurosira_construens	Ehrenberg 1843	4
Staurosira_elliptica	(Schumann) D.M. Williams et Round 1987	4
Staurosira_cf_subsalina	(Hustedt) Lange-Bertalot 2000	1
Staurosira_venter	(Ehrenberg) Grunow in Pantocsek 1889	4
Staurosirella_martyi	(Heribaud) E.A. Morales and K.M. Manoylov 2006	1
Staurosirella_pinnata	(Ehrenberg) Williams and Round 1987	1
Stenopterobia_curvula	(W.Smith) Krammer in Lange-Bertalot and Krammer 1987	1
Stenopterobia_pumila	Lange-Bertalot and U. Rumrich in U. Rumrich, Lange-Bertalot and M. Rumrich 2000	2
Stephanodiscus_agassizensis	Håkansson and Kling 1989	1
Stephanodiscus_binderanus	(Kützing) Krieger 1927	1
Stephanodiscus_hantzschii	Grunow in Cleve et Grunow 1880	2
Stephanodiscus_minutulus	(Kützing) Cleve and Moeller 1878	4
Stephanodiscus_neoastraea	Håkansson and Hickel 1986	1
Stephanodiscus_niagarae	Ehrenberg 1846	2
*Stephanodiscus_*sp.	Ehrenberg 1846	1
Stephanodiscus_yellowstonensis	E.C. Theriot and Stoermer 1984	1
Surirella_angusta	Kützing 1844	4
Surirella_biseriata	Brébisson in Brébisson and Godey 1836	1
Surirella_brebissonii	Krammer et Lange-Bertalot 1987	7
Surirella_capronii	Brébisson ex Kitton 1869	2
Surirella_cf._bifrons	Ehrenberg 1843	11

Taxon	Authority	Number of barcodes [1]
Surirella_cf._biseriata	Brébisson in Brébisson and Godey 1836	1
Surirella_cf._tenuissima	Hustedt 1942	2
Surirella_costata	Jurilj 1949	1
Surirella_helissella	Jurilj 1954	1
Surirella_iconella	A. Jurilj	2
Surirella_imbuta	Jurilj 1949	2
*Surirella_linearis*_var._*helvetica*	(Brun) Meister 1912	4
Surirella_lineopunctata	Jurilj 1949	2
Surirella_minuta	Brébisson ex Kützing 1849	1
Surirella_ovalis	Brébisson 1838	2
Surirella_spiralis	Kützing 1844	1
Surirella_splendida	(Ehrenberg) Kützing 1844	12
Surirella_tenera	Gregory 1856	3
*Synedra_fulgens*_var._*gigantea*	Lobarzewsky 1840	1
Synedra_hyperborea	Grunow 1884	1
Synedropsis_cf._recta	G.R. Hasle, Medlin and E.E. Syvertsen 1994	1
Tabellaria_flocculosa	(Roth) Kützing 1844	9
Tabularia_cf._tabulata	(C. Agardh) Snoeijs 1992	1
Tabularia_fasciculata	(C. Agardh) D.M. Williams and Round 1986	1
Tabularia_laevis	Kützing	1
*Tabularia*_sp._*Naples*	(Kützing) D.M. Williams and Round 1986	1
Thalassiosira_pseudonana	Hasle and Heimdal 1970	1
Thalassiosira_punctigera	(Castracane) Hasle 1983	2
Tryblionella_apiculata	Gregory 1857	1
Tryblionella_constricta	Gregory 1855	1
Tryblionella_debilis	Arnott in O'Meara 1873	1
*Tryblionella*_sp.	W. Smith 1857	1

Taxon	**Authority**	**Number of barcodes** [1]
Ulnaria_acus	(Kützing) Aboal in Aboal, Alvarez Cobelas, Cambra and Ector 2003	6
Ulnaria_ulna	(Nitzsch) P. Compère in Jahn et al. 2001	13

Notes: [1] Number of barcodes assigned to each taxon in the database. Some taxa have multiple representatives of the same barcode (100% sequence identity) and so this table does not represent the genetic variability of barcodes within each taxon.

Appendix 2: List of taxa with indicator values for calculating TDI4, TDI5LM and TDI5NGS

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100041538	AC008A	*Achnanthes exigua*	3.0	3.0	
NBNSYS0100041567	AC095A	*Achnanthes minuscula*	2.0	3.0	
NHMSYS0000523413	AC175A	*Achnanthes rupestoides*	5.0	5.0	
NBNSYS0100041514	AC9999	*Achnanthes* sp.	3.0	3.0	
NHMSYS0020749119	AC161A	*Achnanthes ventralis*	1.0	1.0	
NHMSYS0021166192	ACHD-04	*Achnanthidium caledonicum*	1.0	1.0	
NHMSYS0021166193	ACHD-09	*Achnanthidium coarctatum*	2.0	2.0	2.9
NHMSYS0020475041	ZZZ835	*Achnanthidium minutissimum type*	2.0	2.0	1.4
NHMSYS0020953761	ACHD-02	*Achnanthidium pyrenaicum*	2.0	2.0	1.7
NHMSYS0021166275	ACHD-11	*Achnanthidium reimeri*	2.0	2.0	1.6
NBNSYS0100041600	AD9999	*Achnanthidium* sp.	2.0	1.0	
NHMSYS0020063117	ZZZ911	*Achnanthidium subatomus*	2.0	2.0	
NHMSYS0021166276	ACHD-12	*Achnnathidium rivulare*	2.0	2.0	1.8
NHMSYS0021166247	ADLA-05	*Adlafia brockmannii*			1.9
NHMSYS0020970838	ADLA-01	*Adlafia bryophila*	3.0	2.0	2.2
NHMSYS0020970839	ADLA-03	*Adlafia minuscula*	3.0	3.0	2.3
NHMSYS0020953766	ADLA-02	*Adlafia minuscula* var. *muralis*	5.0	3.0	
NHMSYS0020953767	ADLA-04	*Adlafia suchlandtii*	5.0	4.0	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100041893	AP9970	*Amphipleura lindheimeri*	2.0	2.0	
NBNSYS0100041894	AP001A	*Amphipleura pellucida*	3.0	3.0	2.4
NBNSYS0100041891	AP9999	*Amphipleura* sp.	2.0	2.0	
NHMSYS0021166252	AMPH-15	*Amphora affinis*			4.6
NHMSYS0021166277	AMPH-10	*Amphora atomoides*	4.0	4.0	2.6
NHMSYS0021166422	AMPH-09	*Amphora copulata*	4.0	4.0	4.4
NBNSYS0100041921	AM013A	*Amphora inariensis*	5.0	5.0	
NBNSYS0100041934	AM001A	*Amphora ovalis*	5.0	5.0	4.9
NHMSYS0020953787	AMPH-05	*Amphora pediculus type*	5.0	5.0	4.9
NBNSYS0100041902	AM9999	*Amphora* sp.	4.0	4.0	4.9
NBNSYS0100042051	AN009A	*Anomoeoneis sphaerophora*			3.4
NBNSYS0100042481	BA005A	*Bacillaria paxillifer*	5.0	5.0	5.8
NBNSYS0100042681	BR006A	*Brachysira brebissonii*	1.0	1.0	
NHMSYS0020475076	BR010A	*Brachysira neoexilis*	1.0	1.0	-0.7
NBNSYS0100042678	BR9999	*Brachysira* sp.	1.0	1.0	
NHMSYS0020953822	BRAC-02	*Brachysira vitrea type*	1.0	1.0	-0.2
NHMSYS0021166279	BREB-01	*Brebissonia lanceolata*			3.2
NBNSYS0100042873	CA006A	*Caloneis amphisbaena*	5.0	5.0	4.7
NBNSYS0100042879	CA002A	*Caloneis bacillum*	4.0	4.0	
NBNSYS0100042884	CALO-05	*Caloneis budensis*	3.0	3.0	3.2
NBNSYS0100042892	ZZZ993	*Caloneis hyalina*	5.0	5.0	
NBNSYS0100042894	CA035A	*Caloneis lauta*	3.0	3.0	2.6
NBNSYS0100042900	CA048A	*Caloneis molaris*	4.0	4.0	
NBNSYS0100042906	CA003A	*Caloneis silicula*	4.0	4.0	3.5
NBNSYS0100042870	CA9999	*Caloneis* sp.	3.0	3.0	4.7

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100042967	CP017A	*Campylodiscus hibernicus*			4.9
NBNSYS0100043120	CV001A	*Cavinula cocconeiformis*	2.0	2.0	
NHMSYS0021168082	CLIM-01	*Climaconeis riddleae*			5.6
NHMSYS0021168076	CLIP-01	*Climacosphenia* sp.			-0.6
NHMSYS0020475100	CO006A	*Cocconeis diminuta*	5.0	5.0	
NBNSYS0100043864	CO010A	*Cocconeis disculus*	5.0	4.0	
NBNSYS0100043872	CO005A	*Cocconeis pediculus*	4.0	4.0	4.0
NHMSYS0020953847	COCO-01	*Cocconeis placentula* agg.	3.1	3.1	3.2
NBNSYS0100043882	CO068A	*Cocconeis pseudothumensis*	4.0	4.0	
NBNSYS0100043862	CO9999	*Cocconeis* sp.	3.0	3.0	
NBNSYS0100044576	CI002A	*Craticula accomoda*	4.0	4.0	4.4
NBNSYS0100044578	CI003A	*Craticula ambigua*	4.0	4.0	
NHMSYS0021166284	CRAT-06	*Craticula buderi*			4.4
NBNSYS0100044579	CI004A	*Craticula cuspidata*			4.6
NBNSYS0100044581	CI005A	*Craticula halophila*	4.0	4.0	
NHMSYS0021167631	CRAT-08	*Craticula importuna*			3.7
NHMSYS0020953861	CRAT-02	*Craticula minusculoides*	5.0	5.0	
NHMSYS0020953862	CRAT-01	*Craticula molestiformis*	5.0	4.0	4.5
NHMSYS0021166201	CRAT-07	*Craticula subminuscula*	4.0	4.0	4.4
NBNSYS0100044688	YH001A	*Ctenophora pulchella*	2.0	2.0	1.1
NBNSYS0100044911	CL002A	*Cymatopleura elliptica*	5.0	5.0	4.7
NHMSYS0020063119	CL001A	*Cymatopleura solea*	5.0	5.0	4.4
NBNSYS0100044909	CL9999	*Cymatopleura* sp.	5.0	2.0	
NBNSYS0100044924	CM022A	*Cymbella affinis*	2.0	2.0	
NBNSYS0100044930	CM005A	*Cymbella aspera*	3.0	3.0	1.1
NHMSYS0021166301	CYMB-09	*Cymbella baicalensis*	3.0	3.0	2.2

