Interpreting TDI5NGS data

Version 1.2 20/12/2018
Martyn Kelly

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.

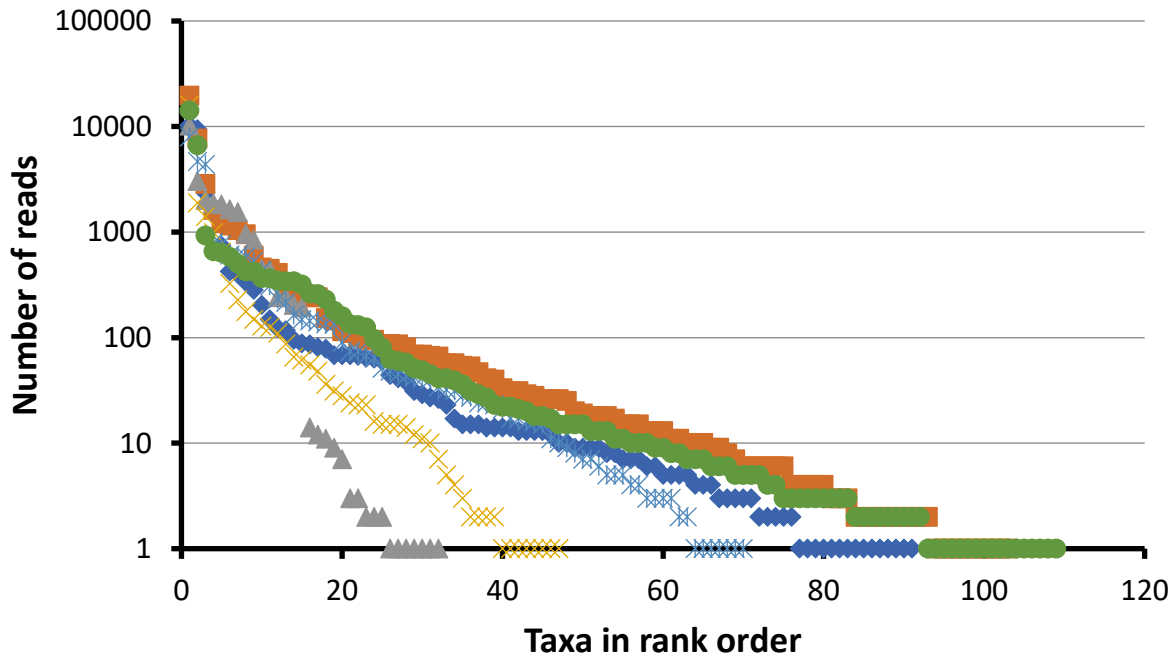

Figure A.4.1 Species abundance curves for 6 NGS samples, selected at random, to illustrate the properties of NGS data

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.

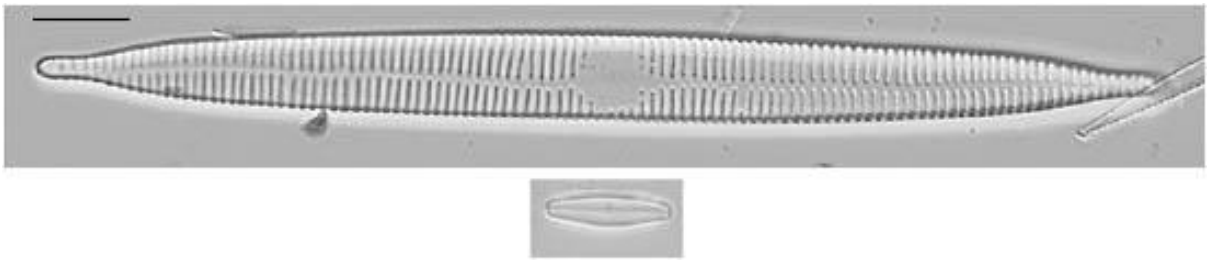

Figure A.4.2 Specimens of *Ulnaria ulna* (top) and *Achnanthidium minutissimum* (bottom)

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).

