Multi-marker metabarcoding resolves subtle variations in freshwater condition: Bioindicators, ecological traits, and trophic interactions

Chloe Victoria Robinson[1,2], Teresita M. Porter[1,3], Victoria Carley Maitland[1], Michael T.G. Wright[1], Mehdrad Hajibabaei[1]*

[1]Centre for Biodiversity Genomics & Department of Integrative Biology, University of Guelph, Guelph, ON Canada

[2]Whales Initiative, Ocean Wise Conservation Association, 21-21 Dallas Road, Victoria, BC Canada

[3]Great Lakes Forestry Centre, Natural Resources Canada, 1219 Queen Street East, Sault Ste. Marie, ON Canada

* Corresponding author

Email: mhajibab@uoguelph.ca

Abstract

Freshwater systems are experiencing rapid biodiversity losses resulting from high rates of habitat degradation. Ecological condition is typically determined through identifying either macroinvertebrate or diatom bioindicator assemblages and comparing them to their known tolerance to stressors. These comparisons are typically conducted at family or genus levels depending on the availability of taxonomic keys and expertise for focal groups. The objective of this study was to test whether a more taxonomically comprehensive assessment of communities in benthic samples can provide a different perspective of ecological conditions. DNA metabarcoding was used to identify macroinvertebrates and diatoms from kick-net samples collected from sites with different habitat status. Sites with ‘good’ condition were associated with higher beta diversity as well as slightly higher directed connectance and modularity indicating higher resilience compared with ‘fair’ condition sites. Indicator value and correlation analyses used DNA metabarcoding data to detect 29 site condition indicator species consistent with known bioindicators and expected relative tolerances. DNA metabarcoding and trophic network analysis also recovered 11 keystone taxa. This study demonstrates the importance of taxonomic breadth across trophic levels for generating biotic data to study ecosystem status, with the potential to scale-up ecological assessments of freshwater condition, trophic stability, and resilience.

Key words: Biomonitoring, DNA metabarcoding, diatom, macroinvertebrate, benthos, bioindicators, water quality, food webs, COI, rbcL

Introduction

We are currently experiencing rapid freshwater biodiversity declines on a global scale, as anthropogenic pressures including habitat destruction, water pollution, overexploitation and climate change continue to escalate. The process of slowing future freshwater biodiversity losses is complicated, due to the influence of surrounding land use, particularly upstream human activities, on environmental conditions within rivers, lakes, wetlands and ponds [1,2]. Protecting and restoring aquatic ecosystems across spatial and temporal scales requires a multi-faceted approach, inclusive of ecological network analyses (e.g. trophic interactions) [1,3–5]. Before we can take the actions necessary to conserve freshwater ecosystems, we need to assess freshwater condition through biomonitoring [4–6] and understand system stability and robustness to biodiversity loss and environmental stressors [7–9]. Reproducible and scalable approaches for monitoring freshwater systems have never been more in demand than they are today.

Freshwater biomonitoring methods have evolved alongside the intensifying biodiversity declines, as demands grow for faster generation of mass data production (i.e. “big data”) [10–12]. Typically, benthic macroinvertebrates are targeted for conducting freshwater health assessments, due to their taxonomic diversity, localized habitat occupancy and taxa-specific responses to a range of environmental gradients [6,12–15]. Across North America, reference sites are used for evaluations across watersheds, to account for variability in macroinvertebrate assemblages across ecoregions [13,16]. Region-specific tolerance values can then be generated via the Hilsenhoff Biotic Index (HBI), which provides a single tolerance value based on the average benthic arthropod community tolerance values to organic pollution (0 for very intolerant to 10 for highly tolerant) [14,17,18].

More recently, freshwater riverine microalgae, also referred to as diatoms, are also being used as bioindicators of rivers and streams, because of their strong response to environmental changes [19–21]. Although microscopic morphological identification is currently the method of choice for diatom biomonitoring, high-throughput DNA metabarcoding of environmental samples has facilitated scaling up, primarily because of the ability of this method to bypass time-consuming morphology-based identifications [22–25]. The combination of newly optimized and species-inclusive sample collection techniques (i.e. benthic kick-net [25]) and reduced time taken to identify taxa [22] highlights the applicability of DNA metabarcoding as the 'catch all' approach for understanding freshwater condition. However, the ecological value of this 'catch all' method has not been investigated in real-world biomonitoring analyses.

DNA metabarcoding overcomes biomonitoring bottlenecks and enhances the amount of species-level diversity detected from environmental samples [10,11,26–28]. The field is now in a position to move beyond simple biodiversity inventory measures such as richness, beta diversity, and community composition to associate species detections with known biological or ecological traits [29–31]. One way to integrate and visualize this data, is through network analysis, a systems-level approach useful for integrating many layers of data. For example, metabarcoding data can be used to identify known trophic interactions, these interactions can then be used to build directed networks, food webs, to examine trophic relationships, identify keystone species, and clusters of potentially interacting species that can then be associated with their ecological traits such as tolerance to water pollution [9,12,32–36]. Although trophic analysis is widely utilized in, for example, pollinator-plant, predator-prey systems; and network analysis has been widely utilized in the analysis of microbiome data [37] these approaches, are under-utilized in biomonitoring, particularly in river systems [9]. Using species interactions to build trophic networks- can facilitate freshwater health assessments by visualizing the overall structural and functional relationships within a system [38,39]. More general network properties corresponding with ecosystem resilience and stability, such as connectedness and modularity, may also function like an early warning system for system collapse [32,40–44].

Considering the role of multiple taxonomic groups (i.e. macroinvertebrates and diatoms) simultaneously as bioindicators can broaden the impact of freshwater assessments [45,46]. Multi-taxa approaches provide a more holistic representation of freshwater ecosystem health through the combination of taxonomic diversity, different species' environmental tolerances and network properties (i.e. connectedness, modularity) from more than one traditional bioindicator kingdom [45–48]. This ultimately enables the detection of keystone taxa that have a very large effect on their environment without which the community would be very different or not exist [49]. Despite the evidence that multi-taxa biomonitoring approaches should be adopted, this is rarely the case due to logistical, time and cost restrictions involved with collecting representative samples for each taxonomic group [46,50]. There remains a lack of integration between DNA-based sample collection techniques for biomonitoring of macroinvertebrates and diatoms [25]. The lengthy and multi-step nature of field sampling techniques for multiple taxa, can be overall detrimental to the amount of freshwater data collection [50] and is particularly incompatible with community-based monitoring (CBM), which is fast becoming a driving force for freshwater health data generation [51].

The objective of this study is to leverage trans-kingdom metabarcoding data generated from the same benthic kick-net samples to identify species associated with site condition in relation to known site condition and bioindicator taxa. Specifically, we: 1) use the cytochrome c oxidase subunit I (COI) (macroinvertebrate mitochondrial DNA marker) and ribulose bisphosphate large subunit (rbcL) (diatom chloroplast DNA marker) for the metabarcoding of benthic kick-net samples to generate biodiversity metrics (richness, effective number of exact sequence variants (ESVs), beta diversity) to assess subtly varying site condition (fair/good), 2) use multi-marker metabarcodes to identify site condition bioindicators for comparison with known stress tolerance, and 3) conduct an exploratory analysis of known trophic interactions to further assess the structure and stability of trophic networks across site conditions. We expect to determine unique bioindicators and keystone taxa, in addition to the well-known groups of bioindicator taxa because metabarcoding results are expected to both reflect and complement traditional sampling methods. We also predicted that ‘good’ quality sites would be more complex networks reflecting their ability to support a diverse array of taxa and functions.

Results

A total of 3.2 million COI and 3.9 million rbcL sequence reads were generated for this study (Supplementary Tables 2 and 3). Following bioinformatic processing of raw reads, removing rare clusters, noise, chimeras, and pseudogenes a total of 4,026 COI and 1,573 rbcL ESVs (1,304,473 and 574,866 reads, respectively) were retained. Rarefaction curves indicate that the sequencing depth was sufficient to capture the ESV diversity for both diatoms and macroinvertebrates across all four sites (Supplementary Fig. 1). After the COI and rbcL datasets were rarefied and normalized to the 15th percentile of library sizes and merged, 45,937 reads in 2,933 ESVs were retained for further ESV level analyses.

Diversity Analyses

At the order level, 'fair' sites show higher richness than 'good' sites for both diatoms and macroinvertebrates (Supplementary Fig. 3 & Supplementary Fig. 4). Diatoms from 12-17 orders were detected from samples from 'fair' sites and 9-16 orders from 'good' sites. Orders Naviculales, Cymbellales, Fragiliariales and Thalassiosirales were most prevalent across both 'good' and 'fair' sites. Genera within Thalassiosirales, Thalassiophysales and Bacillariales are known tolerant taxa [52–54] and had higher read abundance in 'fair' versus 'good' sites (Supplementary Fig. 3). We detected macroinvertebrates from 5 phyla: Platyhelminthes (flat worms), Nematoda (roundworms), Mollusca (molluscs), Arthropoda, and Annelida. Macroinvertebrates from 30-51 orders were detected from samples from 'fair' sites and 12-29 orders from 'good' sites. Traditional indicators of poorer water quality in river systems [16], including

Haplotaxida, Gastropoda and Diptera and Odonata had higher read abundance in ‘fair’ sites. Although we appreciate that read abundance does not necessarily reflect organismal abundance in the environment, due to known issues with primer-bias and differential recovery of taxa with different body sizes, the relative abundance of these taxa across site conditions does correlate with what we would expect based on known species tolerances to pollution [55,56].