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100044939	CM006A	*Cymbella cistula*	2.0	2.0	2.6
NBNSYS0100044945	CM007A	*Cymbella cymbiformis*	3.0	3.0	
NHMSYS0021166300	CYMB-08	*Cymbella excisa*	2.0	2.0	2.8
NBNSYS0100044955	CM013A	*Cymbella helvetica*	2.0	2.0	1.6
NHMSYS0021166303	CYMB-11	*Cymbella janischii*	3.0	3.0	4.4
NBNSYS0100044965	CM041A	*Cymbella lanceolata*	4.0	4.0	3.7
NHMSYS0021166304	CYMB-12	*Cymbella mexicana*	3.0	3.0	2.1
NBNSYS0100044980	CM030A	*Cymbella proxima*			2.3
NBNSYS0100044918	CM9999	*Cymbella* sp.	3.0	3.0	1.5
NHMSYS0021166305	CYMB-13	*Cymbella stuxbergia*	3.0	3.0	2.7
NBNSYS0100044996	CM042A	*Cymbella tumida*	3.0	3.0	2.7
NBNSYS0100045002	CN001A	*Cymbellonitzschia diluviana*	4.0	4.0	
NHMSYS0020953871	CYMB-02	*Cymbopleura amphicephala*	3.0	3.0	
NHMSYS0021166259	CYMB-15	*Cymbopleura inaequalis*			4.7
NHMSYS0020953875	CYMB-01	*Cymbopleura naviculiformis*	3.0	3.0	1.6
NHMSYS0020953871	CYMB-16	*Cymbopleura* sp.			2.1
NHMSYS0020953877	CYMB-05	*Cymbopleura subaequalis*	5.0	1.0	
NHMSYS0021166258	CYMB-14	*Cymbopleura subcuspidata*			2.8
NHMSYS0020953879	DELI-01	*Delicata delicatula*	1.0	1.0	
NBNSYS0100045140	DE001A	*Denticula tenuis*	2.0	2.0	
NBNSYS0100045199	DA005A	*Diadesmis contenta*	5.0	4.0	
NHMSYS0020953881	ZZZ844	*Diadesmis contenta fo. biceps*	3.0	3.0	
NBNSYS0100045202	DA002A	*Diadesmis gallica*	3.0	3.0	
NBNSYS0100045195	DA007A	*Diadesmis perpusilla*	2.0	3.0	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NHMSYS0020063121	DT010A	*Diatoma ehrenbergii*	1.0	1.0	
NHMSYS0000719189	DT021A	*Diatoma mesodon*	1.0	1.0	
NHMSYS0020953884	DT022A	*Diatoma moniliformis*	1.0	1.0	1.4
NHMSYS0020749132	ZZZ941	*Diatoma problematica*	2.0	2.0	
NBNSYS0100045220	DT9999	*Diatoma* sp.	2.0	2.0	
NBNSYS0100045227	DT004A	*Diatoma tenue*	1.0	1.0	2.6
NHMSYS0020953886	DIAT-01	*Diatoma vulgare* agg.	5.0	4.0	3.7
NHMSYS0021167649	DIDY-01	*Didymosphenia dentata*	1.0	1.0	0.9
NBNSYS0100045326	DD001A	*Didymosphenia geminata*	1.0	1.0	0.4
NBNSYS0100045378	DP009A	*Diploneis elliptica*	5.0	4.0	
NBNSYS0100045382	DP012A	*Diploneis marginestriata*	5.0	5.0	
NBNSYS0100045386	DP007A	*Diploneis oblongella*	5.0	3.0	
NBNSYS0100045389	DP001A	*Diploneis ovalis*	3.0	3.0	
NBNSYS0100045367	DP9999	*Diploneis* sp.	5.0	4.0	
NHMSYS0021166200	DP061A	*Diploneis subovalis*	5.0	5.0	5.6
NBNSYS0100045630	EL001A	*Ellerbeckia arenaria*	4.0	4.0	
NHMSYS0020953907	ENCY-06	*Encyonema 'ventricosum'* agg.	2.6	2.6	
NBNSYS0100045681	EY002A	*Encyonema caespitosum*	4.0	3.0	1.2
NBNSYS0100045684	EY017A	*Encyonema gracile*	1.0	1.0	
NBNSYS0100045694	EY011A	*Encyonema minutum*	2.6	2.6	2.0
NHMSYS0020749137	EY004A	*Encyonema prostratum*	5.0	3.0	4.5
NBNSYS0100045702	EY015A	*Encyonema reichardtii*	2.6	2.6	
NBNSYS0100045704	EY016A	*Encyonema silesiacum*	2.6	2.6	1.9
NBNSYS0100045676	EY9999	*Encyonema* sp.	4.0	2.0	2.1
NBNSYS0100045709	ENCY-10	*Encyonema triangulum*	4.0	4.0	1.3
NHMSYS0020953911	ENCS-03	*Encyonopsis cesatii*	2.0	1.0	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NHMSYS0020953912	ENCS-07	*Encyonopsis falaisensis*	1.0	1.0	0.3
NHMSYS0020953913	ENCS-01	*Encyonopsis microcephala*	2.0	1.0	0.4
NBNSYS0100045744	EN001A	*Entomomeis ornata*			2.3
NBNSYS0100045742	EN9999	*Entomoneis* sp.			5.4
NHMSYS0020970860	EOLI-01	*Eolimna minima*	3.0	3.0	3.0
NBNSYS0100045821	EP007A	*Epithemia adnata*	5.0	3.0	
NBNSYS0100045836	EP001A	*Epithemia sorex*	3.0	3.0	
NBNSYS0100045820	EP9999	*Epithemia* sp.	2.0	1.0	
NBNSYS0100045838	EP004A	*Epithemia turgida*	1.0	2.0	
NBNSYS0100046052	EC001A	*Eucocconeis flexella*	1.0	1.0	
NHMSYS0020970862	EUCO-01	*Eucocconeis laevis*	1.0	1.0	1.0
NBNSYS0100046149	EU013A	*Eunotia arcus*	1.4	1.4	1.0
NHMSYS0020749138	EU070A	*Eunotia bilunaris*	1.4	1.4	0.2
NBNSYS0100046170	EU016A	*Eunotia diodon*	1.4	1.4	
NBNSYS0100046174	EU009A	*Eunotia exigua*	1.4	1.4	0.0
NHMSYS0020749140	EU009C	*Eunotia exigua* var. *tridentula*	1.4	1.4	
NBNSYS0100046177	EU025A	*Eunotia fallax*	1.4	2.4	
NBNSYS0100046182	EU018A	*Eunotia formica*	1.4	1.4	2.2
NBNSYS0100046185	EU024A	*Eunotia glacialis*	1.4	1.4	1.0
NBNSYS0100046189	EU107A	*Eunotia implicata*	1.4	1.4	0.5
NBNSYS0100046190	EU047A	*Eunotia incisa*	1.4	1.4	
NBNSYS0100046191	EU108A	*Eunotia intermedia*	1.4	1.4	
NHMSYS0021166541	EUNO-05	*Eunotia latitaenia*	1.4	1.4	2.0
NBNSYS0100046199	EU020A	*Eunotia meisteri*	1.4	1.4	
NHMSYS0020063133	EU110A	*Eunotia minor*	1.4	1.4	1.0
NHMSYS0020953923	EU008A	*Eunotia monodon fo. monodon*	1.4	2.4	
NBNSYS0100046207	EU114A	*Eunotia muscicola*	1.4	1.4	
NBNSYS0100046209	EU114B	*Eunotia muscicola* var. *tridentula*	1.4	1.4	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NHMSYS0020970866	EU040B	*Eunotia paludosa* var. *trinacria*	1.4	1.4	
NBNSYS0100046217	EU002A	*Eunotia pectinalis*	1.4	1.0	0.4
NBNSYS0100046234	EU011A	*Eunotia rhomboidea*	1.4	1.4	
NBNSYS0100046244	EUNO-01	*Eunotia silvahercynia*	1.4	1.4	
NBNSYS0100046142	EU9999	*Eunotia* sp.	1.4	1.4	-0.4
NBNSYS0100046247	EU105A	*Eunotia subarcuatoides*	1.4	1.4	
NBNSYS0100046251	EU004A	*Eunotia tenella*	1.4	1.4	
NBNSYS0100046254	EU053A	*Eunotia tridentula*	1.4	1.4	
NBNSYS0100046312	FA009A	*Fallacia helensis*	5.0	5.0	
NBNSYS0100046314	ZZZ866	*Fallacia indifferens*	2.0	2.0	
NBNSYS0100046315	FA013A	*Fallacia insociabilis*	4.0	4.0	
NHMSYS0020953936	FALL-01	*Fallacia lenzii*	5.0	5.0	
NBNSYS0100046318	FA016A	*Fallacia monoculata*	4.0	4.0	5.0
NBNSYS0100046323	FA001A	*Fallacia pygmaea*	5.0	5.0	4.3
NBNSYS0100046308	FA9999	*Fallacia* sp.	5.0	3.0	
NBNSYS0100046325	FA021A	*Fallacia subhamulata*	5.0	5.0	
NHMSYS0020953937	MAYA-04	*Fistulifera / Mayamaea*	3.9	3.9	
NHMSYS0020970868	FIST-02	*Fistulifera pelliculosa*	3.9	3.9	
NHMSYS0020970869	FIST-01	*Fistulifera saprophila*	3.9	3.9	3.7
NHMSYS0021166203	FIST-04	*Fistulifera solaris*	3.9	3.9	3.5
NBNSYS0100046434	FR026A	*Fragilaria bidens*	3.0	3.0	1.8
NBNSYS0100046435	FR009A	*Fragilaria capucina*	1.2	1.2	1.0
NHMSYS0020953939	FRAG-02	*Fragilaria capucina* agg.	1.2	1.2	
NHMSYS0000523730	FR009L	*Fragilaria capucina* var. *amphicephala*	1.2	1.2	
NHMSYS0020749145	FR009I	*Fragilaria capucina* var. *austriaca*	1.2	1.2	
NHMSYS0021166220	FRAG-06	*Fragilaria delicatissima*	3.0	3.0	
NHMSYS0021168236	FRAG-10	*Fragilaria famelica*	2.0	5.0	2.9

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NHMSYS0021166310	FRAG-11	*Fragilaria gracilis*	1.2	1.2	1.0
NHMSYS0020970881	FR040A	*Fragilaria mesolepta*	1.2	1.2	
NHMSYS0020970882	FRAG-07	*Fragilaria nanana*	1.0	1.0	
NHMSYS0020953951	FRAG-01	*Fragilaria nanoides*			1.7
NHMSYS0021166204	FRAG-03	*Fragilaria pararumpens*	1.2	1.2	1.0
NHMSYS0020063125	ZZZ842	*Fragilaria perminuta*	1.2	1.2	1.5
NBNSYS0100046450	FR059A	*Fragilaria radians*	1.2	1.2	2.2
NHMSYS0021166206	FRAG-08	*Fragilaria recapitellata*	2.0	3.0	
NHMSYS0021166311	FRAG-12	*Fragilaria rumpens*	1.2	1.2	1.8
NBNSYS0100046431	FR9999	*Fragilaria* sp.	2.0	2.0	0.9
NHMSYS0020063126	FR060A	*Fragilaria tenera*	1.0	1.0	
NBNSYS0100046453	FR007A	*Fragilaria vaucheriae*	2.0	2.0	1.6
NBNSYS0100046456	FF002A	*Fragilariforma bicapitata*	2.5	2.5	
NBNSYS0100046457	FF003A	*Fragilariforma constricta*	2.5	1.5	
NBNSYS0100046468	ZZZ841	*Fragilariforma exigua*	2.5	2.5	
NBNSYS0100046454	FRFO-01	*Fragilariforma* sp.	2.5	2.5	
NBNSYS0100046465	FF001A	*Fragilariforma virescens*	2.5	1.5	4.7
NHMSYS0020953956	FRUS-03	*Frustulia crassinveria*	1.0	1.0	-0.3
NHMSYS0020953958	FRUS-04	*Frustulia erifuga*	1.0	1.0	0.0
NHMSYS0021167685	FRUS-07	*Frustulia gondwana*			0.2
NHMSYS0020953959	FRUS-01	*Frustulia krammeri*	1.0	1.0	
NBNSYS0100046485	FU002A	*Frustulia rhomboides*	1.0	1.0	
NBNSYS0100046482	FU9999	*Frustulia* sp.	5.0	2.0	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100046492	FU001A	*Frustulia vulgaris*	2.0	2.0	2.8
NHMSYS0020953964	FU037A	*Frustulia weinholdii*	1.0	1.0	
NHMSYS0020970886	GEIS-05	*Geissleria acceptata*	4.0	4.0	
NHMSYS0020970887	GEIS-02	*Geissleria decussis*	5.0	3.0	4.4
NHMSYS0020970888	GEIS-03	*Geissleria ignota*	4.0	4.0	
NHMSYS0020953968	GEIS-01	*Geissleria schoenfeldii*	3.0	3.0	
NHMSYS0021166383	GEIS-06	*Geissleria* sp.			2.0
NHMSYS0021166313	GOMP-10	*Gomphoneis minuta*			4.4
NHMSYS0021166261	GM9999	*Gomphoneis sp*			2.9
NHMSYS0020953974	ZZZ834	*Gomphonema ‘intricatum’ type*	3.6	2.0	2.5
NBNSYS0100046689	GO006A	*Gomphonema acuminatum*	2.0	2.0	1.8
NBNSYS0100046696	GO020A	*Gomphonema affine*	5.0	2.0	1.8
NBNSYS0100046698	GO003A	*Gomphonema angustatum*	3.0	3.0	0.0
NBNSYS0100046701	GO003E	*Gomphonema angustatum* var. *sarcophagus*	3.0	3.0	2.5
NBNSYS0100046704	GO019A	*Gomphonema augur*	5.0	5.0	
NHMSYS0021166314	GOMP-11	*Gomphonema bourbonense*	3.0	3.0	3.8
NBNSYS0100046711	GO029A	*Gomphonema clavatum*	2.0	2.0	0.7
NBNSYS0100046712	GO024C	*Gomphonema clevei*	1.0	1.0	4.6
NHMSYS0021166209	GOMP-09	*Gomphonema exilissimum*	3.0	2.0	
NBNSYS0100046717	GO004A	*Gomphonema gracile*	1.0	1.0	1.6
NBNSYS0100046722	GO074A	*Gomphonema hebridense*			0.5
NBNSYS0100046725	GO043A	*Gomphonema insigne*	5.0	3.0	
NHMSYS0021166316	GOMP-13	*Gomphonema lagenula*	3.0	3.0	1.1