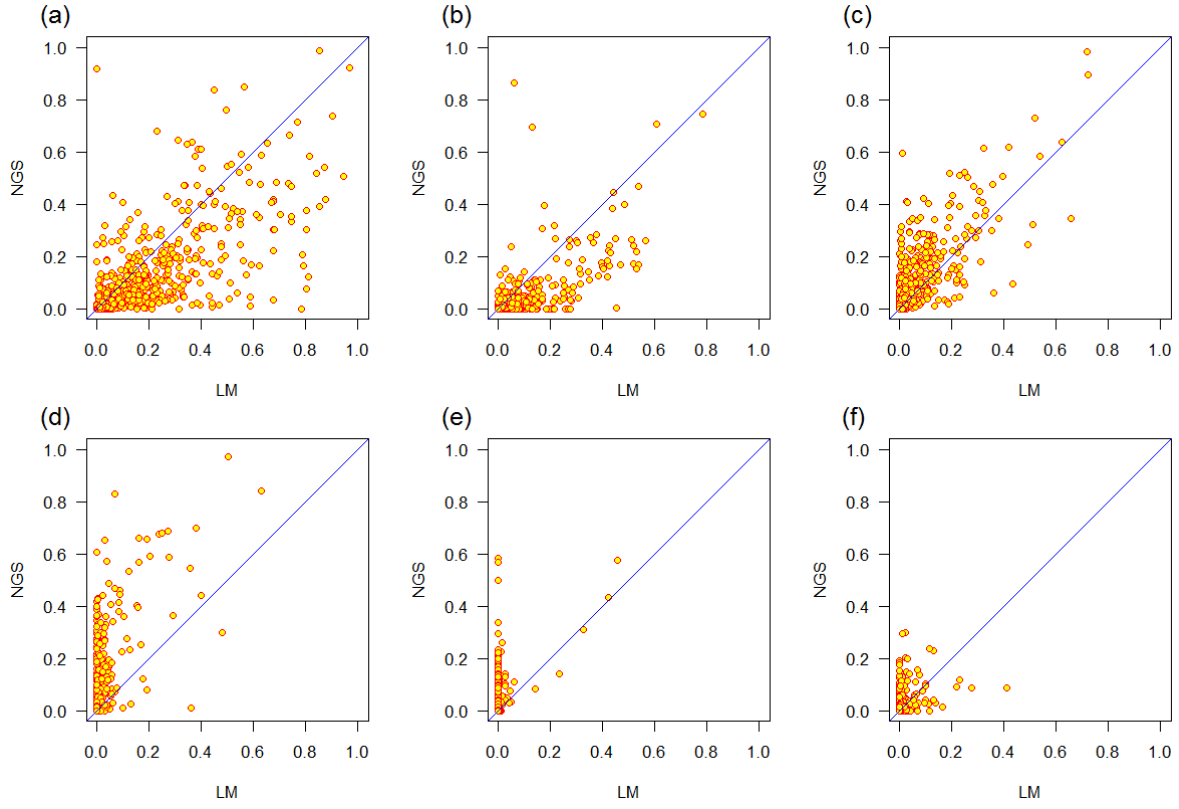

Figure A.4.3 Differences between representation of common taxa in LM (x axis) and NGS (y axis) on a proportional scale: (a) *Achnanthidium minutissimum* type (small, one chloroplast); (b) *Amphora pediculus* (small, one chloroplast); (c) *Navicula lanceolata* (medium sized, 2 chloroplasts); (d) *Melosira 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.