The higher richness of ESVs compared to effective number of ESVs shows that the diversity across both site conditions is driven largely by many rare ESVs. The diversity detected in ‘fair’ sites was about twice as high in ‘good’ sites when measured using richness (macroinvertebrates: t-test, p.adj = 0.0039; diatoms: t-test, p.adj = 0.0240) but the effective number of ESVs were not found to be significantly different among site conditions (macroinvertebrates: t-test, p.adj = 0.92; diatoms: t-test. p.adj = 0.39; Fig. 1). Within each site condition diversity was similar between macroinvertebrates and diatoms (t-test, p.adj > 0.05; Fig. 1).

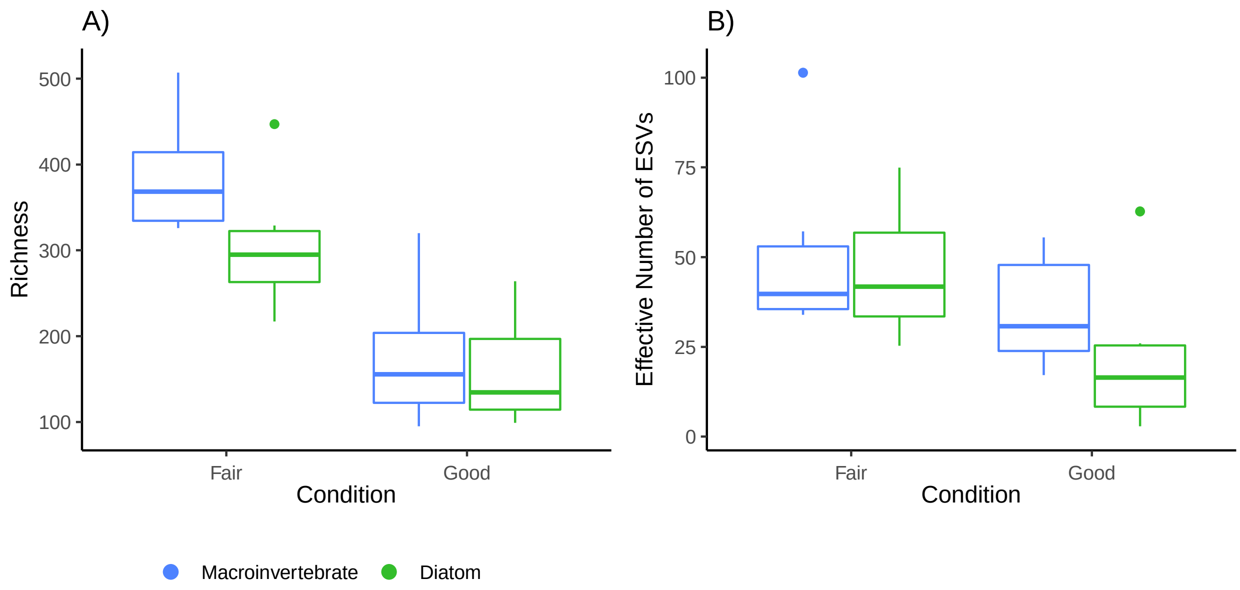

Figure 1. Observed richness is driven by a large number of rare ESVs. A) ESV richness and B) The effective number of ESVs is shown for each condition and taxonomic group.

NMDS plots based on binary Bray-Curtis dissimilarities of ESVs across sites show good separation among fair and good sites using either diatoms or macroinvertebrates (diatoms stress = 0.04, linear R^2 = 0.99; macroinvertebrates stress = 0.06, linear R^2 = 0.98; Fig. 2). Diatoms and macroinvertebrates are both correlated with dissolved oxygen, turbidity, and pressure (mmHg). Additionally, macroinvertebrates are also correlated with pH and temperature. PERmutational ANalysis Of Variance (PERMANOVA) for diatoms showed that habitat status (good or fair) explained 22% of the variation in beta diversity (p-value = 0.001), whereas sampling site explains 25% of the variation (p-value = 0.004; Supplementary Table 4). For macroinvertebrates, habitat status explained 19% of the variation in beta diversity (p-value = 0.001), whereas site explained 35% of the variation (p-value = 0.002;

Supplementary Table 4). The PERMANOVA reflects a combination of both dispersion (diatom site and status; macroinvertebrate site) and location effects as shown in the ordinations. Within each site condition (fair or good), dissimilarities were lower in fair sites and higher in good sites for diatoms but had a similar overlapping distribution for macroinvertebrates (Supplementary Fig. 5).

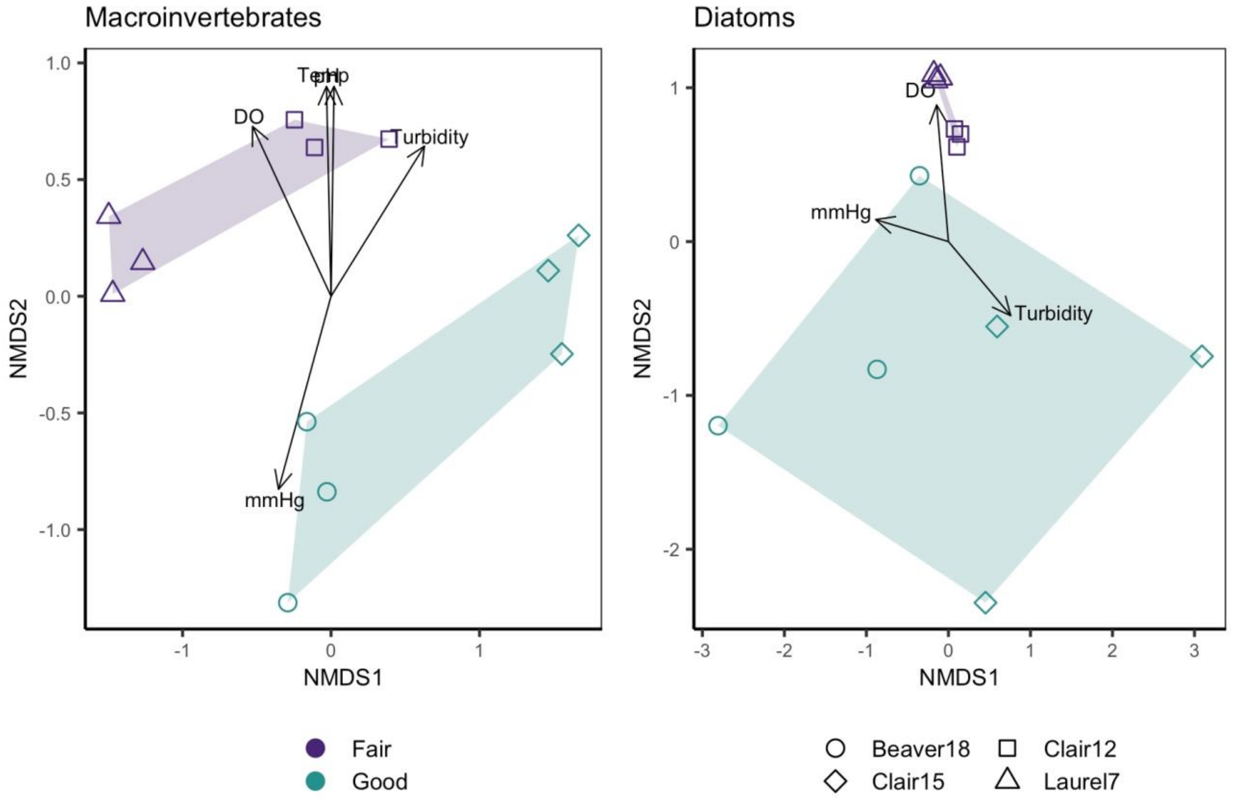

Figure 2. Beta diversity was greater within 'good' sites, especially for diatoms. Beta diversity within 'fair' sites, tended to be lower, especially for diatoms. Sample replicates cluster by site, and sites cluster by site status. Environmental variables that correlate with beta diversity patterns are shown if they have a p-value < 0.05. Based on binary Bray Curtis dissimilarities from libraries where read counts were rarefied to the

15th percentile. Abbreviations: dissolved oxygen (DO); pressure (mmHg), temperature (Temp).

Bioindicators

Results of indicator species analyses based on a rarefied read count matrix detected 29 site condition indicator species (Table 1). We recovered 28 fair condition indicators and one good condition indicator. The site condition indicators detected using the indicator value method (IndVal) and the point biserial correlation coefficient (*r*) were largely similar. The main differences between these statistics lie in their interpretation (Supplementary Fig. 6). The A and B components of the IndVal method indicate the predictive value and sensitivity/fidelity of the indicator [57]. In this study, most of our site condition indicators have very high predictive value (close to 1) but only moderate sensitivity/fidelity. The point biserial correlation coefficient represents the ecological preference of species for a particular site condition and can range from -1 to +1, reflecting negative to positive correlations, and the closer to the absolute value of 1, the stronger the correlation [58]. The advantage of including this measure, is that this method can detect both positive and negative correlations. This method recovered many of the same species as the indicator value method, all positively correlated with site condition. We also populated Table 1 with Biological Condition Gradient (BCG) scores from the Diatoms of North America Database (NADED; https://diatoms.org/) [25] and HBI scores for macroinvertebrates [16].