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100046727	GO086A	*Gomphonema lateripunctatum*	3.6	5.0	
NBNSYS0100046730	GO050A	*Gomphonema minutum*	3.6	3.0	0.6
NHMSYS0021166318	GOMP-15	*Gomphonema narodoense*	3.0	3.0	1.5
NBNSYS0100046733	GO052A	*Gomphonema olivaceoides*	2.0	2.0	
NBNSYS0100046734	GO001A	*Gomphonema olivaceum*	3.0	3.0	
NHMSYS0020953985	GMPH-03	*Gomphonema olivaceum* agg.	3.0	3.0	
NBNSYS0100046740	GO013A	*Gomphonema parvulum*	3.0	3.0	1.4
NBNSYS0100046745	GO055A	*Gomphonema pseudoaugur*	5.0	5.0	
NHMSYS0021166319	GOMP-18	*Gomphonema rosenstockianum*	3.0	3.0	3.1
NHMSYS0021166320	GOMP-16	*Gomphonema saprophilum*	3.0	3.0	2.9
NBNSYS0100046687	GO9999	*Gomphonema* sp.	3.0	3.0	1.5
NHMSYS0021166321	GOMP-17	*Gomphonema subclavatum*	3.0	3.0	0.8
NBNSYS0100046751	GO066A	*Gomphonema tergestinum*	4.0	4.0	
NBNSYS0100046752	GO023A	*Gomphonema truncatum*	3.0	2.0	2.0
NBNSYS0100046753	GO023B	*Gomphonema truncatum* var. *capitatum*	2.0	2.0	1.8
NBNSYS0100046755	GO027A	*Gomphonema ventricosum*	1.0	1.0	
NBNSYS0100046756	GO025H	*Gomphonema vibrio*	3.6	3.6	4.8
NHMSYS0021166315	GOMP-12	*Gomphonema carolinense*	3.0	3.0	2.5
NHMSYS0020953991	GOMP-01	*Gomphosphenia grovei*	4.0	4.0	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NHMSYS0020953992	GOMP-02	*Gomphosphenia grovei* var. *lingulata*	4.0	4.0	
NBNSYS0100046917	GY005A	*Gyrosigma acuminatum*	5.0	5.0	4.8
NBNSYS0100046920	GY001A	*Gyrosigma attenuatum*	5.0	5.0	
NBNSYS0100046929	ZZZ970	*Gyrosigma nodiferum*	4.0	5.0	
NHMSYS0020749156	GY025A	*Gyrosigma scalproides*	5.0	5.0	
NHMSYS0021166196	HALA-02	*Halamphora montana*	5.0	5.0	3.6
NHMSYS0021166197	HALA-03	*Halamphora normanii*	5.0	5.0	4.4
NHMSYS0021166198	HALA-04	*Halamphora oligotraphenta*	1.0	1.0	1.4
NHMSYS0021166199	HALA-05	*Halamphora veneta*	5.0	5.0	3.7
NBNSYS0100047043	HN001A	*Hannaea arcus*	1.0	1.0	0.4
NHMSYS0020749158	ZZZ932	*Hantzschia abundans*	5.0	5.0	
NBNSYS0100047047	HA001A	*Hantzschia amphioxys*	4.0	4.0	
NBNSYS0100047051	HA001F	*Hantzschia amphioxys* var. *major*	4.0	4.0	0.8
NBNSYS0100047046	HA9999	*Hantzschia* sp.	2.0	3.0	
NHMSYS0021166392	HIPP-01	*Hippodonta capitata*	4.0	4.0	4.2
NHMSYS0021166395	HIPP-02	*Hippodonta costulata*	4.0	4.0	
NBNSYS0100047734	ZZZ908	*Karayevia clevei*	5.0	5.0	
NBNSYS0100047735	ZZZ953	*Karayevia laterostrata*	3.0	3.0	
NHMSYS0021166191	KARA-01	*Karayevia oblongella*	1.0	1.0	0.7
NBNSYS0100047733	KARA-04	*Karayevia* sp.			5.3
NBNSYS0100047784	ZZZ899	*Kolbesia kolbei*	5.0	5.0	
NBNSYS0100047785	ZZZ887	*Kolbesia ploenensis*	5.0	5.0	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NHMSYS0020063130	ZZZ900	*Lemnicola hungarica*	5.0	5.0	3.8
NBNSYS0100048361	LU003A	*Luticola goeppertiana*	5.0	5.0	4.7
NHMSYS0020749159	ZZZ840	*Luticola kotschyi*	5.0	2.0	
NBNSYS0100048363	LU001A	*Luticola mutica*	3.0	5.0	
NBNSYS0100048367	LU005A	*Luticola nivalis*	5.0	5.0	
NBNSYS0100048357	LU9999	*Luticola* sp.	4.0	4.0	
NHMSYS0021166323	LUTI-02	*Luticola sparsipunctata*	4.0	4.0	3.4
NBNSYS0100048373	LU009A	*Luticola ventricosa*	3.0	3.0	1.9
NHMSYS0020954056	MAYA-01	*Mayamaea atomus*	3.9	3.9	
NHMSYS0021167808	MAYA-08	*Mayamaea atomus* var. *alcimonica*	3.9	3.9	5.8
NHMSYS0020954057	MAYA-03	*Mayamaea atomus* var. *permitis*	3.9	3.9	3.7
NHMSYS0021166288	MAYA-07	*Mayamaea fossalis*	3.9	3.9	4.8
NHMSYS0020970892	MAYA-05	*Mayamaea lacunolaciniata*	3.9	3.9	
NHMSYS0021166324	MAYA-06	*Mayamaea terrestris*			4.9
NBNSYS0100048612	ME035A	*Melosira moniliformis*			4.9
NBNSYS0100048613	ME007A	*Melosira nummuloides*			3.2
NBNSYS0100048607	ME9999	*Melosira* sp.			3.6
NBNSYS0100048616	ME015A	*Melosira varians*	4.0	4.0	3.8
NBNSYS0100048637	MR001A	*Meridion circulare*	2.0	2.0	0.6
NBNSYS0100048638	MR001B	*Meridion circulare* var. *constrictum*	1.0	1.0	
NBNSYS0100049175	NA037A	*Navicula angusta*	1.0	1.0	0.1
NBNSYS0100049211	NA745A	*Navicula capitatoradiata*	4.0	4.0	4.1
NBNSYS0100049212	NA051A	*Navicula cari*	4.0	4.0	5.1
NBNSYS0100049217	NA021A	*Navicula cincta*	4.0	4.0	5.0
NBNSYS0100049220	ZZZ982	*Navicula claytonii*	3.0	3.0	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100049233	NA007A	*Navicula cryptocephala*	3.0	3.0	3.3
NBNSYS0100049239	NA751A	*Navicula cryptotenella*	4.0	4.0	4.1
NHMSYS0020954084	NAVI-04	*Navicula cryptotenella* agg.	4.0	4.0	
NBNSYS0100049250	NA060A	*Navicula digitoradiata*	4.0	4.0	
NBNSYS0100049293	NA023A	*Navicula gregaria*	4.0	4.0	3.9
NHMSYS0020749171	NA004A	*Navicula hungarica*	5.0	5.0	
NBNSYS0100049316	NA760A	*Navicula ingenua*	5.0	5.0	
NBNSYS0100049331	NA009A	*Navicula lanceolata*	4.0	4.0	3.8
NHMSYS0020749176	NA769A	*Navicula lundii*	3.0	3.0	4.8
NBNSYS0100049353	NA030A	*Navicula menisculus*	4.0	4.0	
NBNSYS0100049374	NA547A	*Navicula oppugnata*	4.0	4.0	
NHMSYS0021166410	NAVI-08	*Navicula perminuta*	4.0	4.0	3.2
NBNSYS0100049384	NA058A	*Navicula phyllepta*			4.2
NHMSYS0021167832	NAVI-09	*Navicula pseudacceptata*	4.0	4.0	4.1
NBNSYS0100049398	NA079A	*Navicula pseudolanceolata*	2.0	3.0	
NBNSYS0100049402	NA003A	*Navicula radiosa*	3.0	3.0	3.1
NHMSYS0020749178	NA773A	*Navicula radiosafallax*	5.0	5.0	
NHMSYS0021166386	NA059A	*Navicula ramosissima*	4.0	5.0	6.0
NBNSYS0100049404	NA762A	*Navicula recens*	5.0	5.0	
NBNSYS0100049406	NA768A	*Navicula reichardtiana*	4.0	4.0	
NBNSYS0100049408	NA026A	*Navicula reinhardtii*	4.0	4.0	4.2
NBNSYS0100049413	NA008A	*Navicula rhynchocephala*	2.0	2.0	
NHMSYS0020749179	ZZZ847	*Navicula rhynchotella*	4.0	4.0	4.5
NBNSYS0100049420	NA035A	*Navicula salinarum*	5.0	5.0	
NBNSYS0100049428	NA764A	*Navicula schroeteri*	5.0	5.0	
NBNSYS0100049439	NA080A	*Navicula slesvicensis*	3.0	3.0	2.7
NBNSYS0100049162	NA9999	*Navicula* sp.	4.0	4.0	3.2

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100049465	NA743A	*Navicula subrhynchocephala*	4.0	4.0	
NBNSYS0100049472	NA675A	*Navicula tenelloides*	4.0	4.0	
NBNSYS0100049481	NA095A	*Navicula tripunctata*	5.0	5.0	4.5
NBNSYS0100049483	NA063A	*Navicula trivialis*	4.0	4.0	4.3
NHMSYS0021166325	NAVI-10	*Navicula upsaliensis*	4.0	5.0	4.9
NBNSYS0100049494	NA054A	*Navicula veneta*	4.0	4.0	4.4
NBNSYS0100049497	NA027A	*Navicula viridula*	4.0	4.0	4.1
NHMSYS0020954117	NAVD-01	*Navicula(dicta) schmassmannii*	3.0	2.0	
NBNSYS0100049516	NE003A	*Neidium affine*	3.0	3.0	
NHMSYS0021166238	NE015A	*Neidium amphigomphus*	3.0	3.0	1.3
NBNSYS0100049523	NE036A	*Neidium ampliatum*	3.0	3.0	
NBNSYS0100049526	ZZZ974	*Neidium binodeforme*	4.0	4.0	
NBNSYS0100049528	NE004A	*Neidium bisulcatum*			2.6
NBNSYS0100049535	NEID-01	*Neidium dilatatum*			1.1
NHMSYS0021167838	NEID-03	*Neidium fossum*	3.0	3.0	3.0
NBNSYS0100049563	NE002A	*Neidium productum*	5.0	4.0	3.9
NBNSYS0100049515	NE9999	*Neidium* sp.	3.0	3.0	1.6
NHMSYS0021167844	NEID-02	*Neidium tumescens*	3.0	3.0	-0.1
NBNSYS0100049697	NI057A	*Nitzschia acicularioides*	5.0	3.0	
NBNSYS0100049698	NI042A	*Nitzschia acicularis*	3.0	3.0	2.8
NBNSYS0100049699	NI030A	*Nitzschia acidoclinata*			0.8
NHMSYS0020749182	NI021A	*Nitzschia acula*	5.0	5.0	
NHMSYS0020954124	NI060A	*Nitzschia aequorea*			4.2
NHMSYS0020749183	NI061A	*Nitzschia aerophila*	4.0	5.0	
NBNSYS0100049703	NI063A	*Nitzschia agnita*	5.0	5.0	
NBNSYS0100049705	NI014A	*Nitzschia amphibia*	5.0	5.0	4.9
NBNSYS0100049714	NI199A	*Nitzschia angustatula*	4.0	4.0	
NBNSYS0100049716	NI065A	*Nitzschia archibaldii*	2.0	2.0	
NBNSYS0100049725	NI072A	*Nitzschia bremensis*	2.0	3.0	
NBNSYS0100049727	NI073A	*Nitzschia brevissima*	2.0	5.0	
NBNSYS0100049729	NI028A	*Nitzschia capitellata*	4.0	4.0	4.3

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100049730	NI080A	*Nitzschia clausii*	4.0	5.0	
NHMSYS0000523885	NI010A	*Nitzschia communis*	5.0	5.0	3.6
NBNSYS0100049731	NI011A	*Nitzschia commutata*	5.0	3.0	
NBNSYS0100049738	ZZZ989	*Nitzschia denticula*			4.2
NHMSYS0020063151	NI091A	*Nitzschia disputata*	2.0	2.0	
NBNSYS0100049740	NI015A	*Nitzschia dissipata*	3.0	3.0	3.6
NBNSYS0100049741	ZZZ930	*Nitzschia dissipata* var. *media*	3.0	3.0	3.7
NBNSYS0100049743	NI093A	*Nitzschia draveillensis*	2.0	3.0	3.0
NBNSYS0100049744	NI018A	*Nitzschia dubia*	5.0	5.0	
NBNSYS0100049749	NI096A	*Nitzschia epithemoides*	3.0	2.0	
NBNSYS0100049753	NI098A	*Nitzschia filiformis*	5.0	5.0	3.2
NBNSYS0100049755	NI099A	*Nitzschia flexa*	3.0	3.0	
NBNSYS0100049757	NI002A	*Nitzschia fonticola*	4.0	4.0	3.5
NHMSYS0020749184	NI212A	*Nitzschia fossilis*	5.0	5.0	
NBNSYS0100049759	NI008A	*Nitzschia frustulum*	3.0	3.0	4.0
NBNSYS0100049767	NI017A	*Nitzschia gracilis*	3.0	3.0	
NBNSYS0100049769	NI034A	*Nitzschia hantzschiana*	2.0	2.0	
NBNSYS0100049770	NI052A	*Nitzschia heufleriana*	4.0	4.0	
NBNSYS0100049776	NI209A	*Nitzschia incognita*	3.0	3.0	
NBNSYS0100049777	NI043A	*Nitzschia inconspicua*	4.0	4.0	4.9
NBNSYS0100049781	NI044A	*Nitzschia intermedia*	3.0	5.0	
NBNSYS0100049786	NI198A	*Nitzschia lacuum*	2.0	2.0	
NBNSYS0100049787	NI123A	*Nitzschia laevis*	3.0	3.0	3.8
NHMSYS0020749185	NI127B	*Nitzschia levidensis* var. *salinarum*	5.0	5.0	
NBNSYS0100049792	NI203A	*Nitzschia liebetruthii*	1.0	2.0	
NBNSYS0100049794	NI031A	*Nitzschia linearis*	4.0	4.0	3.7
NHMSYS0020749186	NI129A	*Nitzschia littoralis*	4.0	4.0	
NHMSYS0021166387	NI131A	*Nitzschia lorenziana*	3.0	3.0	1.2
NBNSYS0100049800	NI027A	*Nitzschia microcephala*	3.0	3.0	4.7