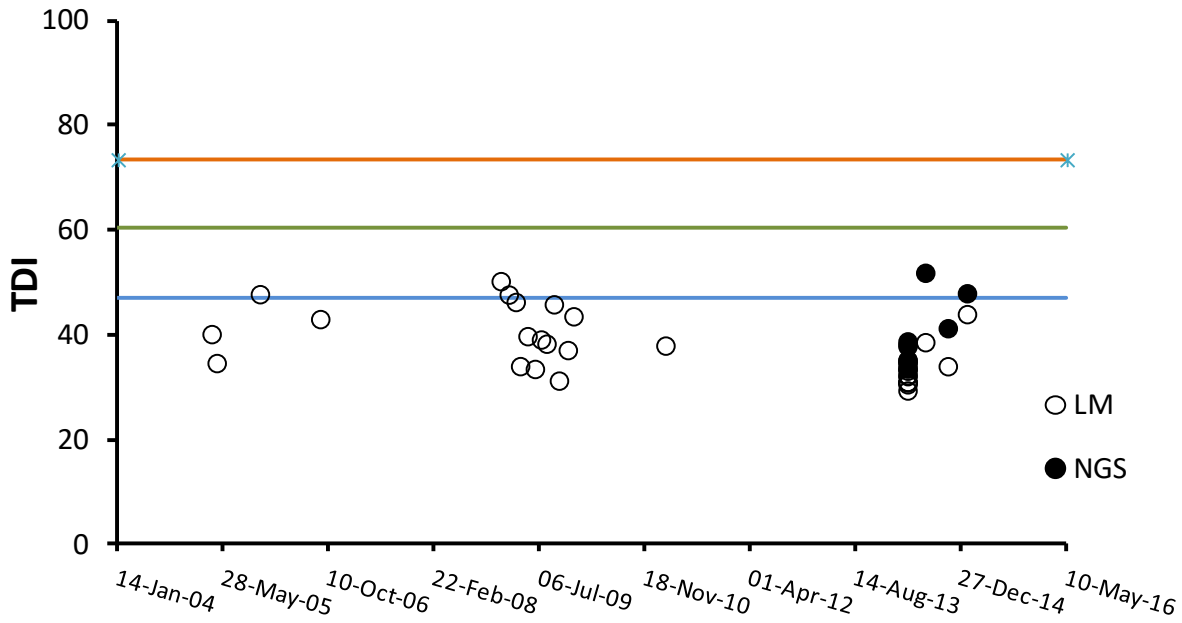

Figure A.4.4 Long-term trends in TDI scores in the River Wear at Wolsingham

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.

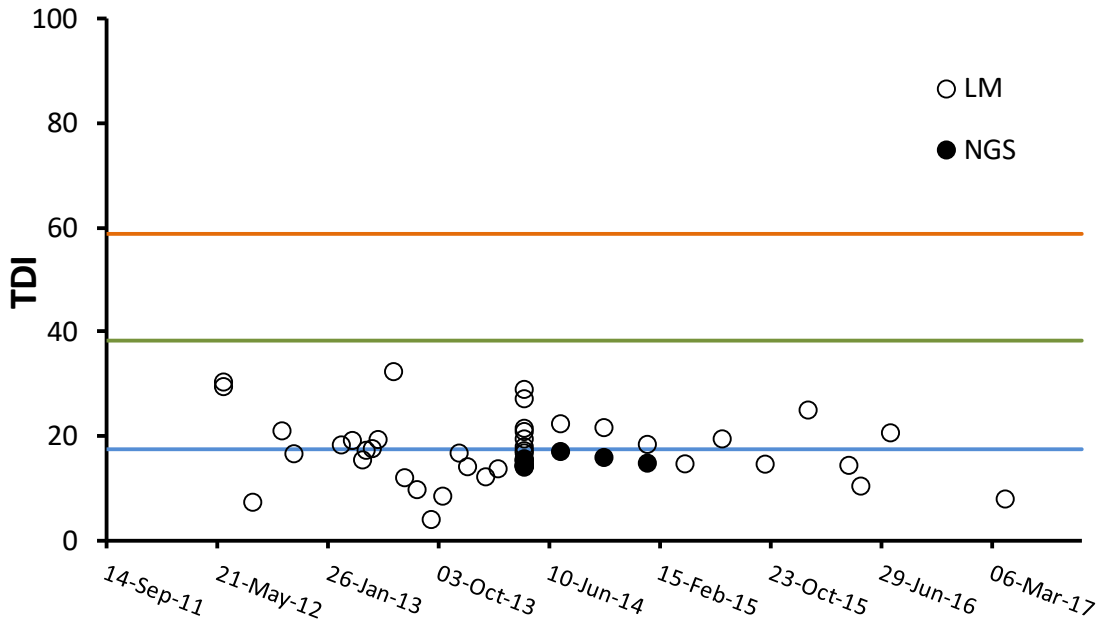

Figure A.4.5 Long-term trends in TDI scores in the River Ehen near Ennerdale Bridge

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).