Trophic interactions

An exploratory analysis using food webs for each site based on the automated retrieval of resource-consumer interactions was conducted (Fig. 3). GloBI annotation of resource to consumer interactions was possible for 71% (548/777) of our target taxa at the species and genus ranks. After filtering out interactions with off-target taxa (E.g., bacteria, fungi, plants, vertebrates), common names and insufficiently identified taxa (E.g., Chironomid, Lumbriculiid, Oligochaeta), and taxonomically unidentified substrates (E.g., CPOM - coarse particulate organic matter, detritus) target taxa representation was reduced to 34% (266/777). After filtering out off-target interactions, 22% (171/777) of our original target taxa were left represented in our interaction list. For each site, this means that 25.8 - 32.3% of the original target taxa were represented in each network. These trophic networks represent the current state of interaction annotations between diatoms and macroinvertebrates in GloBi and were used to visualize the trophic structure within each site and measure the network properties that would allow us to learn more about the stability of each site. Food webs generated from ‘fair’ habitat status sites tend to have more nodes (taxa), links (resource to consumer interactions), greater trophic height (longer food chains), and more clusters (Table 2). Food webs generated from ‘good’ habitat status sites, however, had slightly higher directed connectance (links/species^2) and modularity (strength of divisions of a network into clusters). Similar to other described small-world type networks, our networks are highly clustered with relatively short path lengths [59]. This means that most of the nodes in the network are not connected to each other, but the ones that are connected likely have neighbours that are also connected to each other, i.e., form clusters. Clair15 classified as a ‘good’ site had the smallest and sparsest food web, with lower numbers of nodes and trophic links. The Clair12 site classified as ‘fair’ had the greatest number of trophic links and trophic height. Fair condition sites tended to have more macroinvertebrate predators that are also tolerant to organic wastes (Table 3).

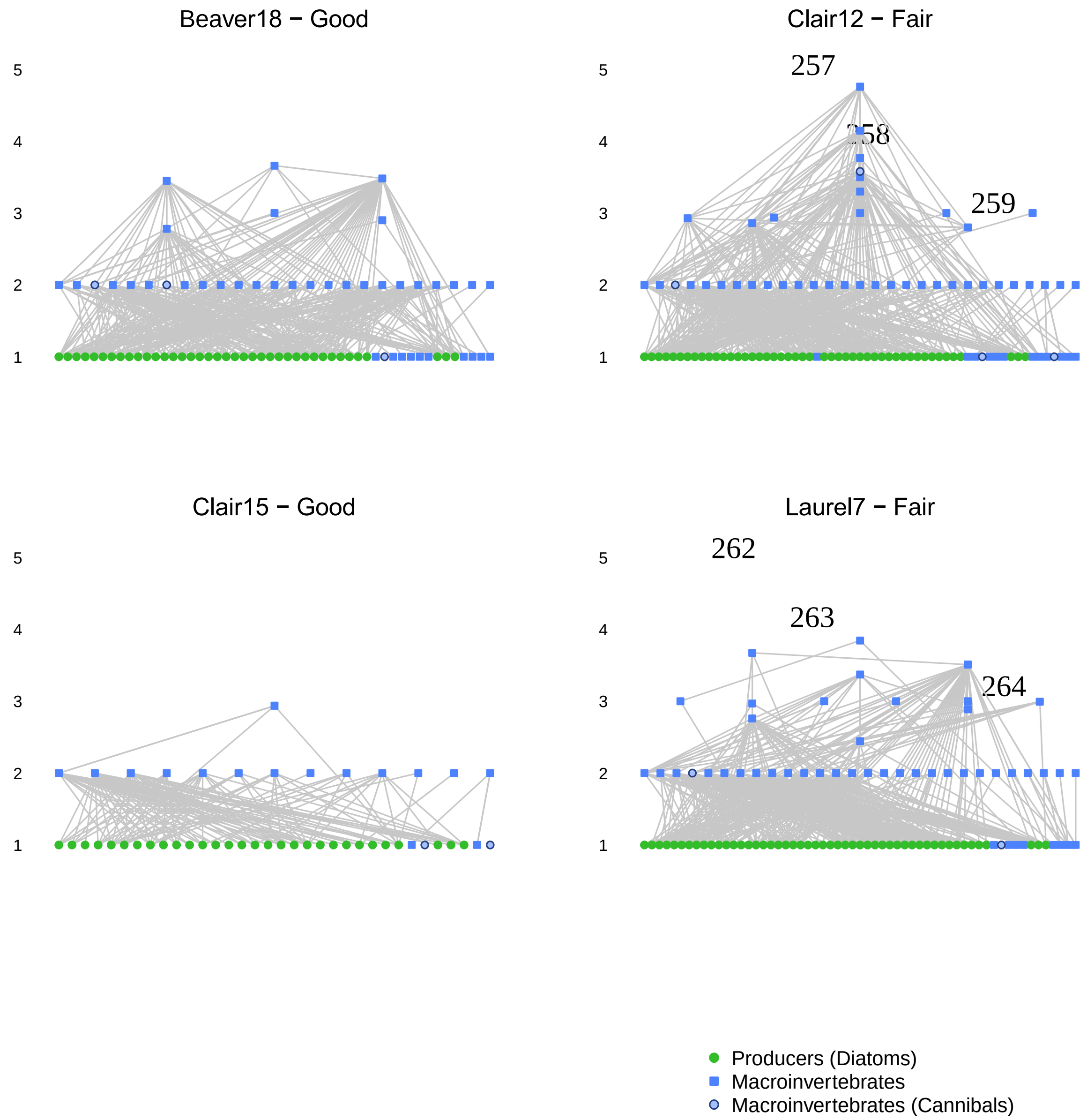

Figure 3. Food webs from 'fair' sites had higher trophic height and more macroinvertebrate predators. Vertical food webs with the lowest trophic level, mainly producers (diatoms), at the bottom. Estimated trophic position for each node in the food web was determined using chain averaged trophic level (trophic height).

Directed graphs of resource-consumer interactions were also generated for each site (Fig. 4). This was done to identify clusters of potentially interacting taxa and to identify potential keystone taxa. Vertices (taxa) belonging to a cluster are encircled by a black line and the taxonomic composition of these clusters are detailed in Supplementary Tables 6-9. 4-5 clusters with more than one taxon were identified in 'good' sites and 7 clusters were identified from each 'fair' site (Table 2). Macroinvertebrate and diatom keystone genera for both 'good' and 'fair' sites were identified using two centrality measures: degree and hub scores and the top 3 scoring taxa from each site are summarized in Table 4. Degree and hub scores for all taxa are shown in Supplementary Tables 5-8. Generally, the distribution of degree and hub scores did not differ among sites assessed as 'good' or 'fair' (Supplementary Tables 5-8). While there were a few outlier invertebrates with a particularly high degree with many links to other taxa, diatoms in general tended to have higher hub scores than macroinvertebrates (Fig. 5). This reflects the large number of links from diatoms to a variety of macroinvertebrate consumers that themselves tend to feed on a variety of diatom species. Using the network terminology described by Kleinburg [60], this makes diatoms ‘hubs’ and diatom-consumers ‘authorities’. Diatoms play a key ecological role in linking macroinvertebrates together in trophic networks and this is reflected by the Kleinburg hub scores. Overall network modularity was assessed as low (0.1-0.15) to medium (0.15-0.2), being slightly higher for good sites (Table 2). In this study, we detected a greater number of clusters (containing more than one taxon) from networks with lower modularity. Though we detected more clusters, the strength of overall graph modularity was relatively weak, i.e. differential density of links within and between clusters not as stark.