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100049804	ZZZ980	*Nitzschia nana*	5.0	4.0	
NBNSYS0100049811	NI009A	*Nitzschia palea*	4.0	4.0	3.3
NBNSYS0100049812	NI009C	*Nitzschia palea* var. *debilis*	3.0	3.0	
NBNSYS0100049814	NI033A	*Nitzschia paleacea*	3.0	3.0	2.4
NBNSYS0100049815	NI139A	*Nitzschia paleaeformis*	3.0	5.0	
NBNSYS0100049817	NI143A	*Nitzschia parvula*	5.0	5.0	
NBNSYS0100049819	NI005A	*Nitzschia perminuta*	3.0	3.0	3.0
NBNSYS0100049826	NI150A	*Nitzschia pumila*	4.0	3.0	
NBNSYS0100049827	NI152A	*Nitzschia pusilla*	4.0	4.0	3.7
NBNSYS0100049828	NI025A	*Nitzschia recta*	4.0	4.0	3.6
NBNSYS0100049834	NI006A	*Nitzschia sigma*	4.0	4.0	3.5
NBNSYS0100049838	NI046A	*Nitzschia sigmoidea*	3.0	5.0	4.5
NBNSYS0100049840	NI164A	*Nitzschia sinuata*	4.0	5.0	
NBNSYS0100049843	NI166A	*Nitzschia sociabilis*	4.0	4.0	1.8
NBNSYS0100049777	NITZ-03	*Nitzschia soratensis*	4.0	4.0	3.8
NBNSYS0100049690	NI9999	*Nitzschia* sp.	3.0	3.0	3.6
NBNSYS0100049849	NI171A	*Nitzschia subacicularis*	4.0	4.0	
NBNSYS0100049853	NI024A	*Nitzschia sublinearis*	4.0	3.0	2.9
NBNSYS0100049856	NI195A	*Nitzschia supralitorea*	5.0	5.0	4.3
NBNSYS0100049862	NI048A	*Nitzschia tubicola*	4.0	4.0	2.6
NBNSYS0100049863	NI184A	*Nitzschia umbonata*	3.0	5.0	
NBNSYS0100049868	NI049A	*Nitzschia vermicularis*	4.0	5.0	
NHMSYS0020954130	NI214A	*Nitzschia wuellerstorffii*	5.0	5.0	
NHMSYS0020954152	PARL-01	*Parlibellus protracta*	4.0	4.0	4.0
NBNSYS0100050611	PE002A	*Peronia fibula*	1.0	1.0	-0.6
NBNSYS0100050841	PI015A	*Pinnularia abaujensis*	2.2	2.2	
NBNSYS0100050847	PINN-06	*Pinnularia acrosphaeria*	2.2	2.2	1.9
NBNSYS0100050848	PINN-24	*Pinnularia acuminata*	2.2	2.2	2.0

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NHMSYS0021167880	PINN-20	*Pinnularia anglica*	2.2	2.2	2.7
NBNSYS0100050852	PI014A	*Pinnularia appendiculata*	2.2	2.2	
NBNSYS0100050861	PI012A	*Pinnularia borealis*	2.2	2.2	4.9
NHMSYS0021167886	PINN-25	*Pinnularia borealis* var. *subislandica*	2.2	2.2	2.7
NBNSYS0100050868	PI048A	*Pinnularia brebissonii*	2.2	3.0	3.1
NBNSYS0100050898	PI016A	*Pinnularia divergentissima*	2.2	1.2	
NHMSYS0021166216	PINN-03	*Pinnularia grunowii*	2.2	2.2	2.4
NHMSYS0021166331	PINN-08	*Pinnularia isselana*	2.2	2.2	1.5
NBNSYS0100050949	PI132A	*Pinnularia lundii*	2.2	2.2	
NBNSYS0100050956	PI011A	*Pinnularia microstauron*	2.2	2.2	0.7
NHMSYS0021166217	PINN-05	*Pinnularia neomajor*	2.2	2.2	2.6
NBNSYS0100050962	PINN-17	*Pinnularia nodosa*	2.2	2.2	1.6
NHMSYS0021166332	PINN-10	*Pinnularia parvulissima*	2.2	2.2	1.4
NHMSYS0021167904	PINN-22	*Pinnularia peracuminata*	2.2	2.2	3.9
NBNSYS0100050972	PI056A	*Pinnularia rupestris*	2.2	4.0	2.7
NBNSYS0100050825	PI9999	*Pinnularia* sp.	2.2	2.2	4.0
NBNSYS0100050982	PI024A	*Pinnularia stomatophora*	2.2	2.2	-0.1
NBNSYS0100050988	PI022A	*Pinnularia subcapitata*	2.2	1.2	0.7
NHMSYS0021166335	PINN-13	*Pinnularia subcommutata* var. *nonfasciata*	2.2	2.2	0.7
NHMSYS0021166336	PINN-14	*Pinnularia subgibba*	2.2	2.2	1.2
NBNSYS0100050997	PI054A	*Pinnularia sudetica*	2.2	3.0	
NBNSYS0100051000	PI164A	*Pinnularia termitina*	2.2	2.2	2.2
NHMSYS0021166400	PINN-16	*Pinnularia viridiformis*	2.2	2.2	4.7
NBNSYS0100051004	PI007A	*Pinnularia viridis*	2.2	2.2	1.0
NHMSYS0021166338	PLAC-09	*Placoneis abiskoensis*			1.4
NBNSYS0100051065	ZZZ872	*Placoneis clementis*	4.0	4.0	1.8

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NHMSYS0021166291	PLAC-10	*Placoneis constans*	2.0	2.0	1.4
NBNSYS0100051067	ZZZ871	*Placoneis elginensis*	5.0	3.0	1.8
NBNSYS0100051063	ZZZ848	*Placoneis* sp.	2.0	3.0	3.9
NHMSYS0021166339	PLAN-07	*Planothidium caputium*	4.0	4.0	3.8
NHMSYS0000523996	ZZZ906	*Planothidium daui*	2.0	2.0	
NBNSYS0100051125	ZZZ905	*Planothidium delicatulum*	5.0	5.0	
NBNSYS0100051126	ZZZ895	*Planothidium dubium*	3.0	3.0	
NBNSYS0100051130	ZZZ896	*Planothidium frequentissimum*	3.0	3.0	3.9
NHMSYS0020063134	ZZZ901	*Planothidium granum*	5.0	5.0	
NBNSYS0100051134	ZZZ897	*Planothidium lanceolatum*	4.0	4.0	3.4
NBNSYS0100051137	ZZZ893	*Planothidium rostratum*	5.0	5.0	
NBNSYS0100051123	ZZZ922	*Planothidium* sp.	4.0	4.0	3.8
NHMSYS0020954177	PLAT-01	*Platessa conspicua*	5.0	5.0	5.8
NHMSYS0021166239	PRES-01	*Prestauroneis integra*		5.0	3.9
NBNSYS0100051674	ZZZ910	*Psammothidium bioretii*	1.0	1.0	1.2
NHMSYS0020749194	ZZZ950	*Psammothidium chlidanos*	2.0	2.0	1.0
NHMSYS0021166340	PSAM-06	*Psammothidium daonense*	2.0	2.0	1.1
NHMSYS0020954187	ZZZ907	*Psammothidium grishunun fo. daonensis*	2.0	2.0	
NBNSYS0100051681	ZZZ852	*Psammothidium helveticum*	2.0	2.0	
NHMSYS0000524017	ZZZ920	*Psammothidium lauenburgianum*	5.0	5.0	
NBNSYS0100051686	ZZZ855	*Psammothidium levanderi*	3.0	3.0	
NBNSYS0100051688	PSAM-05	*Psammothidium pseudoswazi*			-0.3

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100051691	ZZZ883	*Psammothidium rossii*	2.0	2.0	
NBNSYS0100051670	ZZZ921	*Psammothidium* sp.	2.0	2.0	
NBNSYS0100051696	ZZZ949	*Psammothidium subatomoides*	2.0	2.0	
NHMSYS0020954190	PSEU-01	*Pseudostaurosira / Staurosira* agg.	3.7	4.0	
NBNSYS0100051799	PS001A	*Pseudostaurosira brevistriata*	3.7	4.0	3.6
NHMSYS0020954192	PSEU-03	*Pseudostaurosira microstriata*	3.7	4.0	
NBNSYS0100051803	PS002A	*Pseudostaurosira pseudoconstruens*	3.7	4.0	
NBNSYS0100051971	RE001A	*Reimeria sinuata*	3.0	3.0	2.8
NBNSYS0100051970	REIM-01	*Reimeria* sp.	3.0	3.0	
NHMSYS0001388927	ZZZ926	*Reimeria uniseriata*	3.0	3.0	
NBNSYS0100052065	RC002A	*Rhoicosphenia abbreviata*	4.0	4.0	4.4
NBNSYS0100052075	RH001A	*Rhopalodia gibba*	1.0	1.0	0.9
NHMSYS0021166341	ROSS-03	*Rossithidium anastasiae*	1.0	1.0	0.0
NBNSYS0100052171	ZZZ859	*Rossithidium linearis*	1.0	1.0	
NHMSYS0020954205	ZZZ888	*Rossithidium petersenii*	1.0	1.0	
NBNSYS0100052173	ZZZ885	*Rossithidium pusillum*	1.0	2.0	
NBNSYS0100052170	ROSS-01	*Rossithidium* sp.	1.0	1.0	
NHMSYS0021166349	SELL-13	*Sellaphora auldreekie*	4.0	4.0	3.7
NBNSYS0100052452	ZZZ925	*Sellaphora bacillum*	3.0	3.0	5.0
NHMSYS0021166350	SELL-14	*Sellaphora blackfordensis*	4.0	4.0	3.0
NHMSYS0020954214	SELL-01	*Sellaphora joubaudii*	4.0	4.0	4.9
NBNSYS0100052456	ZZZ864	*Sellaphora laevissima*	5.0	5.0	3.8

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NHMSYS0021166353	SELL-17	*Sellaphora lanceolata*	4.0	4.0	3.7
NBNSYS0100052460	SL001A	*Sellaphora pupula*	4.0	4.0	3.6
NHMSYS0021166348	SELL-12	*Sellaphora rotunda*	2.0	1.0	
NBNSYS0100052466	SL002A	*Sellaphora seminulum*	4.0	4.0	4.0
NBNSYS0100052450	SL9999	*Sellaphora* sp.	4.0	4.0	
NHMSYS0021166346	SELL-10	*Sellaphora subrotundata*		4.0	
NHMSYS0021166358	SEMI-02	*Seminavis robusta*			4.3
NBNSYS0100052556	SIM01A	*Simonsenia delognei*	5.0	5.0	
NBNSYS0100053221	SA001A	*Stauroneis anceps*	2.0	2.0	2.8
NHMSYS0021166296	STAR-02	*Stauroneis gracilior*			3.2
NBNSYS0100053239	SA012A	*Stauroneis kriegeri*	2.0	2.0	2.4
NBNSYS0100053245	SA005A	*Stauroneis legumen*	4.0	4.0	
NBNSYS0100053254	SA006A	*Stauroneis phoenicenteron*	4.0	4.0	3.2
NHMSYS0021167976	STAR-04	*Stauroneis schmidiae*			5.3
NBNSYS0100053265	SA003A	*Stauroneis smithii*	4.0	4.0	
NBNSYS0100053215	SA9999	*Stauroneis* sp.	3.0	3.0	
NBNSYS0100053271	SA068A	*Stauroneis thermicola*	3.0	3.0	
NBNSYS0100053280	SR001A	*Staurosira construens*	3.7	4.0	2.5
NBNSYS0100053287	SR002A	*Staurosira elliptica*	3.7	4.7	3.8
NHMSYS0020063135	ZZZ880	*Staurosira oldenburgiana*	3.7	4.0	
NBNSYS0100053279	SR9999	*Staurosira* sp.	3.7	4.0	2.6
NHMSYS0021166359	STAU-03	*Staurosira venter*	3.7	4.7	5.0
NBNSYS0100053292	STAS-01	*Staurosirella lapponica*	5.0	5.0	4.3
NBNSYS0100053293	SS003A	*Staurosirella leptostauron*	4.0	4.0	
NBNSYS0100053295	SS002A	*Staurosirella pinnata*	4.0	4.0	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100053325	SP006A	*Stenopterobia curvula*			0.2
NBNSYS0100053323	SP9999	*Stenopterobia* sp.			4.8
NBNSYS0100053502	SU017A	*Surirella amphioxys*	3.0	3.0	
NBNSYS0100053503	SU001A	*Surirella angusta*	3.0	3.0	3.7
NBNSYS0100053507	SU074A	*Surirella bifrons*			4.2
NBNSYS0100053509	SU004A	*Surirella biseriata*			3.4
NBNSYS0100053515	SU073A	*Surirella brebissonii*	3.0	3.0	3.2
NBNSYS0100053518	SU022A	*Surirella brightwellii*	3.0	3.0	
NBNSYS0100053520	SU024A	*Surirella capronii*	2.0	2.0	4.3
NHMSYS0021166363	SURI-02	*Surirella costata*	2.0	2.0	5.5
NBNSYS0100053525	SU032A	*Surirella crumena*	3.0	3.0	
NHMSYS0021166365	SURI-04	*Surirella iconella*	2.0	2.0	4.0
NBNSYS0100053538	SU005A	*Surirella linearis*	1.0	1.0	
NHMSYS0021166401	SU005C	*Surirella linearis* var. *helvetica*	2.0	2.0	4.5
NHMSYS0021166367	SURI-06	*Surirella lineopunctata*	2.0	4.0	5.8
NBNSYS0100053541	SU016A	*Surirella minuta*	4.0	4.0	
NBNSYS0100053544	SU003A	*Surirella ovalis*	5.0	5.0	4.3
NBNSYS0100053555	SU076A	*Surirella roba*	2.0	2.0	
NBNSYS0100053498	SU9999	*Surirella* sp.	2.0	2.0	3.9
NBNSYS0100053561	SURI-07	*Surirella splendida*	2.0	2.0	4.6
NHMSYS0021166362	SURI-01	*Surirella tenuissima*	2.0	2.0	2.0
NHMSYS0020749201	ZZZ923	*Surirella terricola*	4.0	4.0	
NBNSYS0100053633	SYND-07	*Synedra bacillaris*			1.5
NHMSYS0020954244	SYNE-01	*Synedrella parasitica*	5.0	5.0	
NHMSYS0020954245	SYNE-02	*Synedrella subconstricta*	5.0	4.0	
NBNSYS0100053703	TA002A	*Tabellaria fenestrata*	1.0	1.0	
NBNSYS0100053704	TA001A	*Tabellaria flocculosa*	1.0	1.0	-0.3
NBNSYS0100053706	TA004A	*Tabellaria quadriseptata*	1.0	1.0	
NBNSYS0100053702	TA9999	*Tabellaria* sp.	1.0	1.0	