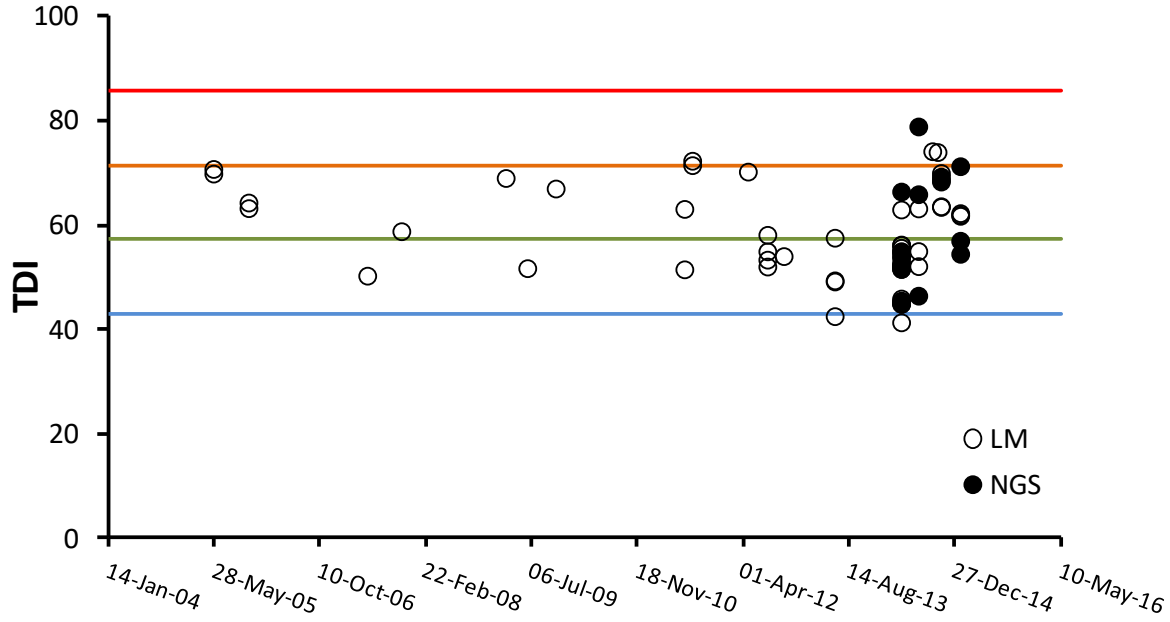

Figure A.4.6 Long-term trend in TDI scores in the River Derwent at Ebchester

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.

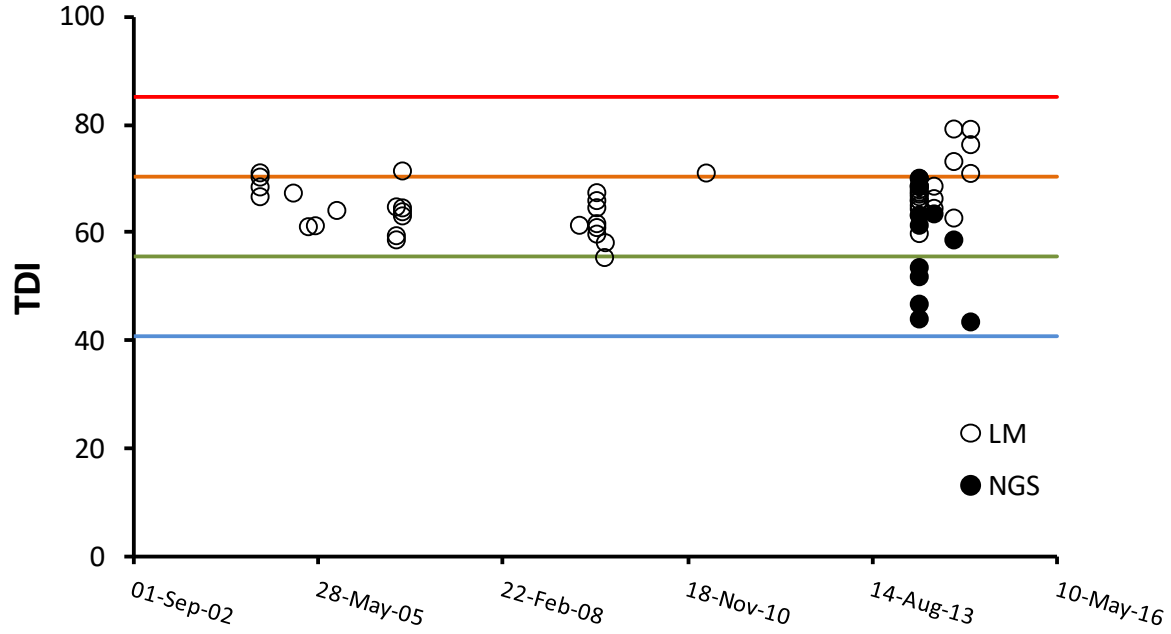

Figure A.4.7 Long-term trend in TDI scores in the River Team at Causey Arch

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.