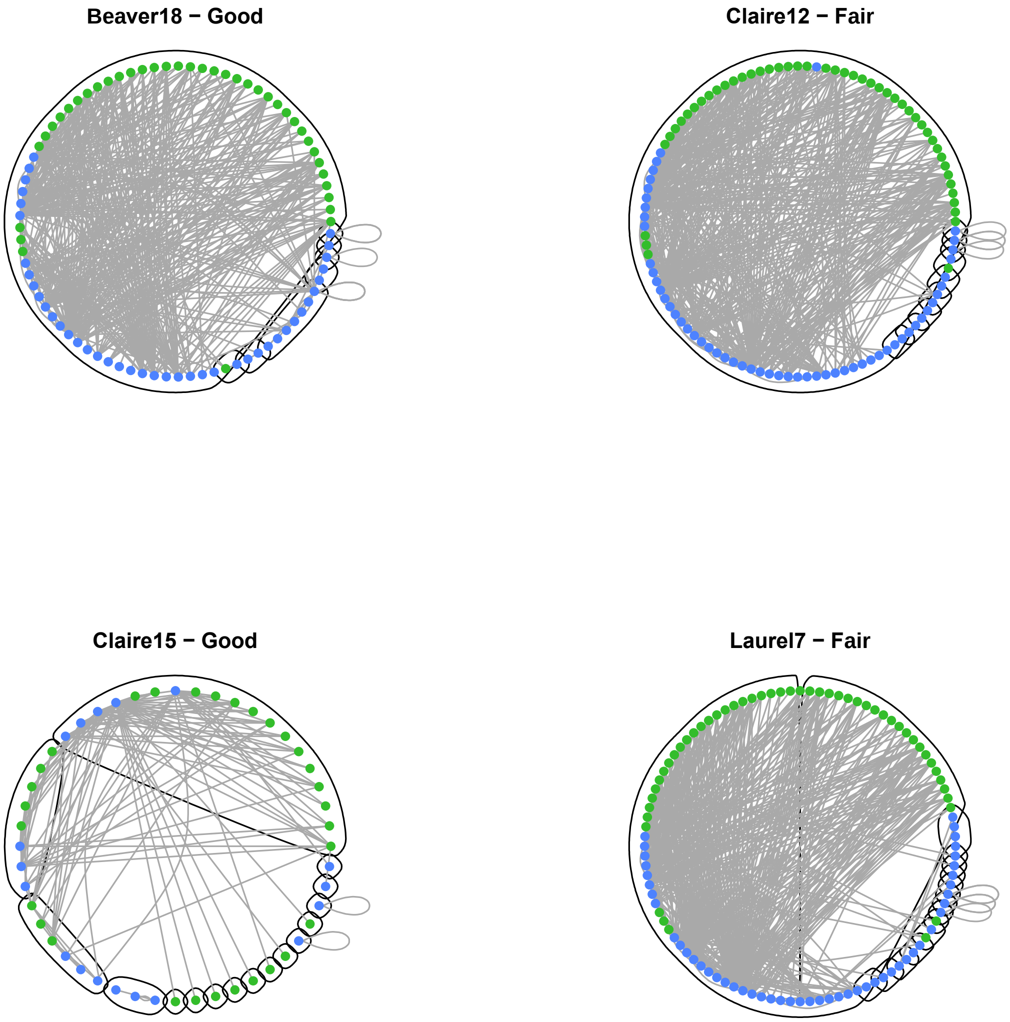

Figure 4. Freshwater benthic food webs show low to medium levels of modularity. Though these are directed graphs with interactions that point from resource to consumer, arrow heads were removed from the plot to improve readability. Trophic links (edges) are shown in grey. The small grey loops indicate cannibals, taxa known to consume members of the same taxon. Nodes (taxa) were arranged in a circle and colored according to taxonomy (blue-macroinvertebrates; green-diatoms). Clusters of taxa, as well as isolated taxa, are circled in black.

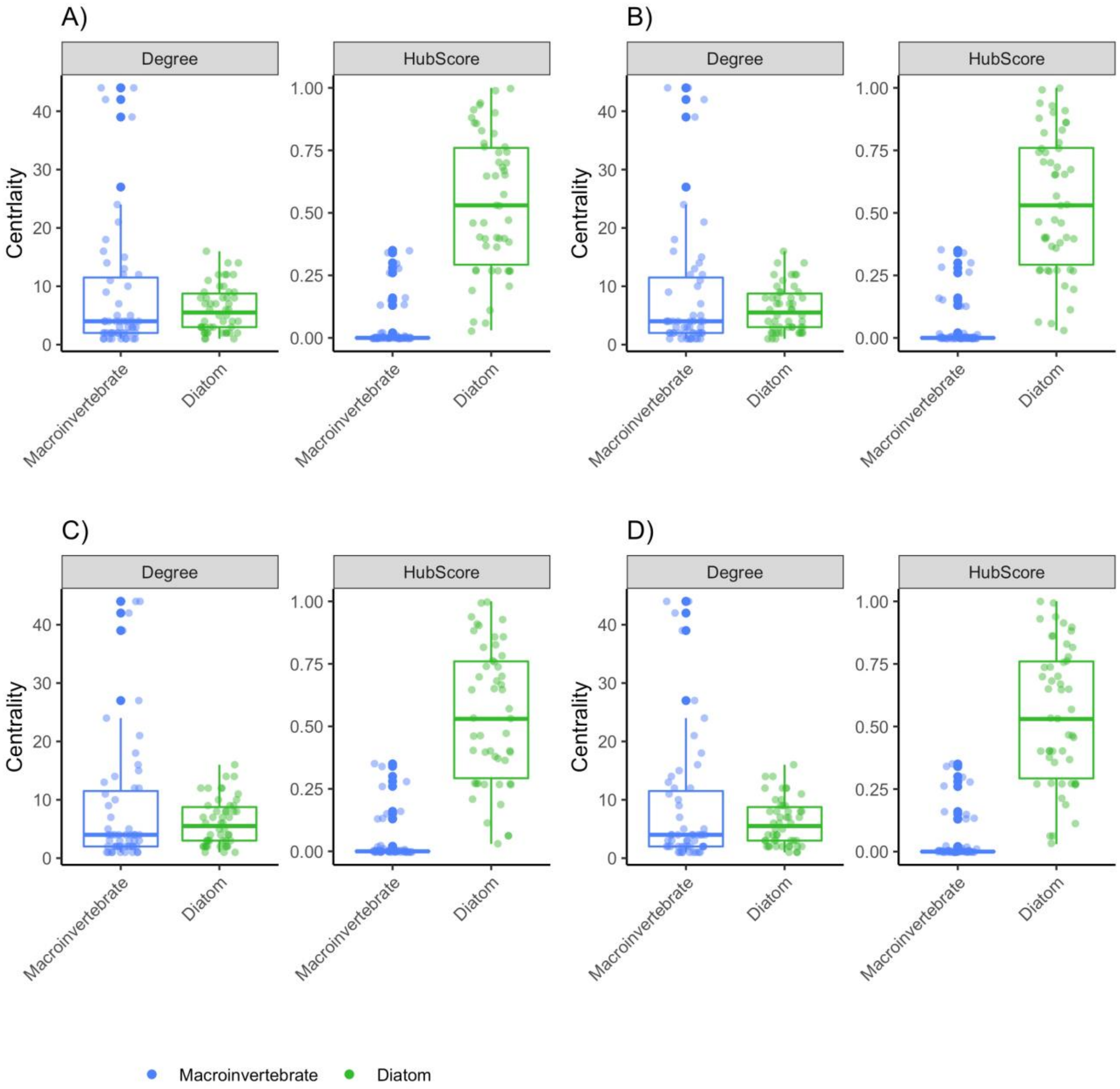

Figure 5. Kleinburg hub scores reflect the importance of diatoms in linking macroinvertebrate consumers together in trophic networks. Two measures used to detect potential keystone taxa are shown: degree centrality and Kleinberg's hub centrality scores. Degree shows the number of connections into and out of each taxon node. Hub scores reflect the number of connections out of each taxon node (hubs), from resource to consumer, weighted more heavily when the consumers are

themselves linked to by many other hubs. Sites are shown as follows: A) Beaver18 (good), B) Clair15 (good), C) Clair12 (fair), and D) Laurel7 (fair).

Discussion

This study presents a trans-kingdom assessment of the parameters associated with a subtle variation in habitat status (fair/good). Multi-marker metabarcoding data detected both macroinvertebrates and diatoms from kick-net samples. We used this data to examine the biodiversity and network parameters associated with fair and good sites.

Macroinvertebrate bioindicator taxa are targeted globally as a means of identifying water quality status through assessment of assemblages in relation to environmental metadata [13]. However, it can be difficult to understand aquatic health in relation to macroinvertebrate assemblages when evaluating ‘test’ sites which do not have corresponding reference sites and only have low-resolution regional tolerance values (e.g. to family and not species level) [48]. This can be especially challenging for DNA metabarcoding-derived bioindicator species data, which requires tandem morphological-based studies for abundance assessments to be made[33].

In our study, we utilized an integrated approach to rapidly identify site-specific bioindicator species. Through a combination of DNA-derived taxonomic assignments and indicator value/correlation analyses, we were able to determine site condition bioindicators without needing to identify species via morphology, or limit analyses to only EPT taxa. Biological Control Gradient (BCG) scores for the 15 diatom bioindicator species identified were between a 3 and 5, with the *Gomphonema* species being the only exception (BCG score of 2). A score between 3 and 5 is representative of an impaired system and reflects the point where diatom assemblages change due to increased human activity [61]. *Gomphonema* as a genus which tends to be located within unimpaired systems [62], however, as this information is based on the ecoregion of California as opposed to eastern Canada [63], this may indicate that *Gomphonema* are more tolerant to poorer water quality in southern Ontario. Similarly, the Hilsenhoff Biotic Index (HBI) scores for identified 'fair' macroinvertebrate bioindicators ranged from 6 to 10, which falls within the 'Fairly Poor' to 'Very Poor' water quality categories [16]. The one 'good' macroinvertebrate bioindicator species identified scored a HBI value of 4.0, whose presence in a system translates to 'Very Good' water quality status [16]. Considering that the HBI was developed to detect organic pollution in aquatic systems through species weighting via relative abundance [16,17], our site condition indicator species have been determined using rarefied read counts from metabarcoding data, without the need to quantify species abundance, and yet is still reflective of this index.

Despite the subtle habitat quality difference between the two site types, we have shown that it is still possible to identify site condition indicator species, especially for diatom taxa, which are lacking BCG metrics for Canadian systems [61]. The habitat quality class used to assign our sites as 'fair' and 'good', are based on an amended HBI equation, which weighted each taxa present based on its tolerance value [16,25]. Unlike the HBI scores for macroinvertebrates, the BCG approach for assessing freshwater health and level of ecological impairment includes nutrient concentrations, other anthropogenic stressors, and possible confounding variables, and facilitates understanding of correlations between diatom assemblages and variables such as percentage of forest in watershed [61].