NBN code*	Taxon ID	Taxon name	TDI4	TDI5LM	TDI5NGS
NBNSYS0100053707	TA006A	*Tabellaria ventricosa*	1.0	2.0	
NBNSYS0100053710	TU003A	*Tabularia fasciculata*	4.0	4.0	
NHMSYS0021166402	TABU-01	*Tabularia tabulata*			3.7
NBNSYS0100053851	TE9999	*Tetracyclus* sp.			3.5
NBNSYS0100054309	TF001A	*Tryblionella acuminata*	5.0	5.0	
NBNSYS0100054311	TF003A	*Tryblionella apiculata*	5.0	5.0	4.5
NBNSYS0100054313	TF013A	*Tryblionella calida*	5.0	5.0	
NBNSYS0100054306	ZZZ985	*Tryblionella debilis*	4.0	4.0	3.3
NBNSYS0100054315	TF015A	*Tryblionella hungarica*	4.0	4.0	
NHMSYS0000524179	TF005A	*Tryblionella levidensis*	5.0	5.0	
NBNSYS0100054303	TF9999	*Tryblionella* sp.	4.0	4.0	4.8
NHMSYS0021166219	ULNA-01	*Ulnaria acus*	3.0	3.0	2.9
NHMSYS0021166267	ULNA-03	*Ulnaria* sp.	3.0	3.0	
NHMSYS0021166396	ULNA-02	*Ulnaria ulna*	2.0	3.0	1.9

Notes: This list only contains taxa identified and analysed as part of this project. An electronic version containg a list of all taxa is contained in the R package DARLEQ3 available at https://github.com/nsj3/darleq3.
*National Biodiversity Network

Appendix 3: List of lakes included in analyses in Section 4

Water body ID	BIOSYS site ID	Water body name	Type
29025	154819	Balderhead Reservoir	LA
28847	108582	Bassenthwaite lake	MA
28172	102044	Broomlee Lough	HA
46279	141857	Burrator Reservoir	LA
29052	141796	Buttermere	LA
43096	157550	Chew Valley Lake	HA
29321	102076	Coniston Water	LA
28220	101997	Crag Lough	HA
35211	149999	Crose Mere	HA
29000	141794	Crummock Water	LA
32359	155189	Derwent Reservoir	LA
28965	154784	Derwent Water (Cumbria)	LA
46232	141855	Dozmary Pool	LA
29222	102083	Elter Water	MA
29062	141730	Ennerdale Water	LA
29328	102075	Esthwaite Water	MA
44031	146017	Frensham Great Pond	HA
29184	133121	Grasmere	MA
28165	102064	Greenlee Lough	MA
28165	102064	Greenlee Lough	MA
29647	102069	Hawes Water	HA
29073	154960	Haweswater Reservoir	LA
35640	152464	Hickling Broad	HA
35640	152465	Hickling Broad	HA
35640	152466	Hickling Broad	HA
30244	101704	Hornsea Mere	HA
46102	145313	Little Sea Mere	MA
46556	99581	Loe Pool	MA

Water body ID	BIOSYS site ID	Water body name	Type
28986	154704	Loweswater	MA
28806	154959	Overwater Reservoir	MA
47017	182026	Rockford Lake	MA
32650	102281	Rostherne Mere	HA
32650	145432	Rostherne Mere	HA
29639	154550	Scar House Reservoir	LA
46472	102541	Slapton Ley	HA
45790	146261	Sowley Pond Lake	MA
32804	102321	Tatton Mere	HA
37306	155072	Thompson Water	HA
28395	154789	Tindale Tarn	MA
28955	146503	Ullswater (North Basin)	MA
36202	150728	Upton Broad	HA
36202	152443	Upton Great Broad	HA
29183	141660	Wastwater	LA
44310	155300	Weir Wood Nature Reserve	HA
29233	102078	Windermere	MA
29233	151706	Windermere	MA
35953	147820	Wroxham Broad	HA

Notes: HA = high alkalinity; LA = low alkalinity; MA = medium alkalinity

Appendix 4: Interpreting TDI5NGS data

The purpose of this appendix is to help end users interpret next generation sequencing (NGS) outputs from DARLEQ3 (https://github.com/nsj3/darleq3).

A.4.1 Introduction

DARLEQ3 offers the capability to perform ecological assessments using data generated by either light microscopy (LM) or NGS. But because the 2 methods will not necessarily give identical results when applied to the same sample, users of DARLEQ3 need to understand:

- how NGS data differ from LM data
- what this means for interpreting ecological status

When considering NGS data for the first time, it is useful to bear in mind the limitations of current methods based on LM (see Box A.4.1). LM based analysis is not perfect, but it is a method that biologists have grown to understand over the years. All ecological assessment methods have limitations and offer insights into the condition of a water body 'as if through a glass darkly'. A clearer view of ecological status is built up by collecting information from a range of different biological, chemical and physical components of a water body over time.

NGS analysis simply offers a different way of generating information about the status of the phytobenthos. While some aspects of the NGS method might offer a clearer view, there will also be information that can be gleaned from LM analysis that cannot (yet) be duplicated with NGS. In the short term, however, it is necessary to understand that NGS data are different to LM data. These differences do not mean that NGS data are wrong, just that it is important to learn to interpret these new data and perhaps to forget some of the preconceptions brought along from interpreting LM data.

The first 3 bullet points in Box A.4.1 apply to assessment of phytobenthos status using NGS as well as to the LM based method. Although the NGS method does not consider cell size, it is possible that the number of rbcL reads offers a more direct measure of the contribution that each species makes to primary productivity (see below). In addition, it is known that DNA can survive outside the cell for some time and so presence in a sample analysed by NGS does not necessarily equate to the presence of a viable population. However, the DNA is less persistent than the silica frustules (diatom cell walls), and so NGS results are likely to give a more direct insight into which species were alive at the time of sampling than LM results.

Box A.4.1 Limitations of LM diatom analysis for ecological status assessment

- Does not capture all phytobenthos diversity
- Assessments are based on lists of species, with no consideration of functional properties or productivity
- Limited quantification (relative not absolute abundance)

- No consideration of cell size
- Cannot differentiate live from dead cells

A.4.2 Sample size

Figure A.4.1 shows the number of reads per species for 6 NGS samples selected at random from the dataset from which DARLEQ3 was developed. It illustrates the following 3 important differences between data generated by NGS and LM:

- NGS samples contain much more potential information than LM samples. It is common for the output from NGS to include over 10,000 separate ‘reads’. In theory, it is possible to identify and count this number of diatoms using LM. However, this would take an extraordinary length of time and, in practice, most analysts name and count between 300 and 500 valves.
- More species are generally recorded using NGS rather than LM. Most samples identified using LM have between 20 and 40 taxa, whereas samples analysed using NGS can have 60 or more. This is partly a consequence of the greater amount of data that are generated. It is also related to the bioinformatics pathways that are used (that is, how stringent are the filters that match reads to species in the barcode database). The size of the barcode database will also be a factor contributing to the number of species recorded.
- Although more species are recorded by NGS, there is a long ‘tail’ of species represented by just a small number of reads. If a typical sample consists of 30,000 reads, then anything with less than 300 reads forms only 1% of the total and will be unlikely to have a major effect on indices based on a weighted averaging equation. Anything with less than 100 reads is unlikely to be detected by a LM analyst. It is also not possible to be sure that taxa represented by a small number of reads represent a viable population living at the site at the time the sample was collected. It is possible that the sample includes some ‘eDNA’ – molecules that are suspended in the river water or tangled in the biofilm but which derive from populations elsewhere in the catchment. Similarly, it is not possible to be sure that very rare diatoms detected by LM represent viable populations rather than dead cells that had drifted into the biofilm from upstream.

A final point that DARLEQ3 users need to understand is that a large number of the total reads (40% on average) are not assigned to species and play no role in assessments. This is partly a consequence of the limited size of the current barcode database and this proportion should decrease as the barcode database increases in size.

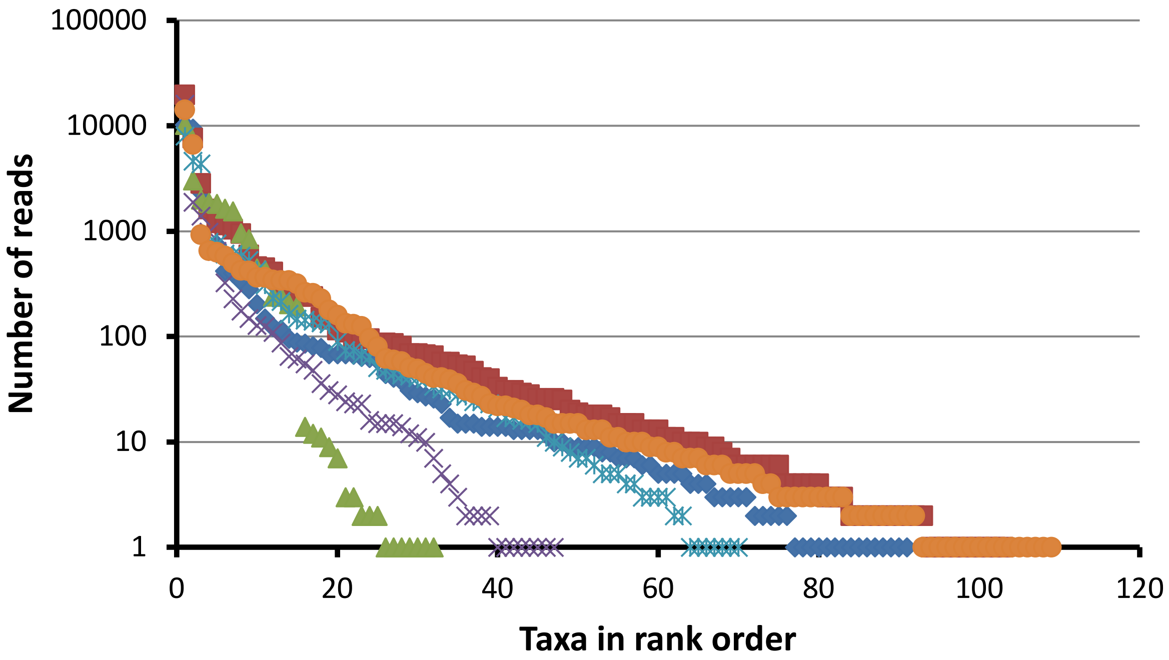

Figure A.4.1 Species abundance curves for 6 NGS samples, selected at random, to illustrate the properties of NGS data

Notes: Species are shown in rank order, with the most abundant on the left.

A.4.3 Expression of individual species

The standard unit of enumeration in LM analyses in the UK and several other countries is the valve (that is, half the cell wall or frustule). However, diatoms can vary considerably in size, both within the cell cycle and between species. Figure A.4.2 shows one of the larger diatoms common in UK waters (*Ulnaria ulna*) alongside one of the smaller ones (*Achnanthidium minutissimum*). The difference in cell volume is 100 times, and it can be assumed that the larger cell contributes substantially more to primary productivity in a sample than the smaller. However, each makes the same contribution to the LM analysis.

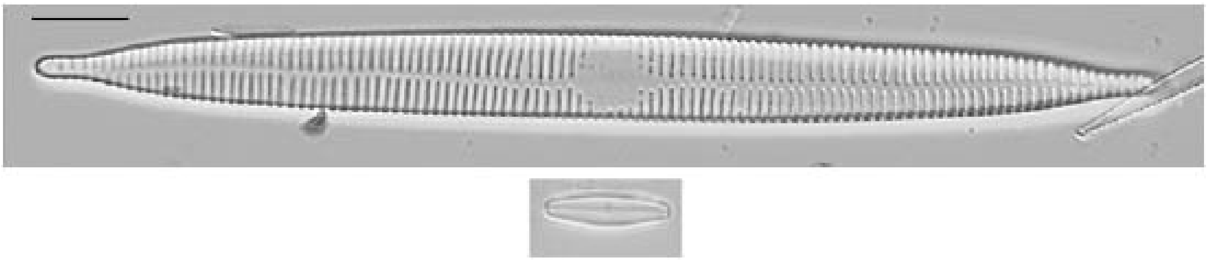

Figure A.4.2 Specimens of *Ulnaria ulna* (top) and *Achnanthidium minutissimum* (bottom)

Notes: Both specimens are from cultures used for obtaining sequences for the barcode database.
Scale bar: 10 µm.
Photographs: Shinya Sato.

