Interpreting TDI5NGS data

Version 1.0   31/07/2017

Martyn Kelly

A journey of 1000 miles starts with a single step ...

Lao Tsu (c. 604-531 BC)

Introduction

DARLEQ3 offers the capability to perform ecological assessments using data generated by either light microscopy (LM) or Next Generation Sequencing (NGS). However, the two methods will not necessarily give identical results when applied to the same sample so DARLEQ3 users need to understand how NGS data differs from LM data, and what this means for interpreting ecological status.

If you are approaching NGS data for the first time, it is useful to bear in mind the limitations of current methods, based on light microscopy (Table 1). LM-based analysis is not perfect, but it is a method that we 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”. We build up a clearer view of ecological status by collecting information from a range of different biological, chemical and physical components of water body over time. NGS analysis simply offers us a different way of generating information about the status of the phytobenthos. Some aspects of the NGS method might offer a clearer view; however, there will also be information that can be gleaned from microscopic analysis that cannot (yet) be duplicated with NGS. In the short term, however, we need to understand that NGS data are **different** to LM data. These differences do not mean that it is **wrong**, just that we need to learn to interpret these new data and, perhaps, to leave behind some of the preconceptions that we brought along when interpreting LM data.

The first three bullet points in Table 1 apply to assessment of phytobenthos status using NGS as well as to the LM-based method. Whilst 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). Finally, we know that DNA can survive outside the cell for some time, 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, so NGS results are likely to be a more direct insight into which species were alive at the time of sampling.

Table 1. Limitations of ecological status assessment using diatoms analysed by light microscopy.

- Does not capture all phytobenthos diversity
- Assessments based on lists of species, no consideration of functional properties or productivity
- Limited quantification (relative not absolute abundance)
- No consideration of cell size
- Cannot differentiate live from dead cells

Sample size

Fig. 1 shows the number of reads per species for six NGS samples selected at random from the dataset from which DARLEQ3 was developed. Species are listed in rank order, with the most abundant on the left. This illustrates three important differences between data generated by NGS and LM:

1. 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 but it would take an extraordinary length of time and, in practice, most analysts name and count between 300 and 500 valves.

2. 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 and will also be related to the bioinformatics pathways that are used (i.e. how stringent are the filters that match reads to species in the barcode library. The size of the barcode library will also be a factor contributing to the number of species that are recorded.

3. 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 one percent 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. We also cannot be sure that taxa represented by a small number of reads represents a viable population living at the site at the time the sample was collected. It is possible that sample includes "eDNA" – molecules that are suspended in the river water or tangled in the biofilm but which derive from populations elsewhere in the catchment. Similarly, we cannot be sure that very rare diatoms detected by LM represented 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 barcode library at present and the proportion should decrease as the barcode library increases in size.