Beyond metrics of water quality, understanding the stability of aquatic ecosystem networks is important for predicting long-term resilience in the face of local and global environmental change scenarios [64–67]. Generating networks of trophic interactions in freshwater systems can provide insight into ecosystem function, structure and robustness [65,68]. It has previously been shown that longer food webs are less stable and top predators more likely to go extinct [69–71]. The trophic networks in our study show small-world characteristics, whereby most nodes are not neighbors of each other, but the neighbors of a node are likely to be neighbors of each other forming clusters [59]. The short path lengths observed in our networks suggest that the effects of perturbations (e.g. species removal) would be distributed rapidly throughout the networks detected in our sites in a non-random fashion [38,59,72]. Our networks also show relatively low modularity, with ‘good’ sites displaying marginally higher modularity. Higher network modularity is suggested to reflect higher stability, often through enhancing species persistence [40,73,74]. By extension, lower network modularity may indicate that food webs in fair sites may be more susceptible to disturbance [41]. In terms of effects of environmental stressors such as pollutants, more modular networks are likely to limit the propagation of both pollutants and their indirect effects through the food web [75]. Despite higher richness in ‘fair’ sites, effective number of ESVs does not significantly differ among site conditions, indicating that rare taxa are more common in these ‘fair’ sites. Analyzing trait data for top predators at each site also indicates that ‘fair’ sites have more predators that also happen to be more resistant to poor water quality. Additionally, beta diversity, directed connectance, and modularity are all higher in ‘good’ sites, meaning these ‘good’ sites are expected to be more stable against persistent pollutant stress compared to ‘fair’ sites [73,75].

In addition to determining stability through modularity, it is vital to determine presence of keystone taxa and trophic hubs, whose loss would likely cause cascading extinctions of many other species within freshwater food webs [76]. Through targeting both diatoms and macroinvertebrates, we were able to determine keystone taxa from both producer and consumer trophic levels. Arthropod keystone taxa included genera from several traditional bioindicator groups (Ephemeroptera, Trichoptera and Chironomidae), of which perform a range of feeding strategies (e.g. collector-gatherer, collector-filterer and shredder). Despite diatoms often being excluded from network studies [48], the higher hub scores obtained for diatoms may reflect the importance of these producers as a food source for many different invertebrates [77]. Several taxa such as *Amphora*, *Gomphonema*, *Crictopus*, and *Polypedilum* were identified as both keystone taxa and site condition bioindicators further reinforcing the ability of eDNA metabarcoding approaches to generate a robust picture of site condition and stability.

Conclusion

There is an urgent need for more effective approaches to decipher biodiversity and ecosystem status as a consequence of environmental change especially due to global warming. Our study demonstrates that multi-taxa metabarcoding, is effective in identifying bioindicators of fine-scale freshwater condition and link these with their known tolerance to stressors. Unlike traditional methods, the use of multi-marker metabarcoding and indicator species analysis does not rely solely on the presence of EPT groups for making assessments of water quality, and enables a holistic measure of ecosystem health, even across previously identified subtle gradients of habitat quality status. While biodiversity analyses allowed us to distinguish site conditions based on alpha and beta diversity, correlation with environmental variables, as well as community composition, the addition of trophic network analyses also allowed us to identify clusters of taxa with known interactions, flag keystone taxa, and to assess ecosystem stability. Trophic networks derived from eDNA data provide information on which key indicator interactions could signal a change in environmental conditions of a site, as opposed to only looking at presence/absence of traditional bioindicator taxa. Despite excluding leaf litter/detritus/fungi/microbes as resources for invertebrates in our food webs, we were still able to reconstruct highly connected systems and present trans-kingdom keystone taxa. In the same way that metabarcoding is considered a scalable approach to biomonitoring by automating the taxonomic assignment process; the annotation of trophic interactions also needs to be automated to be a scalable approach. The continued growth of online biotic interaction databases, from ecological studies, and text mining from the literature [42,78], may one day help make the construction of global food webs a reality. Going forward, applying additional eDNA markers to target taxonomic groups such as fish, amphibians, and mammals, would greatly increase the level of trophic complexity in the networks and potentially identify additional bioindicator and keystone species, which may currently be overlooked in traditional biomonitoring strategies. Our work will set the stage for larger-scale studies involving sampling across a wide range of environmental gradients to further establish site condition bioindicator trends and potential influence of stressors on long-term ecosystem trophic networks.

Methods

Field Sampling

Samples were collected in November 2019 from Grand River tributaries across four study sites in Waterloo, Ontario (Supplementary Table 1; Supplementary Fig. 1). No specific permissions were required for sampling these sites because they are on public land and the field studies did not involve endangered or protected species. Status and location data were provided by Dougan & Associates based on a 2018 benthos biomonitoring project for the City of Waterloo (Supplementary Table 1). Clair15 and Clair12 are close in proximity, however Clair12 is directly downstream of several sewage outflows. The four selected sites were a subset of the sites from this project and were chosen based on accessibility and habitat quality. Hilsenhoff Biotic Index ranges (weighted by species) informed the habitat quality scale [79] which categorized sites into ‘Good’ (4.51-5.50) and ‘Fair’ (5.51-6.50).

Benthic kick-net samples were collected in triplicate within riffles, following the Canadian Aquatic Biomonitoring Network [CABIN] protocol [80], as previously described in [25]. All samples were collected in 1L sample jars and placed in a cooler to transport back to the lab. Upon arrival at the lab, samples (n = 12) were preserved using >99% ethanol and stored in a -20°C freezer until processing.

DNA Extraction

DNA from all samples was extracted following the methods previously detailed in [25]. Briefly, samples were homogenized using blenders decontaminated with ELIMINase1 (Decon Labs, Pennsylvania, USA), rinsed with deionized water, and treated with UV light for 30 minutes. Homogenate was transferred to 50 mL Falcon tubes, one tube was centrifuged at 2400 rpm for two minutes. Supernatant was removed and pellets were dried at 70 ℃. Approximately 300 mg dried tissue was used with the DNeasy Power Soil kit (Qiagen, CA) following the manufacturer's protocol. Final elution was in 50 µL of buffer C6 (TE). Negative controls with no tissue were included with each batch of extractions. All negative controls failed to amplify and were not sequenced.

DNA Amplification, Library Preparation and Sequencing

Diatom rbcL

DNA amplification of samples for generation of diatom sequences is detailed in [25]. Briefly, we targeted a 312 bp region of the chloroplast ribulose bisphosphate carboxylase large chain (rbcL) gene using five diatom specific primers: forward primers Diat_rbcL_708F_1, Diat_rbcL_708F_2 and Diat_rbcL_708F_3 combined in an equimolar mix; reverse primers Diat_rbcL_R3_1 and Diat_rbcL_R3_2 were also combined [23]. The PCR cocktail was comprised of 17.5 µL HyPure^{TM} molecular biology grade water, 2.5 µL 10X reaction buffer (200 mM Tris- HCl, 500 mM KCl, pH 8.4), 1 µL $MgCl_2$ (50 mM), 05. µL dNTPs mix (10 mM), 0.5 µL of both forward (10 mM) and reverse (10 mM) equimolar mixes, 0.5 µL Invitrogen's Platinum Taq polymerase (5 U) and 2 µL of DNA for a final reaction volume totaled 25 µL. The PCR protocol was as follows: 35 cycles of denaturation at 95℃ for 45 seconds, annealing at 55℃ for 45 seconds and extension at 72℃ for 45 seconds. PCR amplification was also performed in two-steps: the first step used the taxon-specific primers listed above, with the second PCR used 2 µL of amplicons from the first PCR as template, with Illumina-adapter tailed taxon-specific primers. One negative PCR control was included with each PCR step, which both came back negative thus were not carried through to sequencing. All PCRs were completed in Eppendorf Mastercycler ep gradient S thermal cycler. Successful amplification was confirmed using 1.5% agarose gel electrophoresis before purifying second PCR amplicons with the MinElute Purification kit (Qiagen).

Macroinvertebrate COI

Three fragments within the standard COI DNA barcode region were amplified with the following primer sets: (B/ArR5 [~310 bp] called BR5, LCO1490/230_R [~230 bp] called F230R, and mICOIintF/jgHCO2198 [~313 bp] called ml-jg [81–84] using a two-step PCR amplification regime as described above, with the exception of the cycler conditions which were: initial denaturation of 95°C for 5min, 35 cycles of 94°C for 40s, 46°C for 1min and 72°C for 30s with a final extension of 72°C for 5min before holding at 10°C until PCRs were removed from the cycler.

Purified amplicons were quantified using a QuantIT PicoGreen daDNA assay kit and all samples were then normalised to 3 ng/µL, pooling the COI fragments for each sample before indexing with Ilumina Nextera adapters (FC-131-2001). Once indexed, samples were pooled into a single library and purified with AMpure magnetic beads. QuantIT PicoGreen dsDNA assay kit was once again used to quantify the library and Bioanalyzer was used to determine fragment length. rbcL and COI fragments were sequenced separately over two partial MiSeq runs. The purified libraries were diluted to 4 nM and sequenced according to manufacturers protocol, using a 10% PhiX spike-in before being sequenced using Illumina MiSeq with a V3 MiSeq sequencing kit (300 bp X 2; MS-102-2003).