Each rbcL read in an NGS analysis represents one copy of the gene that encodes for ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCo), an important enzyme which catalyses the chemical reaction by which inorganic carbon is captured by the chloroplast at the start of the photosynthesis pathway. Consequently, an analysis based on rbcL reads should, in theory, give a better insight into the contribution each species makes to primary productivity than simply counting cell numbers. In practice, however, there is still much that is not understood about:

- the expression of rbcL in diatoms
- how the number of reads for any species relates to the abundance of that species in the original sample

There is some evidence that:

- larger cells have more rbcL reads than smaller ones
- cells with many chloroplasts have more rbcL than cells with single chloroplasts

It is also possible that:

- chloroplast shape influences the number of reads
- read number can vary depending on environmental conditions and through the cell cycle

In addition, the number of chloroplast varies between different groups of diatoms (Table A.4.1).

Table A.4.1 Variation in chloroplast numbers between major groups of diatoms

Group	Number of chloroplasts
Centric diatoms	Mostly many per cell
Araphid diatoms	Many genera have 1 or 2 per cell (for example, *Fragilaria, Hannaea*); a few have many per cell (*Tabellaria, Fragilariforma, Asterionella*)
Raphid diatoms	Most have 1 or 2 per cell; a few have four (*Neidium, Fistulifera*)

Figure A.4.3 shows how the expression of 6 common species differs between LM and NGS. Figures A.4.3a and A.4.3b show *Achnanthidium minutissimum* and *Amphora pediculus*; these small pioneer species each have a single chloroplast and both tend to form a greater part of the LM than the NGS analysis. In contrast*, Navicula lanceolata* (Figure A.4.3c) is a larger diatom with 2 chloroplasts and the proportion recorded in NGS tends to be greater than in LM. *Melosira varians* (Figure A.4.3d) shows a more extreme situation, with proportions in NGS almost always much greater than in LM. This is a species with many chloroplasts, each of which will be contributing to the total number of rbcL copies in the cell. Finally, *Fistulifera saprophila* (Figure A.4.3e) is a very small, weakly silicified diatom with 4 chloroplasts. The higher proportions in NGS may reflect underreporting in LM analyses, particularly if cells do not survive the digestion process, and possibly misidentification with other small species such as *Mayamaea atomus* var. *permitis* (Figure A.4.3f).

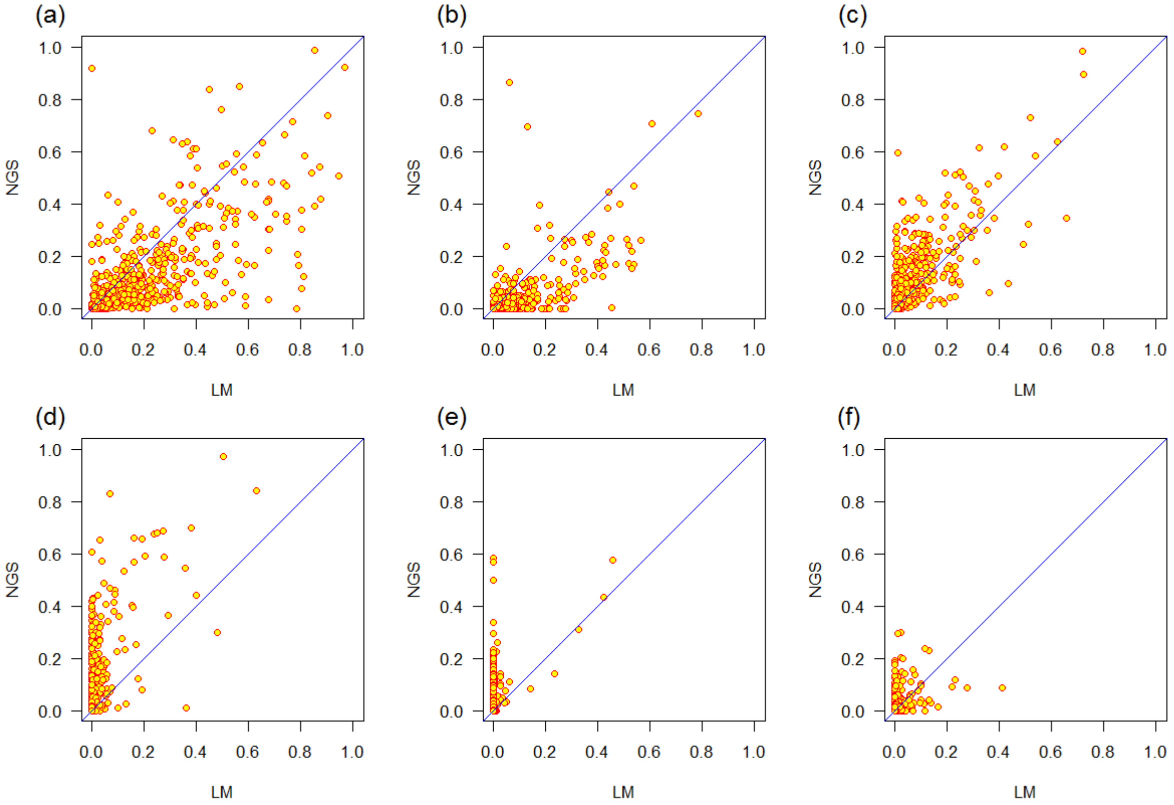

Figure A.4.3 Differences between representation of common taxa in LM (x axis) and NGS (y axis) on a proportional scale: (a) *Achnanthidium minutissimum* type (small, one chloroplast); (b) *Amphora pediculus* (small, one chloroplast); (c) *Navicula lanceolata* (medium sized, 2 chloroplasts); (d) *Melosira varians* (large, many chloroplasts); (e) *Fistulifera saprophila* (very small, 4 chloroplasts, weakly silicified); and (f) *Mayamaea atomus* including var. *permitis* (very small, possibly 2 chloroplasts, weakly silicified)

Notes: The diagonal line shows slope = 1 (that is, equal representation in LM and NGS).

Source: Environment Agency (2018, Figure 6.3).

A.4.4 Interpreting TDI5NGS

Biologists are still learning how to interpret NGS outputs. Problems will be particularly acute in the period following the transition from LM to NGS as users will have to reconcile results produced with NGS with older data collected using LM. This is discussed more in Section A.4.5. The following pointers should help users to understand their NGS output.

A.4.4.1 Cell size and chloroplast number

Cell size and chloroplast number play an important role in determining the representation of a taxon in NGS outputs.

- Do not over-interpret the presence of taxa that are represented by a small number of reads.
- Use the following values as approximate detection limits for presence:
  - Large taxa and those with many chloroplasts: 50–100 reads
  - Other taxa: 10 reads

A.4.4.2 Know your catchment

This applies to all data interpretation, not just to diatoms analysed by NGS. In the case of NGS data, however, it is important to be aware that:

- the sample may contain eDNA from upstream sources
- planktonic taxa may behave differently in NGS compared with LM

Therefore, consider the state of the river upstream when interpreting NGS data, bearing in mind geological changes that might influence the species that are found in different parts of the catchment. Also, look to see if there are fish farms, lakes or ponds that may serve as inocula of planktic taxa to the stream.

A.4.4.3 Gaps in the barcode database

About 2,800 diatom species have been recorded from Britain and Ireland but only around 350 are currently represented in the barcode database. Many of these are only represented by a few barcode sequences, and so it is not possible to be sure that all of the genetic variation within some species complexes will be detected. On average, about 40% of rbcL reads in each NGS analysis cannot be assigned to a species. These issues are likely to be more important when looking in detail at trends over time

Table A.4.2 lists taxa that are abundant in LM analyses but which are not, as yet, represented in the barcode database.

Table A.4.3 lists taxa that are abundant in LM analyses but which have <5 DNA barcode sequences in the barcode database. This is offered as a rough indication of the depth of coverage of each species but needs to be interpreted with caution. *Navicula lanceolata*, for example, is represented by 45 sequences but none differ by more than 3 base pairs across the whole rbcL gene. On the other hand, the *Achnanthidium minutissimum* complex is represented by over 85 sequences, with considerable variation (5% variability between barcodes in the database representing 12 different strains or genotypes), despite not fully capturing all the morphological variation apparent in field material. Several important groups (for example, *Cocconeis placentula* complex) are represented by just a few sequences.

Table A.4.2 List of taxa that have been recorded at a relative abundance of 5% or more in LM analyses but which are missing from the barcode database

Achnanthidium caledonicum	*Fragilaria delicatissima*	*Navicula tenelloides*
Achnanthidium catenatum	*Fragilaria mesolepta*	*Navicula(dicta) schmassmannii*
Achnanthidium subatomus	*Fragilaria recapitellata*	*Nitzschia archibaldii*
Adlafia suchlandtii	*Fragilaria tenera*	*Nitzschia brevissima*
Amphora inariensis	*Fragilariforma* sp.	*Nitzschia disputata*
Brachysira brebissonii	*Frustulia krammeri*	*Nitzschia lacuum*
Caloneis bacillum	*Geissleria schoenfeldii*	*Nitzschia levidensis* var. *salinarum*
Delicata delicatula	*Gomphonema exilissimum*	*Nitzschia liebetruthii*
Denticula tenuis	*Gomphonema olivaceoides*	*Nitzschia umbonata*
Diatoma ehrenbergii	*Gomphonema olivaceum*	*Nupela lapidosa*
Diatoma mesodon	*Gomphonema tergestinum*	*Pinnularia appendiculata*
Diatoma problematica	*Gomphonema varioreduncum*	*Planothidium dubium*
Diploneis sp.	*Gomphosphenia grovei*	*Planothidium granum*
Encyonema gracile	*Karayevia clevei*	*Psammothidium helveticum*
Encyonema reichardtii	*Karayevia laterostrata*	*Psammothidium lauenburgianum*
Epithemia adnata	*Kolbesia kolbei*	*Psammothidium* sp.
Epithemia sorex	*Kolbesia ploenensis*	*Psammothidium subatomoides*
Eucocconeis flexella	*Luticola mutica*	*Rossithidium linearis*
Eunotia muscicola	*Mayamaea atomus*	*Rossithidium petersenii*
Eunotia paratridentula	*Mayamaea lacunolaciniata*	*Staurosirella pinnata*
Eunotia subarcuatoides	*Meridion circulare* var. *constrictum*	*Surirella linearis*
Fallacia subhamulata	*Navicula claytonii*	*Surirella ovata* var. *minuta*
Fistulifera / Mayamaea	*Navicula ingenua*	*Surirella roba*
Fragilaria amphicephala	*Navicula menisculus*	*Tabellaria ventricosa*
Fragilaria austriaca	*Navicula reichardtiana*	*Simonsenia delognei*

Table A.4.3 List of taxa that have been recorded at a relative abundance of 5% or more in LM analyses but which are represented by ≤5 barcode sequences in the database

Amphora copulata	*Fragilaria famelica*	*Nitzschia capitellata*
Brachysira neoexilis	*Frustulia vulgaris*	*Nitzschia dissipata*
Brachysira vitrea type	*Gomphonema ‘intricatum’ type*	*Nitzschia filiformis*
Cocconeis pediculus	*Gomphonema angustatum*	*Nitzschia frustulum*
Cocconeis placentula agg.	*Gomphonema clevei*	*Nitzschia paleacea*
Craticula accomoda	*Gomphonema gracile*	*Nitzschia pusilla*
Craticula molestiformis	*Gomphonema minutum*	*Nitzschia recta*
Craticula subminuscula	*Halamphora montana*	*Nitzschia sociabilis*
Ctenophora pulchella	*Halamphora oligotraphenta*	*Nitzschia* sp.
Diatoma tenue	*Halamphora veneta*	*Nitzschia sublinearis*
Diatoma vulgare agg.	*Hannaea arcus*	*Pinnularia subcapitata*
Didymosphenia geminata	*Karayevia oblongella*	*Planothidium frequentissimum*
Encyonema silesiacum	*Luticola goeppertiana*	*Platessa conspicua*
Encyonopsis microcephala	*Mayamaea atomus* var. *permitis*	*Psammothidium chlidanos*
Eolimna minima	*Meridion circulare*	*Psammothidium daonense*
Eunotia exigua	*Navicula capitatoradiata*	*Pseudostaurosira brevistriata*
Eunotia formica	*Navicula cari*	*Reimeria sinuata*
Eunotia implicata	*Navicula cincta*	*Rhoicosphenia abbreviata*
Eunotia minor	*Navicula cryptotenella*	*Sellaphora seminulum*
Eunotia pectinalis	*Navicula phyllepta*	*Staurosira construens*
Eunotia sp.	*Navicula slesvicensis*	*Staurosira elliptica*
Fallacia pygmaea	*Navicula tripunctata*	*Staurosira venter*
Fistulifera saprophila	*Navicula veneta*	*Surirella angusta*
Fragilaria bidens	*Navicula viridula*	*Tryblionella apiculata*
Fragilaria capucina	*Nitzschia acicularis*	*Tryblionella debilis*

A.4.4.4 Different in species behaviour in the 2 methods

Individual species may behave differently in NGS compared with LM.