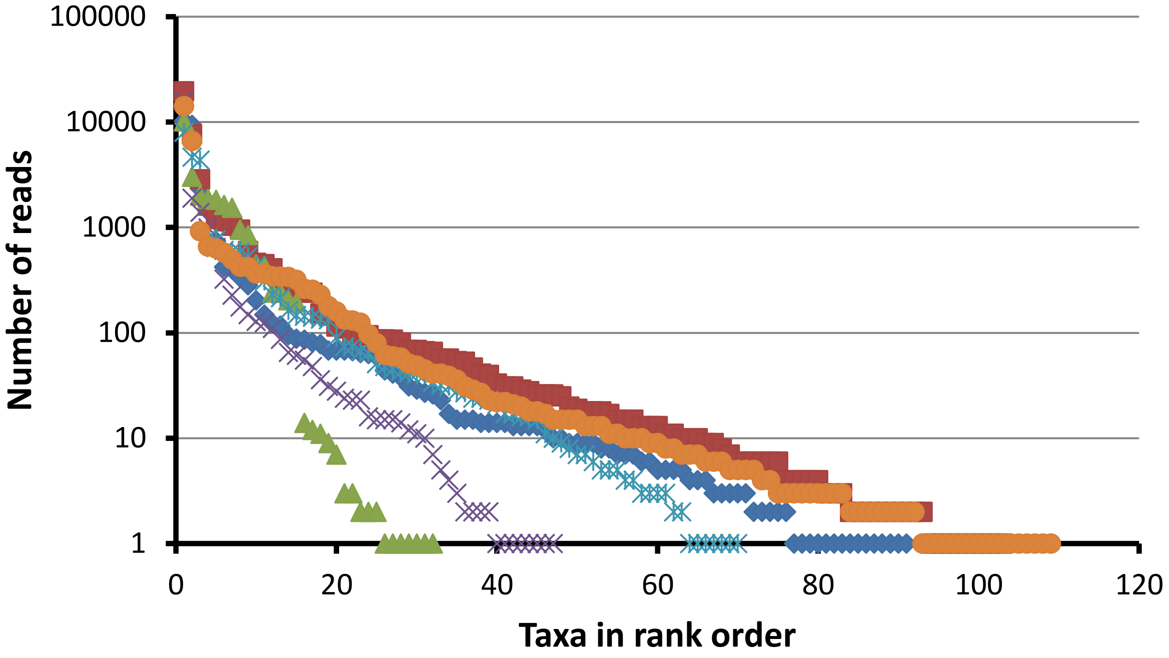

Figure 1: Species abundance curves for six NGS samples, selected at random, to illustrate the properties of NGS data.

Expression of individual species

The standard unit of enumeration in LM analyses in the UK and several other countries is the valve (i.e. half the cell wall, or frustules). However, diatoms can vary considerably in size, both within the cell cycle and between species. Figure 2 shows one of the larger diatoms that is common in UK waters alongside one of the smaller diatoms. The difference in cell biovolume is 100 times, and we can assume that the larger cell contributes substantially more to primary productivity in a sample than the smaller. However, each make the same contribution to the LM analysis.

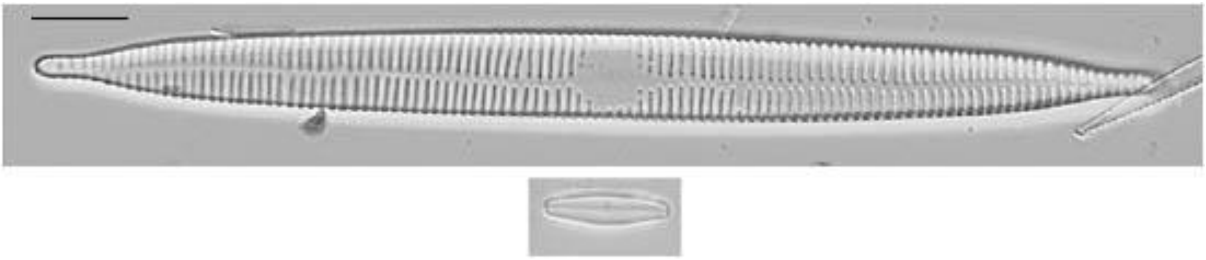

Figure 2. Specimens of *Ulnaria ulna* (top) and *Achnanthidium minutissimum* (bottom). Both are from cultures used for obtaining sequences for the barcode library. Scale bar: 10 µm. Photographs: Shinya Sato.

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 should, in theory, give us a better insight into the contribution each species makes to primary productivity than simply counting cell numbers.

In practice, however, there is still a lot that we do not understand about the expression of rbcL in diatoms, and about 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, and also that cells with many chloroplasts have more rbcL than cells with single chloroplasts. It is also possible that chloroplast shape influences the number of reads, and that read number can vary depending on environmental conditions and through the cell cycle.

Table 2: Variation in chloroplast numbers between major groups of diatoms

Group	**Number of chloroplasts**
Centric diatoms	Mostly many per cell
Araphid diatoms	Many genera have one or two per cell (e.g. *Fragilaria, Hannaea*); a few have many per cell (*Tabellaria, Fragilariforma, Asterionella*)
Raphid diatoms	Most have one or two per cell; a few have four (*Neidium, Fistulifera*)

Figure 3 shows how the expression of six common species differs between LM and NGS. Figures 3a and 3b show *Achnanthidium minutissimum* and *Amphora pediculus*, two small pioneer species each with a single chloroplast and both of these tend to form a greater part of the LM than the NGS analysis. By contrast*, Navicula lanceolata* (Figure 3c) is a larger diatom with two chloroplasts and the proportion recorded in NGS tends to be greater than in LM. *Melosira varians* (Fig. 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* (Fig. 3e) is a very small, weakly silicified diatom with four 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* (Fig. 3f).