Bioinformatic Processing

Illumina MiSeq paired-end reads for both COI and rbcL were processed using the MetaWorks-1.3.1 pipeline [85,86] available from https://github.com/terrimporter/MetaWorks. MetaWorks is an automated Snakemake [87] bioinformatic pipeline that runs in a conda [88] environment. SeqPrep v1.3.2 [89] was used to pair raw reads requiring a minimum Phred score of 20 in the overlap region to ensure 99% base-calling accuracy and a minimum of 25 bp overlap. CUTADAPT v2.6 was used to trim primers from sequences, using a Phred score cutoff of 20 at the ends, leaving a minimum fragment length of at least 150 base pairs, no more than 3 N's permitted [90]. Global exact sequence variants (ESV) [91] were generated for the primer-trimmed reads. Reads were dereplicated using the 'derep_fulllength' command with the 'sizein' and 'sizeout' options of VSEARCH v2.14.1 [92]. VSEARCH was also used to denoise the data using the unoise3 algorithm [93]. These steps were taken to remove sequences with errors and rare reads (clusters with only one or two reads) [94]. Putative chimeric sequences were removed using the 'uchime3_denovo' algorithm in VSEARCH. An ESV x sample table was created using the 'search_exact' method in VSEARCH.

Diatom rbcL ESVs were classified using the rbcL diatom reference set available from https://github.com/terrimporter/rbcLdiatomClassifier [25,95]. The reference sequence set is based on rbcL sequences from the Diat.barcode project [96,97] and reformatted to train a naive Bayesian classifier to make rapid, accurate taxonomic assignments [98]. This method makes assignments to the species rank and produces a statistical measure of confidence for each taxon up to the domain rank to help reduce false positive taxonomic assignments. Species level assignments used a 90% bootstrap support cutoff, no cutoff was needed at the genus rank, to expect at least 90% correct taxonomic assignments assuming the query sequences are represented in the reference database.

Macroinvertebrate COI ESVs were classified using the COI Classifier v4 available from https://github.com/terrimporter/CO1Classifier/releases/tag/v4 [99], comprised of a curated reference sequence set mined from BOLD [100] and GenBank [101], and uses the RDP classifier v2.12 that uses a naive Bayesian algorithm [98,102]. We used a 0.70 bootstrap support cutoff at the species rank (90% correct), and no cutoff at the genus rank was needed, to expect 95% correct taxonomic assignments assuming the query sequences are represented in the reference database.

As we were using protein coding markers in this study, we also screened out obvious pseudogenes to try to reduce noise in the dataset and avoid inflating richness estimates [103]. For rbcL, we removed putative pseudogenes using removal method 1: rbcL ESVs were translated into every possible reading frame, plus strand only, using ORFfinder v0.4.3, keeping the longest open reading frame (ORF). ORFs with shorter or longer outlier sequence lengths were removed as putative pseudogenes. For COI, putative pseudogenes were identified and removed in the METAWORKS pipeline using removal method 2 since a hidden Markov model (HMM) was available for this marker. COI ESVs were translated into ORFs as described above, and for each ESV, the longest ORF was retained. Amino acid ORFs were used for HMM profile analysis and ORFs with low outlier sequence bit scores were removed as putative pseudogenes.

Data Analyses

RStudio v1.3.1093 was used with R v4.0.3 to analyze the data [104,105]. Plots were created with ggplot2 except for specialized plots where indicated [106]. Sites were plotted using the 'ggmap' package with stamen maps (Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL. Available from http://maps.stamen.com/#watercolor/12/37.7706/-122.3782) [107].To account for variable reads in each library, sample read number was normalized to the 15th percentile library size using the 'rrarefy' function in the vegan package [108,109]. Rarefaction curves were plotted using a slightly modified 'rarecurve' function (Supplementary Fig. 2). Sequencing depth was found to be sufficient to capture amplicon diversity across samples, as each curve reached a plateau, even after rarefying to the 15th percentile read depth. Unless otherwise stated, all further analyses are based on rarefied data.

We checked for differences in richness, the average number of ESVs per sample, site conditions (fair or good, 2 sites per condition, 3 replicates per site), and taxonomic groups (Arthropoda, Bacillariophyta). However, since richness is insensitive to species frequencies (rare species are weighted equally to common species), we also calculated the numbers equivalent, i.e., the effective number of equally-frequent species, using the exponential of Shannon entropy [110] in the vegetarian package [111], which is less sensitive to rare 'species'. Our richness and effective number of ESVs were found to be normally distributed using a quantile-quantile plot using the ggqqplot function in R as well as the Shapiro-Wilk test using the shapiro.test function in R ($p = 0.27$). T-tests were used to compare sample means across site conditions and across taxa within site conditions. The Holm method was used to adjust for multiple comparisons. To illustrate the range of biodiversity detected using multi-marker metabarcoding, as well as among-sample variability, we plotted a heatmap to visualise the detection of diatom and macroinvertebrate orders (Supplementary Fig. 3 and Supplementary Fig. 4). We reviewed our taxa and removed any non-freshwater taxa from our indicator lists. This may due to taxonomic misidentification in the underlying reference database [112].

A non-metric multi-dimensional scaling (NMDS) analysis on binary Bray-Curtis (Sorensen) dissimilarity matrix was conducted using the vegan 'metaMDS' function [113].

A scree plot was run using the ‘dimcheckMDS’ command from the goeveg package to determine the number of dimensions (k=2) to use with the vegan metaMDS function [114]. Shephard’s curve and goodness of fit plots were created using the vegan ‘stressplot’ and ‘goodness’ functions. The vegan ‘vegdist’ command was used to build a binary Bray Curtis dissimilarity matrix. We checked for heterogeneous distribution of dissimilarities using the ‘betadisper’ function. We used the ‘adonis’ function to perform permutational multivariate analysis of variance (PERMANOVA). PERMANOVA [115] was performed to assess whether sites or site status had any significant interactions or explained any variation in beta diversity. Environmental variables (temperature, percentage dissolved oxygen, pressure, specific conductance, pH and turbidity) were fitted to the NMDS plot using the ‘envfit’ function in vegan with 999 permutations. Only significant variables ($p<0.05$) were plotted.

We used the ‘multipatt’ function from the indicspecies R package to: identify species that could be used as indicators of site quality using the Indicator Value method (IndVal); and to identify species correlated with environmental conditions at fair and poor sites using the point biserial correlation coefficient [57,58]. Both functions used the “.g” version to correct for unequal group sizes and we set duleg=TRUE to avoid considering site group combinations. We included taxa at the species rank where we could, otherwise we retained the genus level assignment and appended the ESV ID. Each test was run using a taxon x sample table containing rarefied read counts. We analyzed the diatom and macroinvertebrate species assemblages independently, to determine the strongest indicators within each taxonomic group.

For trophic analyses, we worked with taxa at the species and genus rank as described above except that we did not append ESV IDs to genera. The list of target taxa was manually reviewed and edited to account for insufficiently identified species assignments with ‘sp.’, ‘cf.’, or an alphanumeric code [116]. We also summarized identifications at the variety level, containing ‘var.’, to the species rank. We obtained biotic interactions for each taxon from the Global Biotic Interactions (GloBI) database [31] using the ‘get_interactions_by_taxa’ function in the rglobi package in R [117]. In our first search, we retrieved interactions for some of the species and genera in our target list. For species that were not detected, we collapsed the taxonomic assignment to the genus rank and conducted a second search. For all searches conducted at the genus rank, the retrieved interactions were pooled across the species within the genus. We filtered interactions to only keep ones where both the resource and consumer were detected in a site. We then filtered the interactions to only keep those that described the directed resource to consumer relationship (ex. “eatenBy”, “preyedUponBy”). With the cheddar package, we used our directed resource to consumer interactions to calculate several measures of trophic complexity including number of nodes (species), trophic links (interactions), linkage density (links/species), directed connectance (links/species^2), characteristic path length (average of path lengths from each node to a basal species), as well as trophic height (chain averaged trophic length) [118–120]. The food webs were visualized using the ‘PlotWebByLevel’ function with the chain averaged trophic level method.

For network analysis, we used the same resource to consumer relationships from above to build directed graphs using the 'graph.data.frame' function in the igraph package. We used the walktrap method to identify communities of potentially interacting (co-occurring) taxa, also referred to as modules, clusters, groups, or subgraphs in the literature. The 'cluster_walktrap' function identifies clusters via random walks, with the assumption that short random walks tend to stay in the same community, and edges within a cluster are denser than edges between clusters. For each site, we also assessed overall network modularity, the strength of divisions of a network into modules (clusters). We categorized modularity as follows: very low (< 0.1), low (0.1-0.15), medium (0.15-0.2), high (0.2-0.3), and very high (> 0.3) [121]. We used two different measures to identify keystone taxa, that is, nodes with high centrality. First, we calculated degree by recording the number of edges (co-occurrences) into and out of a vertex (species). Second, we calculated Kleinberg's hub centrality scores (hub scores) that takes into consideration the authority and hubbiness of a vertex [60]. An authority refers to the number of links into a vertex and hubbiness refers to the number of links out to other vertices with high authority. Degree and hub scores were calculated for every taxon in the network and for each measure, the top 3 taxa from each site were retained as a potential keystone taxon. The network was circularized with vertices ordered by cluster membership.

References

1. Ormerod, S. J., Dobson, M., Hildrew, A. G. & Townsend, C. R. Multiple stressors in freshwater ecosystems. *Freshw. Biol.* **55**, 1–4 (2010).