A.4.4.5 Occasional 'misfires'

Both methods can produce occasional misfires.

For LM analyses, most analysts participated in a ring test scheme. However, there were instances where samples were contracted out to analysts who were not part of this scheme. Remember, too, that the ring test ensured the general competence of analysts rather than the quality of each individual analyst. When comparing data collected by LM and NGS, do not automatically assume that LM analyses are 'right' and NGS analyses are 'wrong'.

NGS analyses are subject to quality control before results are released and, if necessary, samples are re-run. Although this will catch most instances of rogue samples, treat samples with low numbers of reads (<3,000) with caution.

In over 80% of cases, the difference between LM and NGS analyses will be <10 TDI units. However, exceptions do occur (see Section A.4.5 for an example); care should therefore be taken if a TDI value computed with NGS data is very different (for example, >1 ecological status class) from what might be expected.

A.4.4.6 Limitations of current reference model

Both DARLEQ2 and DARLEQ3 use a reference model that is not very effective in hard water. These models should not therefore be used in water where alkalinity is >120mgL^{-1} CaCO$_3$. TDI4 and TDI5 may be useful in investigations in harder water, but should be interpreted with care.

Table A.4.4 compares LM and NGS results for one sample as an illustration of the practicalities of data interpretation. It is important to emphasise that not all differences can be readily explained. Why, for example, was *Nitzschia palea* abundant in LM but absent from NGS, despite a number of barcodes in the database? Similarly, *Cyclotella meneghiniana* should in theory have been more abundant in NGS than LM (it is a medium sized cell with many chloroplasts). Other differences, however, do match expectations, and the overall difference in TDI is within the expected range.

Table A.4.4 Comparison of TDI scores from LM and NGS data from River Browney, County Durham, B6301 bridge, August 2014

Species	LM	NGS	Comments
Achnanthidium minutissimum	29.2	4.2	← lower representation in NGS is typical for this species
Navicula gregaria	12.0	2.1	
Cyclotella meneghiniana	9.8	1.0	
Nitzschia palea	8.6	0.0	

Cocconeis placentula complex	8.3	2.1	
Rhoicosphenia abbreviata	6.2	0.0	← limited number of barcode sequences available for a morphologically diverse species complex
Amphora pediculus	4.0	9.4	
Melosira varians	4.0	25.0	← species with many chloroplasts: may explain greater abundance in NGS
Surirella brebissonii	3.7	10.4	← species with single large, lobed chloroplast: may explain greater abundance in NGS
Navicula tripunctata	2.8	2.1	
TDI	**57.4**	**67.7**	← difference of about 10 TDI units is within expected range

Notes: Only species present at >5% in at least one analysis are presented.

A.4.5 Effect of changing to NGS analyses on long-term trends in TDI

A very reasonable question to ask before adopting a NGS based diatom method is whether the change from LM to NGS will affect the classifications of water bodies. This question can only be answered where there are data showing a long-term trend based on LM plus sufficient NGS data to permit a comparison.

Project SC140024 generated NGS data over space and time for 4 water bodies in northern England for which long-term LM data were also available. These 4 rivers are considered below in order of decreasing ecological status.

A.4.5.1 River Wear, Wolsingham, County Durham

This site is located at the eastern edge of the Pennines and diatom based EQRs generally suggest high to good ecological status. Figure A.4.4 plots NGS samples collected throughout 2014 against LM data that extend back to 2004. The NGS data reflect this trend, with most samples reporting high status and 2 suggesting good status.

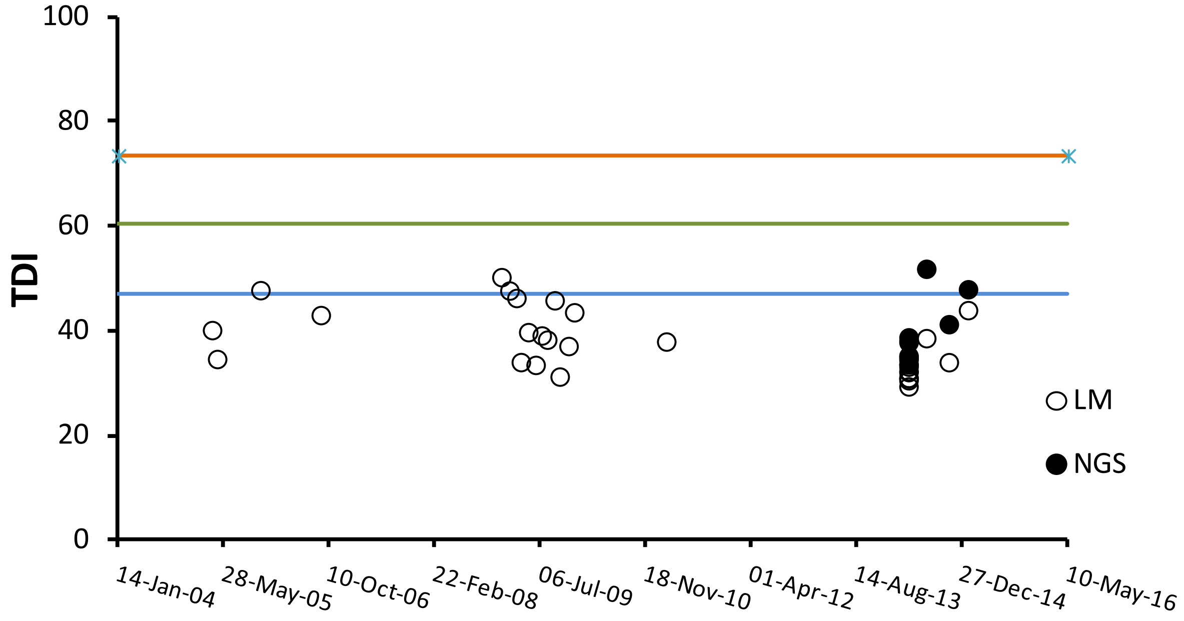

Figure A.4.4 Long-term trends in TDI scores in the River Wear at Wolsingham

Notes: Horizontal lines show the position of high to good (blue), good to moderate (green) and moderate to poor (orange) ecological status class boundaries.

A.4.5.2 River Ehen, just above Ennerdale Bridge, Cumbria

This is another high status site and, again, samples collected as part of SC140024 fit into the longer term trend of LM data from this site (Figure A.4.5). The alkalinity at this site is much lower, and so the ecological status class boundaries are correspondingly lower than in the River Wear.

The upper River Ehen has a challenging assemblage of diatoms that is responsible for more variation in LM analyses than is normal. The relatively consistent results for NGS may reflect some gaps in the barcode database rather than suggesting that the method is more reproducible than LM here.

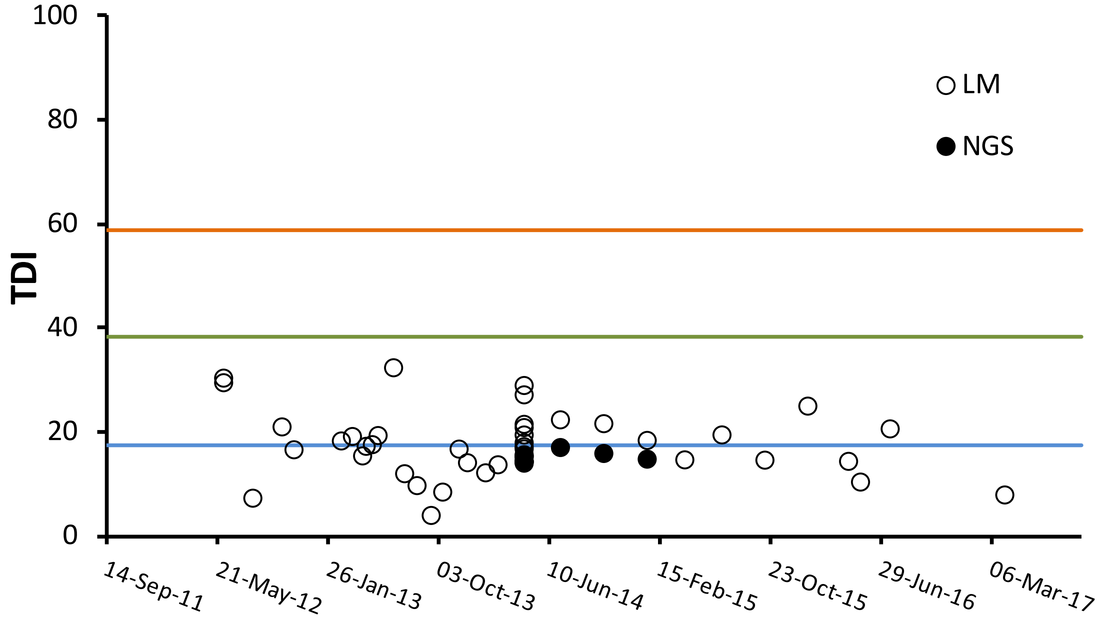

Figure A.4.5 Long-term trends in TDI scores in the River Ehen near Ennerdale Bridge

Notes: Horizontal lines show the position of high to good (blue), good to moderate (green) and moderate to poor (orange) ecological status class boundaries.

A.4.5.3 River Derwent, Ebchester, County Durham

The River Derwent, a tributary of the Tyne, also flows off the eastern Pennines. The sampling site used is downstream of Consett STW and the river shows signs of enrichment. Both LM and NGS analyses fluctuate across good and moderate ecological status, with occasional results in poor status (Figure A.4.6).

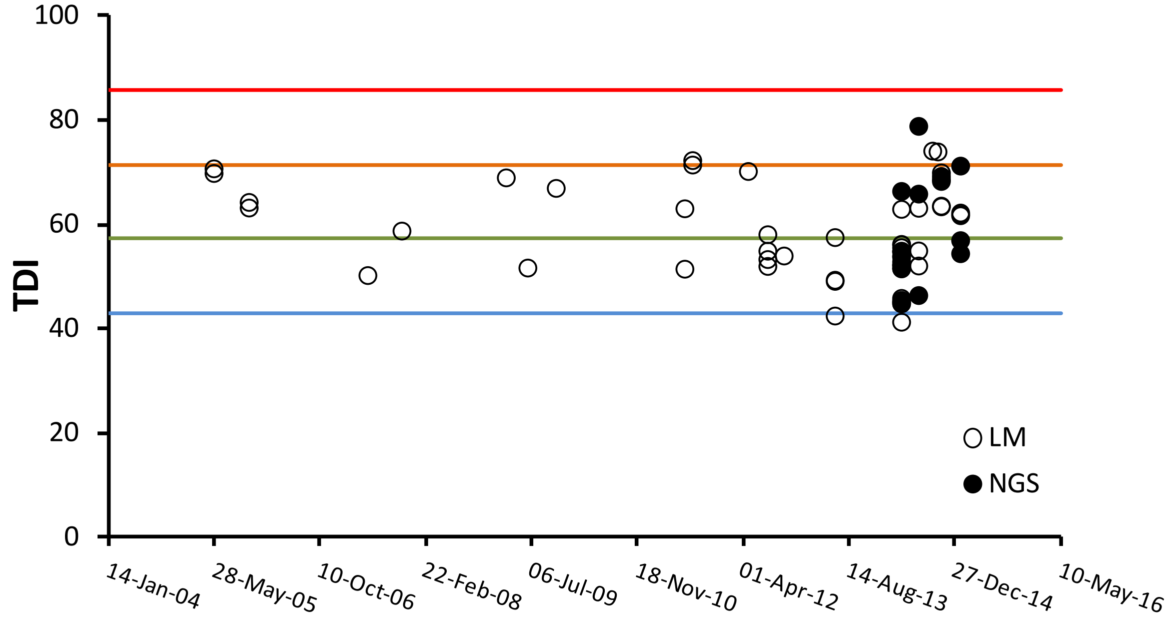

Figure A.4.6 Long-term trend in TDI scores in the River Derwent at Ebchester

Notes: Ecological status class boundaries as in Figure A.4.5, with the addition of poor to bad (red).

A.4.5.4 River Team, Causey Arch, County Durham

The River Team is a lowland tributary of the River Tyne that flows through a former industrial region with a variety of pollution sources including minewater, sewage and contaminated land. The river contains prolific growths of *Cladophora* and *Vaucheria* and, sometimes, sewage fungus. LM samples are consistently less than good ecological status, with some falling to poor status (Figure A.4.7). Most NGS samples follow this trend, but there were also a few outliers for reasons that cannot be fully explained (see Environment Agency 2018, Section 7). Some of the outliers, however, had very low read numbers following NGS. Following improved quality control (QC) procedures these samples would now fail QC and be reanalysed. Therefore, in this instance, getting classifications of good status from a river where all previous evidence points to less than good status should prompt further investigation of the data and the procedure leading to data generation.