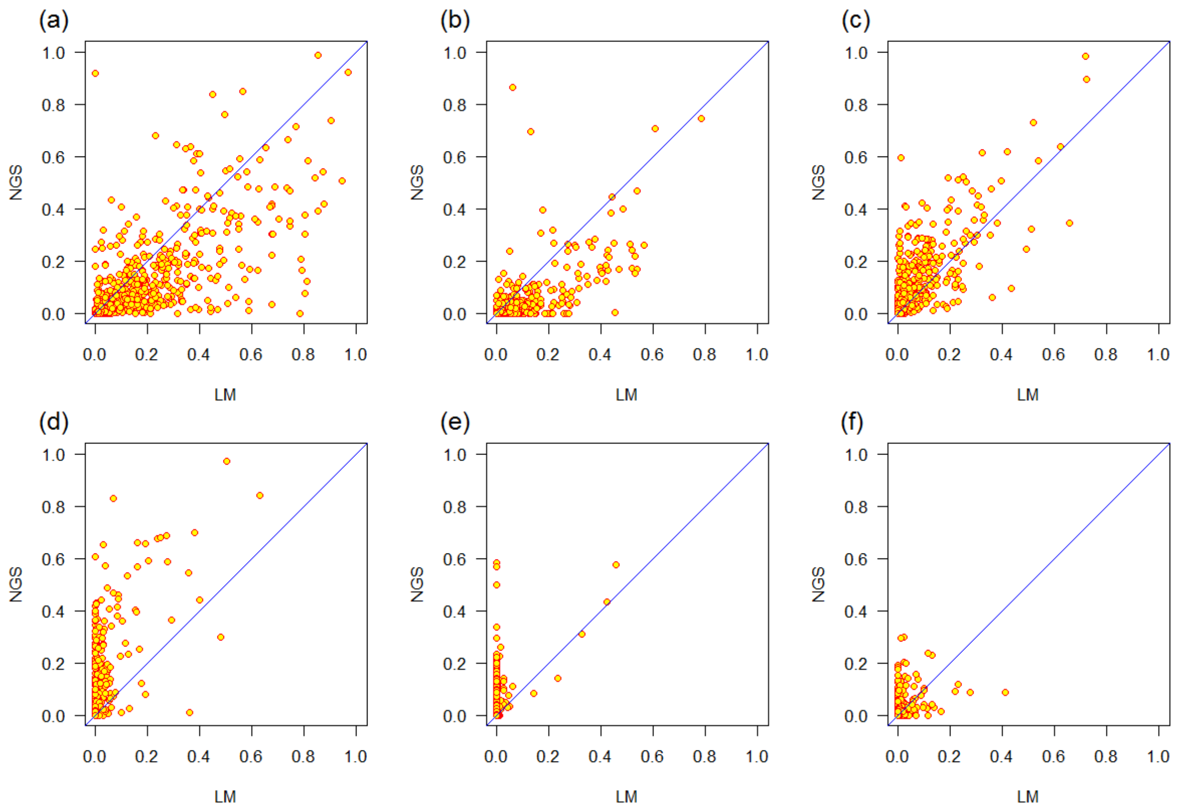

Figure 3. Differences between representation of common taxa in LM (x axis) and NGS (y axis) on a proportional scale: a) *Achnanthidium minutissimum*-type (small, one chloroplast); b) *Amphora pediculus* (small, one chloroplast); c) *Navicula lanceolata* (medium sized, two chloroplasts); d) *Melosira varians* (large, many chloroplasts); e) *Fistulifera saprophila* (very small, four chloroplasts, weakly silicified); f) *Mayamaea atomus* (including var. *permitis* (very small, possibly two chloroplasts, weakly silicified). The diagonal line shows slope = 1 (i.e. equal representation in LM and NGS). Source: SC140024.

Interpreting TDI5NGS

We are still learning how to interpret NGS outputs. Problems will be particularly acute in the period following the transition from LM to NGS as you will have to reconcile results produced with NGS with older data collected using LM. This is discussed more in the next section. The following pointers should help you understand your NGS output:

- **Cell size and chloroplast number** play a role in determining the representation of a taxon in NGS output. Do not overinterpret 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

- **Know your catchment**. This applies to all data interpretation, not just to diatoms analysed by NGS. In the case of NGS data, however, we need to be aware that the sample may contain eDNA from upstream sources, and also that planktonic taxa may behave differently in NGS compared to 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.