2. Reid, A. J. *et al.* Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biol. Rev.* **94**, 849–873 (2019).
3. Geist, J. Integrative freshwater ecology and biodiversity conservation. *Ecol. Indic.* **11**, 1507–1516 (2011).
4. Geist, J. Seven steps towards improving freshwater conservation. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **25**, 447–453 (2015).
5. Kroll, S. A. *et al.* Large-scale protection and restoration programs aimed at protecting stream ecosystem integrity: the role of science-based goal-setting, monitoring, and data management. *Freshw. Sci.* **38**, 23–39 (2019).
6. Park, Y.-S. & Hwang, S.-J. Ecological monitoring, assessment, and management in freshwater systems. *Water* **8**, 234 (2016).
7. Estrada, E. Food webs robustness to biodiversity loss: The roles of connectance, expansibility and degree distribution. *J. Theor. Biol.* **244**, 296–307 (2007).
8. Gilbert, A. J. Connectance indicates the robustness of food webs when subjected to species loss. *Ecol. Indic.* **9**, 72–80 (2009).
9. Compson, Z. G. *et al.* Network-based biomonitoring: exploring freshwater food webs with stable isotope analysis and DNA metabarcoding. *Front. Ecol. Evol.* **7**, 395 (2019).
10. Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C. & Baird, D. J. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLOS ONE* **6**, e17497 (2011).

11. Baird, D. J. & Hajibabaei, M. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* **21**, 2039–2044 (2012).

12. Keck, F., Vasselon, V., Tapolczai, K., Rimet, F. & Bouchez, A. Freshwater biomonitoring in the Information Age. *Front. Ecol. Environ.* **15**, 266–274 (2017).

13. *Freshwater biomonitoring and benthic macroinvertebrates*. (Springer US, 1993).

14. Chang, F.-H., Lawrence, J. E., Rios-Touma, B. & Resh, V. H. Tolerance values of benthic macroinvertebrates for stream biomonitoring: assessment of assumptions underlying scoring systems worldwide. *Environ. Monit. Assess.* **186**, 2135–2149 (2014).

15. Bush, A. *et al.* Studying ecosystems with DNA metabarcoding: lessons from biomonitoring of aquatic macroinvertebrates. *Front. Ecol. Evol.* **7**, 434 (2019).

16. Mandaville, S. M. *Benthic Macroinvertebrates in Freshwaters- Taxa Tolerance Values, Metrics, and Protocols*. 128 http://lakes.chebucto.org/H-1/tolerance.pdf (2002).

17. Hilsenhoff, W. L. Rapid field assessment of organic pollution with a family-level biotic index. *J. North Am. Benthol. Soc.* **7**, 65–68 (1988).

18. Carter, J. L. & Resh, V. H. *Analytical approaches used in stream benthic macroinvertebrate biomonitoring programs of State agencies in the United States*. 56 http://pubs.er.usgs.gov/publication/ofr20131129 (2013).

19. Vasselon, V., Rimet, F., Tapolczai, K. & Bouchez, A. Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). *Ecol. Indic.* **82**, 1–12 (2017).

20. Wu, N. *et al.* Using river microalgae as indicators for freshwater biomonitoring: Review of published research and future directions. *Ecol. Indic.* **81**, 124–131 (2017).

21. Chonova, T. *et al.* Benthic diatom communities in an alpine river impacted by waste water treatment effluents as revealed using DNA metabarcoding. *Front. Microbiol.* **10**, 653 (2019).

22. Kermarrec, L. *et al.* A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshw. Sci.* **33**, 349–363 (2014).

23. Rivera, S. F. *et al.* Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia* **807**, 37–51 (2018).

24. Tapolczai, K. *et al.* Diatom DNA metabarcoding for biomonitoring: strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. *Front. Ecol. Evol.* **7**, 409 (2019).

25. Maitland, V. C., Robinson, C. V., Porter, T. M. & Hajibabaei, M. Freshwater diatom biomonitoring through benthic kick-net metabarcoding. *PLOS ONE* **15**, e0242143 (2020).

26. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **21**, 2045–2050 (2012).

27. Yu, D. W. *et al.* Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.* **3**, 613–623 (2012).

28. Creer, S. *et al.* The ecologist's field guide to sequence-based identification of biodiversity. *Methods Ecol. Evol.* **7**, 1008–1018 (2016).
29. U.S. EPA. *Freshwater Biological Traits Database (Final Report)*. https://www.epa.gov/risk/freshwater-biological-traits-database-data-sources (2012).
30. Schmidt-Kloiber, A. & Hering, D. www.freshwaterecology.info – An online tool that unifies, standardises and codifies more than 20,000 European freshwater organisms and their ecological preferences. *Ecol. Indic.* **53**, 271–282 (2015).
31. Poelen, J. H., Simons, J. D. & Mungall, C. J. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecol. Inform.* **24**, 148–159 (2014).
32. Menezes, S., Baird, D. J. & Soares, A. M. V. M. Beyond taxonomy: a review of macroinvertebrate trait-based community descriptors as tools for freshwater biomonitoring. *J. Appl. Ecol.* **47**, 711–719 (2010).
33. Pawlowski, J. *et al.* The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci. Total Environ.* **637–638**, 1295–1310 (2018).
34. Taberlet, P., Bonin, A., Zinger, L. & Coissac, E. Environmental DNA for functional diversity. in *Environmental DNA* (Oxford University Press, 2018). doi:10.1093/oso/9780198767220.003.0010.
35. Caputi, S. S. *et al.* Seasonal food web dynamics in the Antarctic benthos of Tethys Bay (Ross Sea): implications for biodiversity persistence under different seasonal sea-ice coverage. *Front. Mar. Sci.* **7**, 594454 (2020).

36. Sentis, A., Montoya, J. M. & Lurgi, M. Warming indirectly increases invasion success in food webs. *Proc. R. Soc. B Biol. Sci.* **288**, 20202622 (2021).

37. Faust, K. Open challenges for microbial network construction and analysis. *ISME J.* **15**, 3111–3118 (2021).

38. Dunne, J. A., Williams, R. J. & Martinez, N. D. Food-web structure and network theory: The role of connectance and size. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12917–12922 (2002).

39. Thompson, R. M. *et al.* Food webs: reconciling the structure and function of biodiversity. *Trends Ecol. Evol.* **27**, 689–697 (2012).

40. Grilli, J., Rogers, T. & Allesina, S. Modularity and stability in ecological communities. *Nat. Commun.* **7**, 12031 (2016).

41. Gilarranz, L. J., Rayfield, B., Liñán-Cembrano, G., Bascompte, J. & Gonzalez, A. Effects of network modularity on the spread of perturbation impact in experimental metapopulations. *Science* **357**, 199–201 (2017).

42. Compson, Z. G. *et al.* Chapter two - Linking DNA metabarcoding and text mining to create network-based biomonitoring tools: a case study on boreal wetland macroinvertebrate communities. in *Advances in Ecological Research* (eds. Bohan, D. A., Dumbrell, A. J., Woodward, G. & Jackson, M.) vol. 59 33–74 (Academic Press, 2018).

43. Delmas, E. *et al.* Analysing ecological networks of species interactions. *Biol. Rev.* **94**, 16–36 (2019).

44. Sousa, L. L. de, Silva, S. M. & Xavier, R. DNA metabarcoding in diet studies: Unveiling ecological aspects in aquatic and terrestrial ecosystems. *Environ. DNA* **1**, 199–214 (2019).

45. Everard, M., Fletcher, M. S., Powell, A. & Dobson, M. K. The feasibility of developing multi-taxa indicators for landscape scale assessment of freshwater systems. *Freshw. Rev.* **4**, 1–19 (2011).

46. Chon, T.-S. *et al.* Evaluation of stream ecosystem health and species association based on multi-taxa (benthic macroinvertebrates, algae, and microorganisms) patterning with different levels of pollution. *Ecol. Inform.* **17**, 58–72 (2013).

47. Duarte, I. A., Reis-Santos, P., França, S., Cabral, H. & Fonseca, V. F. Biomarker responses to environmental contamination in estuaries: A comparative multi-taxa approach. *Aquat. Toxicol. Amst. Neth.* **189**, 31–41 (2017).

48. Seymour, M. *et al.* Executing multi-taxa eDNA ecological assessment via traditional metrics and interactive networks. *Sci. Total Environ.* **729**, 138801 (2020).

49. Mills, L. S., Soulé, M. E. & Doak, D. F. The keystone-species concept in ecology and conservation. *BioScience* **43**, 219–224 (1993).

50. Resh, V. H. Which group is best? Attributes of different biological assemblages used in freshwater biomonitoring programs. *Environ. Monit. Assess.* **138**, 131–138 (2008).

51. Robinson, C. V. *et al.* Combining DNA and people power for healthy rivers: Implementing the STREAM community-based approach for global freshwater monitoring. *Perspect. Ecol. Conserv.* **19**, 279–285 (2021).

52. Spaulding, S. *Amphora. Diatoms of North America.* https://diatoms.org/genera/amphora (2011).

53. Spaulding, S. & Edlund, M. *Cyclotella. In Diatoms of North America*. https://diatoms.org/genera/cyclotella (2008).

54. Spaulding, S. & Edlund, M. *Nitzschia. Diatoms of North America*. https://diatoms.org/genera/nitzschia (2008).

55. Elbrecht, V. & Leese, F. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PLOS ONE* **10**, e0130324 (2015).

56. Elbrecht, V., Peinert, B. & Leese, F. Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecol. Evol.* **7**, 6918–6926 (2017).

57. Dufrêne, M. & Legendre, P. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol. Monogr.* **67**, 345–366 (1997).

58. Cáceres, M. D., Jansen, F. & Dell, N. *indicspecies: Relationship Between Species and Groups of Sites*. (2020).

59. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).

60. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. in *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms* 668–677 (Society for Industrial and Applied Mathematics, 1998).