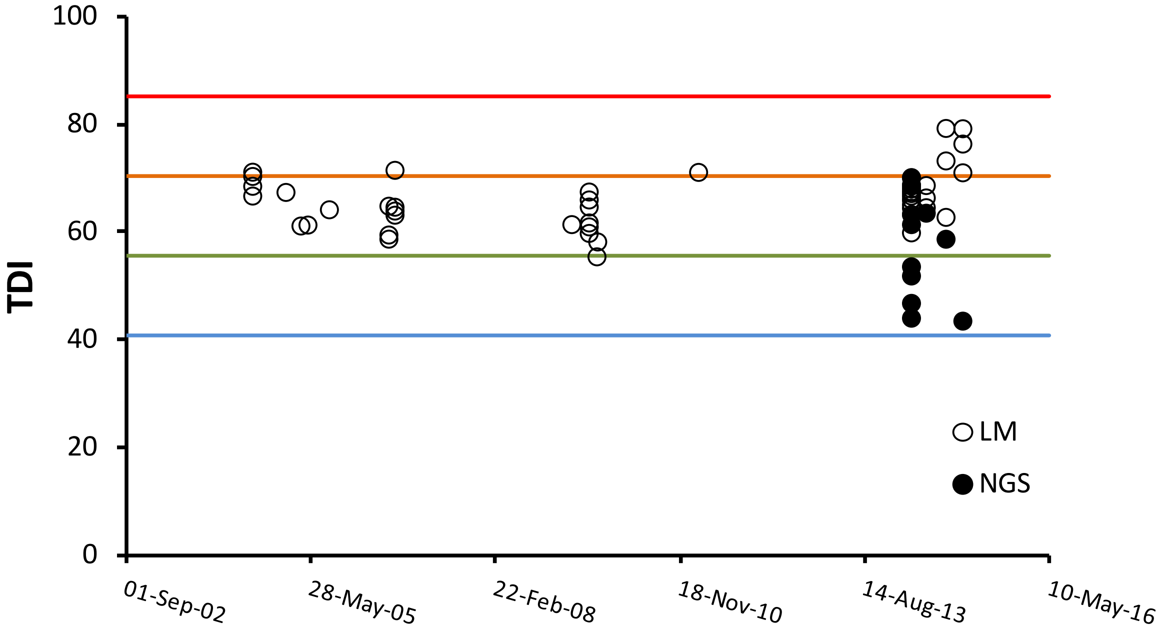

Figure A.4.7 Long-term trend in TDI scores in the River Team at Causey Arch

Notes: Ecological status class boundaries as in Figure A.4.6.

A.4.6 Other metrics

DARLEQ3, like earlier versions of the DARLEQ software, contains a number of metrics in addition to versions of the TDI, which can be useful when interpreting data.

A.4.6.1 Percentage planktic valves

This metric is the sum of all the individuals belonging to taxa that are predominately planktic in habit. These usually form just a small part of the total valve count, but can be elevated at sites downstream of lakes and in slow-flowing rivers or canals where there are phytoplankton blooms. NGS equivalents of these metrics are included in DARLEQ3 and the following notes are provided to guide interpretation.

For percentage planktic valves, there is a poor relationship between LM and NGS outputs (Figure A.4.8a). Development of the barcode database has focused on assembling as many possible representatives of benthic flora and, as a result, barcodes of planktic taxa are largely derived from publicly available sequences. Mismatches between LM and NGS results probably arise for the following reasons.

- There are gaps in the barcode database, leading to over-representation in LM relative to NGS.
- Many planktic taxa have several chloroplasts per cell and so, when there is a good match with a sequence in the barcode database, relatively high representation in the NGS sample should be expected.
- As planktic taxa do not influence ecological status metrics, some analysts did not upload data for these taxa in the past – meaning that LM records may underestimate the true situation.

A high proportion of planktic taxa, whether in LM or NGS, should provoke the curiosity of anyone interpreting data. In most cases, there will be a simple explanation (that is, the sample came from a location close to a lake/reservoir outfall during the spring bloom period) and there is not always a clear distinction between ‘planktic’ and ‘benthic’ taxa (several *Aulacoseira* spp., for example, thrive in the loose epiphyton around macrophytes).

Do not over-interpret patterns in this metric: for those wanting to follow patterns in phytoplankton, there are better ways of doing this than analysing the benthos!

A.4.6.2 Percentage organic tolerant valves

There is a positive relationship between LM and NGS outputs for this metric, but with considerable scatter and a slight tendency for values computed using NGS data to be higher than those computed using LM data (Figure A.4.8b). This metric was included with the first version of the TDI to help users screen out sites where organic pollution effects were likely to confound any causal relationships between nutrients and diatoms. It has not been updated since 1995 and provides only an approximate indication of the scale of organic pollution.

A.4.6.3 Percentage motile valves

There is again a positive relationship between LM and NGS outputs with this metric, but with considerable scatter (Figure A.4.8c). This metric replaced % organic tolerant valves in the second version of the TDI, recognising that organically enriched sites could be identified by other means (water chemistry, invertebrates) and that interpretation of TDI outputs should focus on how biofilms differed in structure between sites and over time.

The % motile diatoms should not be used as an absolute measure of the condition of the biofilm; rather, it should be used to qualify interpretations of change. The emphasis should be on looking for consistent patterns of change (that is, ‘site B has consistently more motile valves than site A’). This should prompt questions on factors (hydrological, grazing, shade and so on) that might be responsible for this.

Do not make direct comparisons between % motile valves calculated on LM and NGS data.

A.4.6.4 Percentage saline valves

This metric was introduced into DARLEQ2 as a means of identifying sites with a brackish influence. Values computed on NGS data tend to be lower than those computed using LM data (Figure A.4.8d). This probably reflects gaps in the barcode database.

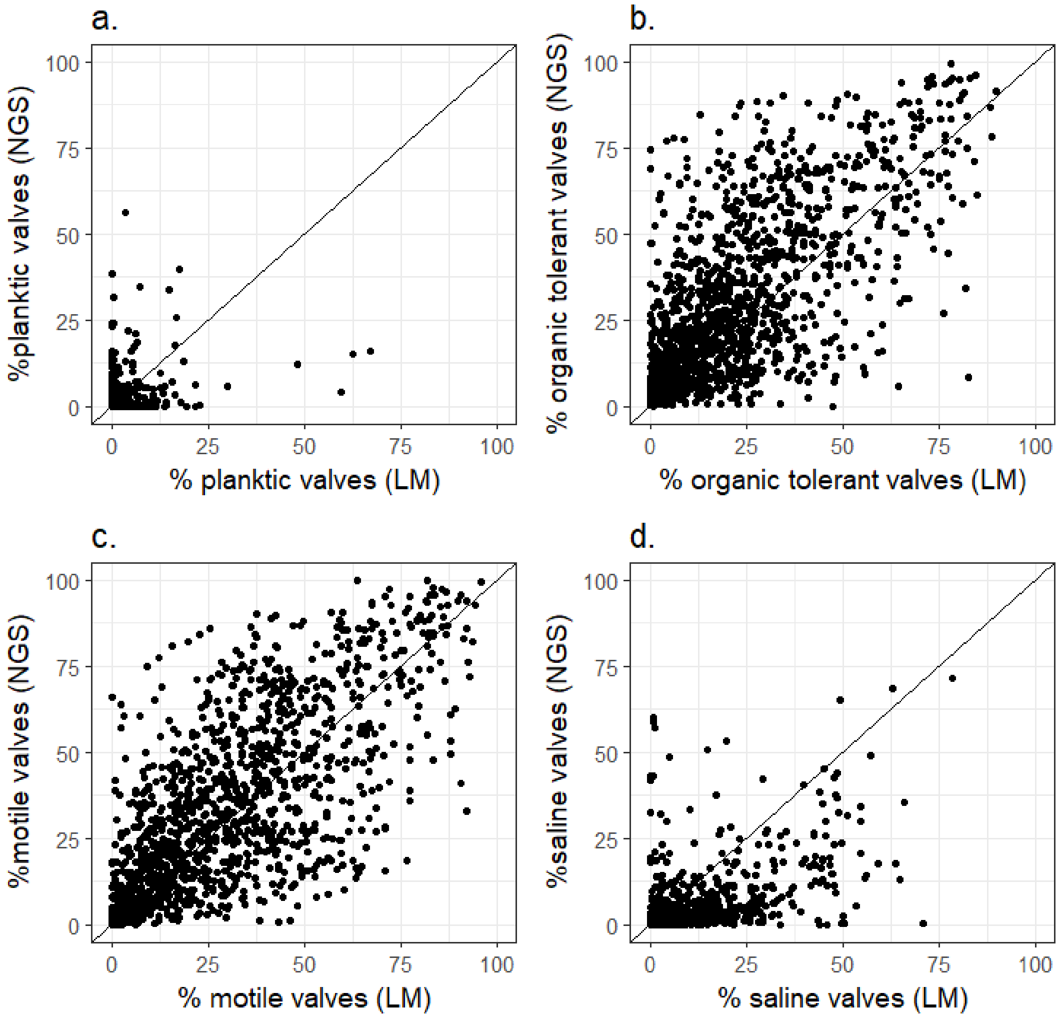

Figure A.4.8 Relationship between values of supporting metrics in LM and NGS outputs in the datasets used to derive TDI5LM and TDI5NGS: (a) %

planktic valves; (b) % organic pollution tolerant valves; (c) % motile valves; and (d) % saline valves

A.4.6.4 Hill's N2 diversity

This metric was not included in previous versions of DARLEQ. It has been included in DARLEQ3 to compensate for the loss of the ability to detect distorted valves when NGS data are used. Abundant numbers of distorted valves can be a sign that there are toxic pollutants present and they have been used in investigations into the effects of heavy metal pollution. Distorted valves encountered during routine surveillance monitoring have prompted checks on likely sources of contaminants within catchments or subcatchments.

Low biological diversity is another sign of toxic pollution and, for this reason, Hill's N2 diversity index has been included in DARLEQ3. Diversity will vary for many reasons within a site and occasional samples with low diversity is not a cause for concern; heavy grazing, for example, can result in a small number of fast-growing taxa thriving at the expense of others. Although there is little relationship between this metric computed with LM and NGS data (Figure A.4.9), a site that consistently returns TDI values <5 is worthy of investigation.

Measures of diversity based on the diatom assemblage alone should be interpreted with care, as diatoms are one part of a larger phytobenthos assemblage (potentially including representatives of several other algal phyla). As is the case for motile taxa, Hill's N2 diversity is not an absolute measure of the condition of the phytobenthos, but does offer useful supplementary information under some circumstances.

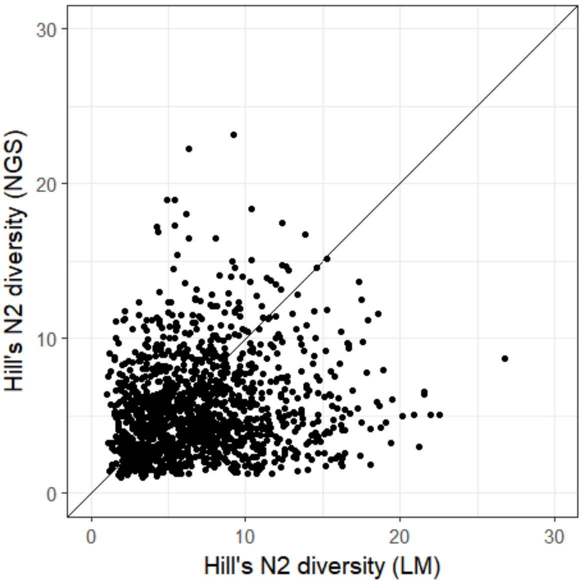

Figure A.4.9 Relationship between values of Hill's N2 diversity computed using LM and NGS data

A.4.6.5 Diatom Acidification Metric (DAM)

This was first included in DARLEQ2. It is not currently used for ecological status classification, although it has been used for investigations. It also provides useful supplementary information when interpreting data, particularly from low alkalinity sites. Great care should be taken when interpreting the TDI in situations where there may be anthropogenic acidification and it is recommended that:

- DAM is also calculated on all samples where alkalinity is <10mgL^{-1} $CaCO_3$
- inferences of trophic status are made only when acidification effects are absent or minimal (that is, when DAM indicates high or good ecological status)

DAM has not yet been tested using NGS data. However, there are likely to be a high rate of mismatches between LM and NGS due to the absence of a large number of important softwater/low pH indicators from the barcode database.

A.4.5.6 Lake Trophic Diatom Index (LTDI2)

A limited amount of testing of the NGS method has been carried out on littoral samples from lakes. These show reasonable agreement between values obtained by LM and NGS analysis (see Section 4 of this report). There are no plans at present to develop an NGS-specific metric but the LM metric does give reasonable results when computed using NGS data (albeit with a slight tendency to predict higher LTDI2 values).

DAM and LTDI2 can be computed for NGS data by following instructions for LM data. Users need to be fully aware of the issues outlined above before proceeding.

A.4.7 Uncertainty

DARLEQ3 includes the same uncertainty module as earlier versions and will calculate risk of misclassification and confidence of class for all sites included in the dataset.

These uncertainty calculations are not used by the Environment Agency or Natural Resources Wales, both of whom use the VISCOUS software package to account for spatial variation in water bodies during classification. The DARLEQ uncertainty module should only be used to support interpretation of LM and NGS data.

Uncertainty calculations for TDI5NGS are based on the same parameters as for LM based metrics. Although analytical uncertainty is lower for samples analysed by NGS, other sources of uncertainty are of a similar magnitude in both LM and NGS. This justifies the use of the LM uncertainty module in the short term. But as the DARLEQ uncertainty module is still used to underpin ecological status classifications in Scotland and Northern Ireland, it may need to be revisited and optimised before too long.

Would you like to find out more about us or 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 **(24 hours)**

Find out about call charges (https://www.gov.uk/call-charges)

Environment first

Are you viewing this onscreen? 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.