- **Gaps in the barcode library:** about 2800 diatom species have been recorded from Britain and Ireland but only 350 are represented in the barcode library. Many of these are only represented by a few strains, so we cannot 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
    - Table3 lists taxa that are abundant in LM analyses but which are not, as yet, represented in the barcode library.
    - Table 4 lists taxa that are abundant in LM analyses but which have < 5 strains in the barcode library.

Table 3 ...[use taxa with max RA >=5% in LM]

Table 4 ...[use taxa with max RA >=5% in LM]

- **Individual species may behave differently** in NGS compared to LM
- **There are occasional “misfires” of both methods.**
    - For LM analyses, most analysts participated in a ring test scheme; however, we know of 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 analyses. 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. If necessary, samples are re-run. This will catch most instances of rogue samples; however, treat samples with low numbers of reads (< 3000) with caution.
    - In over 80% of cases, the difference between LM and NGS analyses will be < 10 TDI units. However, exceptions do occur (see next section for an example) and you should take care if a TDI value computed with NGS data is very different (e.g. > 1 status class) from what you expected.
- **Limitations of current reference model:** both DARLEQ2 and DARLEQ3 use a reference model that is not very effective in hard water. You should not use these models in water where alkalinity is > 120 mg L^{-1} $CaCO_3$. TDI4 and TDI5 may be useful in investigations in harder water, but should be interpreted with care.

Table 5 compares LM and NGS results for one sample, in order to illustrate 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 library? Similarly, *Cyclotella meneghiniana* should, in theory (medium sized cell with many chloroplasts) have been more abundant in NGS than LM. Other differences, however, do match expectations, and the overall difference in TDI is within the expected range.

Table 5: Comparison results from light microscopy (LM) and Next Generation Sequencing (NGS) from River Browney, Co. Durham, B6301 bridge, August 2014. Only species present at >5% in at least one analysis are presented.

Species	**LM**	**NGS**	
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 strains 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

The effect of changing to NGS analyses on long-term trends in TDI

A very reasonable question to ask prior to adoption of a NGS-based diatom method is if the change from LM will affect the classifications of water bodies. This question can only be answered where there is data showing a long-term trend based on light microscopy plus sufficient NGS data to permit a comparison. Project SC140024 generated NGS data over space and time for four water bodies in northern England for which long-term LM data were also available. These four rivers are considered in order of decreasing ecological status.

River Wear, Wolsingham, Co. Durham

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

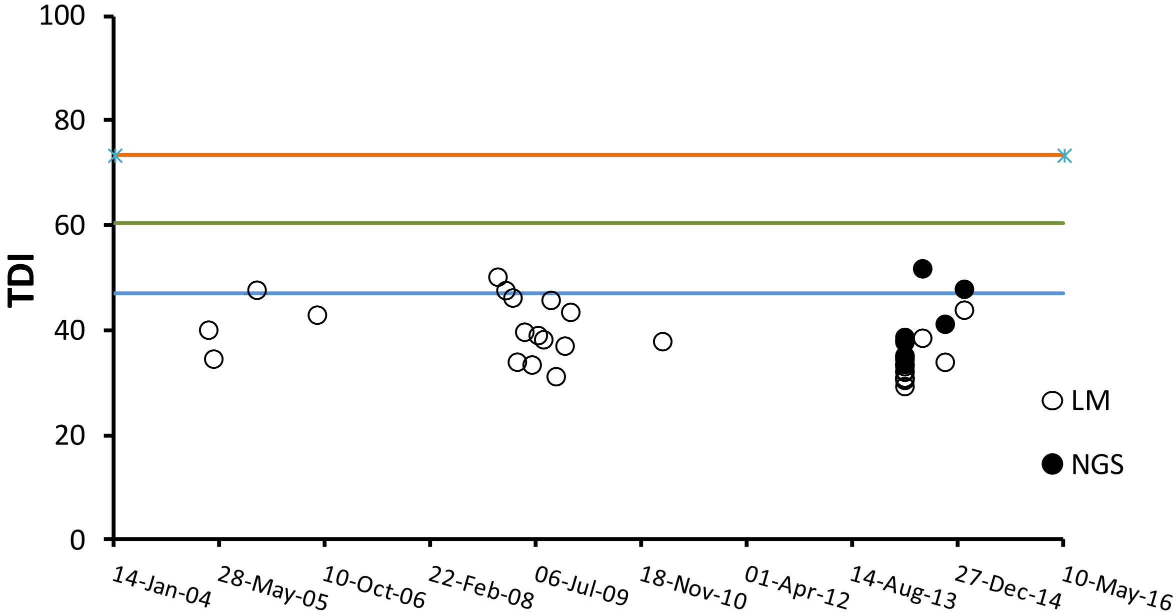

Figure 4: Long-term trends in TDI at the River Wear, Wolsingham. Horizontal lines show the position of high/good (blue), good/moderate (green) and moderate/poor (orange) status class boundaries.