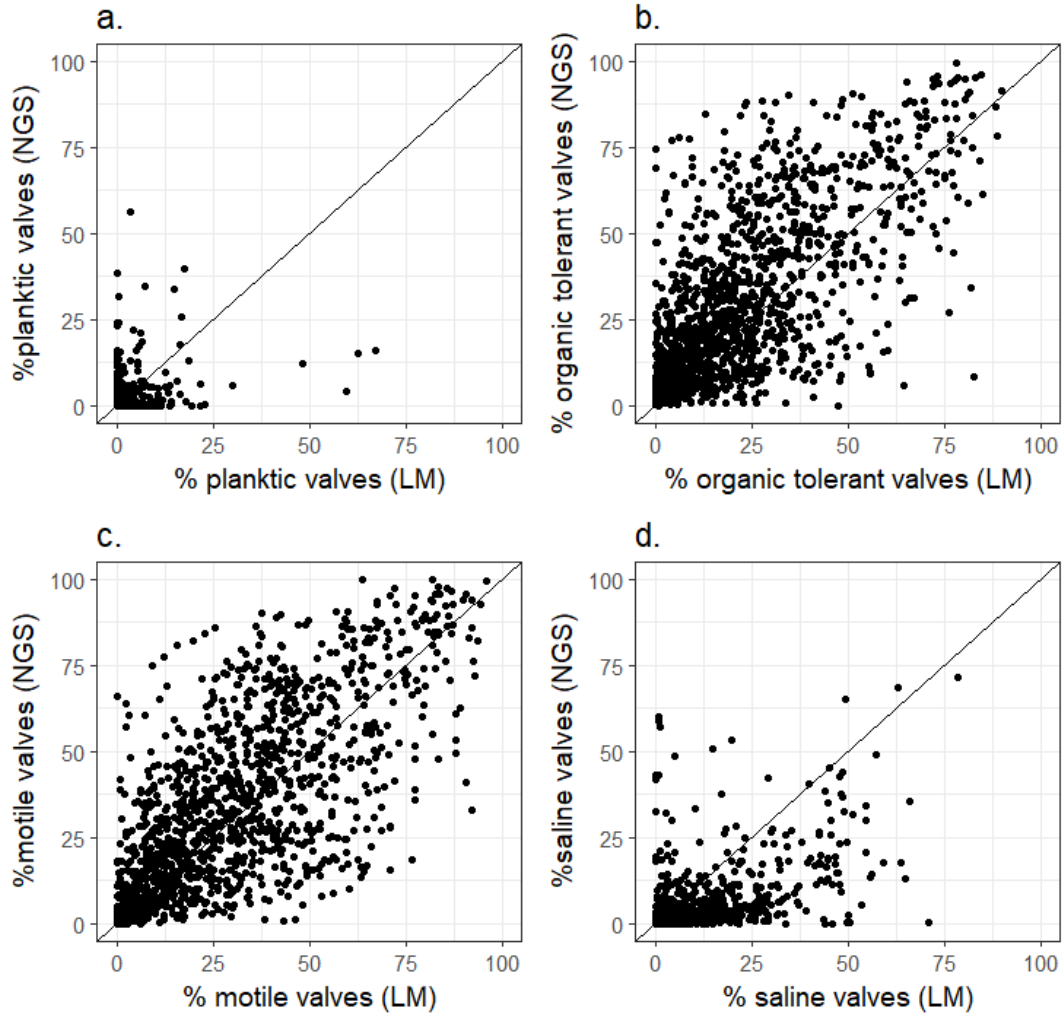

Figure A.4.8 Relationship between values of supporting metrics in LM and NGS outputs in the datasets used to derive TDI5LM and TDI5NGS: (a) % planktic valves; (b) % organic pollution tolerant valves; (c) % motile valves; and (d) % saline valves

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.

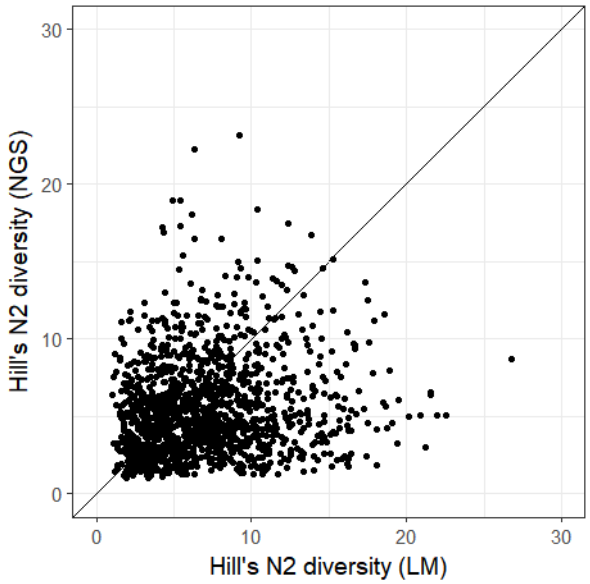

Figure A.4.9 Relationship between values of Hill's N2 diversity computed using LM and NGS data

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.