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. The alkalinity at this site is much lower, so the 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 and the relatively consistent results for NGS may reflect some gaps in the barcode library rather than suggesting that the method is more reproducible than LM here.

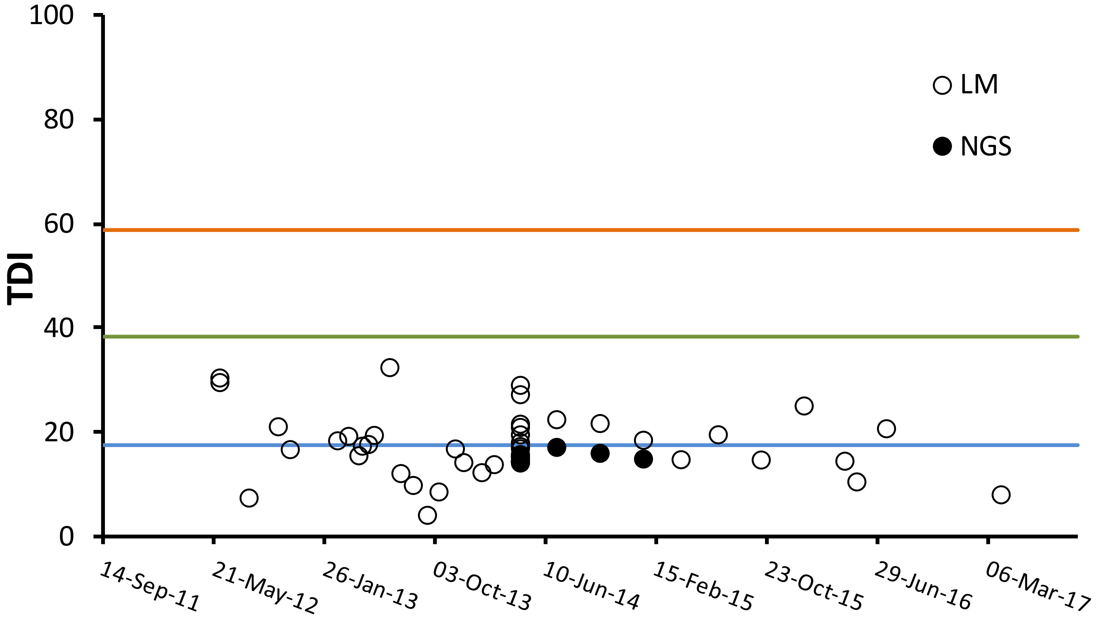

Figure 5: Long-term trends in TDI at the River Ehen near Ennerdale Bridge. Horizontal lines show the position of high/good (blue), good/moderate (green) and moderate/poor (orange) status class boundaries.

River Derwent, Ebchester, Co. 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 status, with occasional results in poor status.

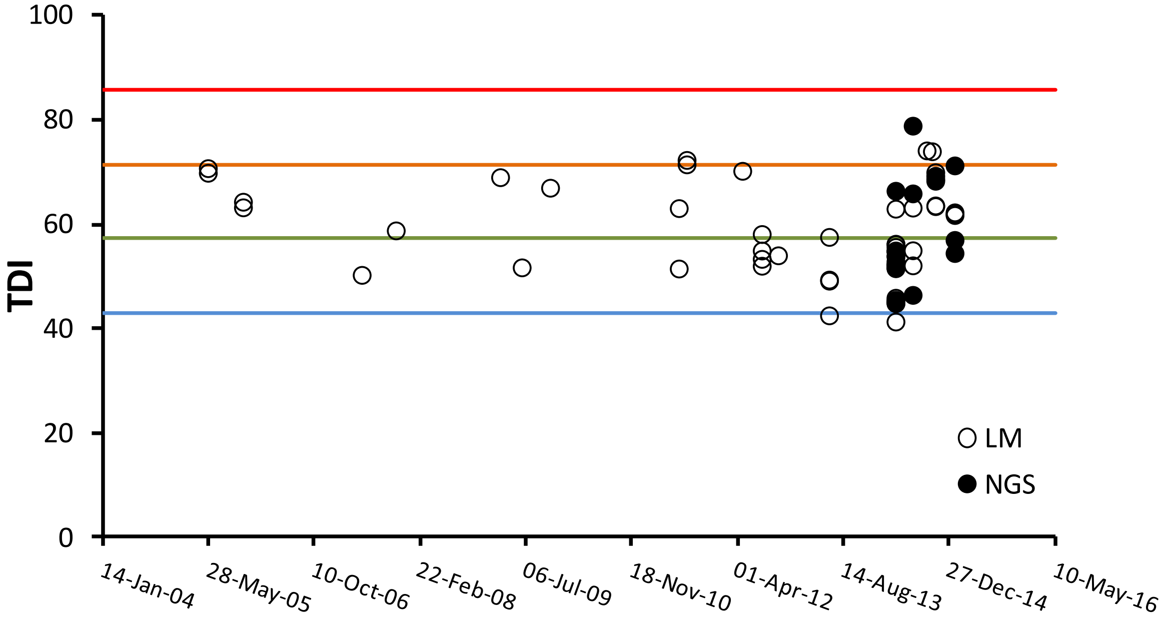

Figure 6: Long term trend in TDI in the River Derwent at Ebchester Boundaries as above, with the addition of poor/bad (red).

River Team, Causey Arch, Co. 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 status, with some falling to poor status. Most NGS samples follow this trend, but there were also a few outliers, for reasons that cannot be fully explained (see SC140024). In this instance, getting classifications of good status from a river where all a priori evidence points to less than good status should sound alarm bells.

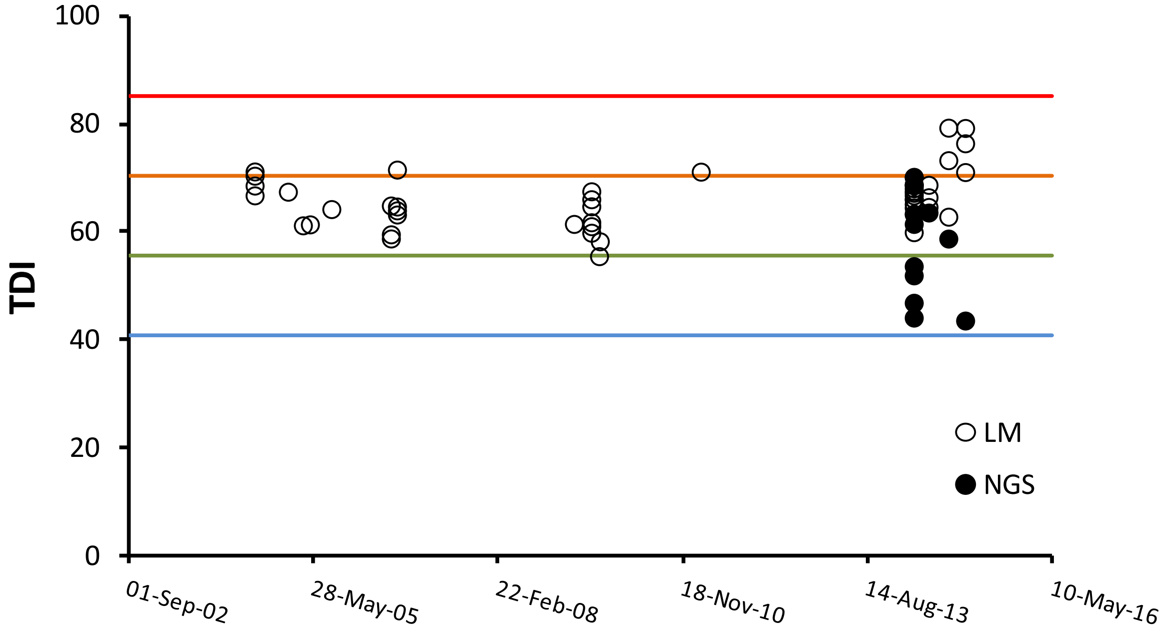

Figure 7. Long term trend in TDI in the River Team at Causey Arch. Boundaries as above.

Other metrics