61. Hausmann, S., Charles, D. F., Gerritsen, J. & Belton, T. J. A diatom-based biological condition gradient (BCG) approach for assessing impairment and developing nutrient criteria for streams. *Sci. Total Environ.* **562**, 914–927 (2016).

62. Spaulding, S. & Edlund, M. *Gomphonema. In Diatoms of North America.* https://diatoms.org/genera/gomphonema (2009).

63. Davies, S. P. & Jackson, S. K. The biological condition gradient: a descriptive model for interpreting change in aquatic ecosystems. *Ecol. Appl.* **16**, 1251–1266 (2006).

64. Aoki, I. & Mizushima, T. Biomass diversity and stability of food webs in aquatic ecosystems. *Ecol. Res.* **16**, 65–71 (2001).

65. Dunne, J. A., Williams, R. J. & Martinez, N. D. Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecol. Lett.* **5**, 558–567 (2002).

66. Araújo, M. B., Rozenfeld, A., Rahbek, C. & Marquet, P. A. Using species co-occurrence networks to assess the impacts of climate change. *Ecography* **34**, 897–908 (2011).

67. Beauchard, O., Veríssimo, H., Queirós, A. M. & Herman, P. M. J. The use of multiple biological traits in marine community ecology and its potential in ecological indicator development. *Ecol. Indic.* **76**, 81–96 (2017).

68. Johnson, S., Domínguez-García, V., Donetti, L. & Muñoz, M. A. Trophic coherence determines food-web stability. *Proc. Natl. Acad. Sci.* **111**, 17923–17928 (2014).

69. Long, Z. T., Bruno, J. F. & Duffy, J. E. Food chain length and omnivory determine the stability of a marine subtidal food web. *J. Anim. Ecol.* **80**, 586–594 (2011).

70. LeCraw, R. M., Kratina, P. & Srivastava, D. S. Food web complexity and stability across habitat connectivity gradients. *Oecologia* **176**, 903–915 (2014).

71. Shanafelt, D. W. & Loreau, M. Stability trophic cascades in food chains. *R. Soc. Open Sci.* **5**, 180995 (2018).

72. Montoya, J. M. & Solé, R. V. Small world patterns in food webs. *J. Theor. Biol.* **214**, 405–412 (2002).

73. May, R. M. Will a large complex system be stable? *Nature* **238**, 413–414 (1972).

74. Teng, J. & McCann, K. S. Dynamics of compartmented and reticulate food webs in relation to energetic flows. *Am. Nat.* **164**, 85–100 (2004).

75. Garay-Narváez, L., Flores, J. D., Arim, M. & Ramos-Jiliberto, R. Food web modularity and biodiversity promote species persistence in polluted environments. *Oikos* **123**, 583–588 (2014).

76. Zhao, L. *et al.* Weighting and indirect effects identify keystone species in food webs. *Ecol. Lett.* **19**, 1032–1040 (2016).

77. Frauendorf, T. C. *et al.* Energy flow and the trophic basis of macroinvertebrate and amphibian production in a neotropical stream food web. *Freshw. Biol.* **58**, 1340–1352 (2013).

78. Muñoz, G. *Literature thesis: Building a framework for retrieving information on multispecies interactions from published literature*. (2017).

79. Gazendam, E., Gharabaghi, B., Jones, F. C. & Whiteley, H. Evaluation of the Qualitative Habitat Evaluation Index as a Planning and Design Tool for Restoration of Rural Ontario Waterways. *Can. Water Resour. J. Rev. Can. Ressour. Hydr.* **36**, 149–158 (2011).

80. Environment Canada. Canadian aquatic Biomonitoring Network -Field Manual: Wadeable Streams. (2013).

81. Hajibabaei, M., Spall, J. L., Shokralla, S. & van Konynenburg, S. Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecol.* **12**, 28 (2012).

82. Geller, J., Meyer, C., Parker, M. & Hawk, H. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Mol. Ecol. Resour.* **13**, 851–861 (2013).

83. Gibson, J. *et al.* Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proc. Natl. Acad. Sci.* **111**, 8007–8012 (2014).

84. Gibson, J. F. *et al.* Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLOS ONE* **10**, e0138432 (2015).

85. Porter, T. M. *MetaWorks: A Multi-Marker Metabarcode Pipeline*. (Zenodo, 2020). doi:10.5281/zenodo.4741407.

86. Porter, T. M. & Hajibabaei, M. METAWORKS: A flexible, scalable bioinformatic pipeline for multi-marker biodiversity assessments. *bioRxiv* 2020.07.14.202960 (2020).

87. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

88. Anon. *Anaconda Software Distribution*. (2020).

89. St John, J. *SeqPrep: Tool for stripping adaptors and/or merging paired reads with overlap into single reads.* (2021).

90. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

91. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).

92. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

93. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 081257 (2016) doi:10.1101/081257.

94. Reeder, J. & Knight, R. The ‘rare biosphere’: a reality check. *Nat. Methods* **6**, 636–637 (2009).

95. Porter, T. M. *RbcL Diat.barcode Reference Set For The RDP Classifier*. (Zenodo, 2020). doi:10.5281/zenodo.4741478.

96. Rimet, F. *et al.* R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. *Database J. Biol. Databases Curation* **2016**, baw016 (2016).

97. Rimet, F. *et al.* Diat.barcode, an open-access curated barcode library for diatoms. *Sci. Rep.* **9**, 15116 (2019).

98. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).

99. Porter, T. M. *Eukaryote CO1 Reference Set For The RDP Classifier*. (Zenodo, 2017). doi:10.5281/zenodo.4741447.

100. Ratnasingham, S. & Heber, P. D. N. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes* **7**, 355–364 (2007).

101. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).

102. Porter, T. M. & Hajibabaei, M. Automated high throughput animal CO1 metabarcode classification. *Sci. Rep.* **8**, 4226 (2018).

103. Porter, T. M. & Hajibabaei, M. Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. *BMC Bioinformatics* **22**, 256 (2021).

104. R Studio Team. *RStudio: Integrated Development Environment for R*. (2021).

105. R Core Team. *R: A language and environment for statistical computing.* (2020).

106. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, 2009). doi:10.1007/978-0-387-98141-3.

107. Kahle, D. *ggmap: A package for plotting maps in R with ggplot2*. (2021).

108. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).

109. Oksanen, J. *et al. vegan: Community Ecology Package*. (2020).

110. Jost, L. Entropy and diversity. *Oikos* **113**, 363–375 (2006).

111. Charney, N. & Record, S. *Vegetarian: Jost Diversity Measures for Community Data*. (2021).

112. Meiklejohn, K. A., Damaso, N. & Robertson, J. M. Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. *PLOS ONE* **14**, e0217084 (2019).

113. Hajibabaei, M., Porter, T. M., Wright, M. & Rudar, J. COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PLOS ONE* **14**, e0220953 (2019).

114. Goral, F. & Schellenberg, J. *goeveg R-package: Functions for Community Data and Ordinations*. (2018).

115. Robinson, C. V., Porter, T. M., Wright, M. T. G. & Hajibabaei, M. Propylene glycol-based antifreeze is an effective preservative for DNA metabarcoding of benthic arthropods. *Freshw. Sci.* **40**, 77–87 (2021).

116. Nilsson, R. H. *et al.* Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLOS ONE* **1**, e59 (2006).

117. Poelen, J. H., Gosnell, S. & Slyusarev, S. *rglobi:R library to access species interaction data of http://globalbioticinteractions.org*. (rOpenSci, 2021).

118. Williams, R. J. & Martinez, N. D. Limits to trophic levels and omnivory in complex food webs: theory and data. *Am. Nat.* **163**, 458–468 (2004).

119. Jonsson, T., Cohen, J. E. & Carpenter, S. R. Food webs, body size, and species abundance in ecological community description. in *Advances in Ecological Research* vol. 36 1–84 (Academic Press, 2005).

120. Hudson, L. N. *et al.* Cheddar: analysis and visualisation of ecological communities in R. *Methods Ecol. Evol.* **4**, 99–104 (2013).

121. Jacobson, D. K. *et al.* Functional diversity of microbial ecologies estimated from ancient human coprolites and dental calculus. *Philos. Trans. R. Soc. B Biol. Sci.* **375**, 20190586 (2020).

Acknowledgements

This study is funded by the Government of Canada through Genome Canada and Ontario Genomics. Teresita Porter was funded by the Government of Canada through the Genomics Research and Development Initiative (GRDI) Metagenomics-Based Ecosystem Biomonitoring Project (Ecobiomics).

Author Contributions and Competing Interests

C.V.R and M.H. designed the study, C.V.R contributed to sampling, bioinformatic processing and statistical analyses, T.M.P created the bioinformatic pipeline and trained the classifiers, contributed to bioinformatic processing and performed data analysis, V.C.M. collected samples and conducted molecular analyses, M.T.G.W. conducted molecular analysis and sequenced the samples. All authors helped to write/edit the manuscript.

Additional Information

The authors have declared that no competing interests exist.

Figure Legends

Figure 1. Observed richness is driven by a large number of rare ESVs. A) richness and B) The effective number of ESVs is shown for each condition and taxonomic group.

Figure 2. Beta diversity was greater within ‘good’ sites, especially for diatoms. Beta diversity within ‘fair’ sites, tended to be lower, especially for diatoms. Sample replicates cluster by site, and sites cluster by site status. Environmental variables that correlate with beta diversity patterns are shown if they have a p-value < 0.05. Based on binary Bray Curtis dissimilarities from libraries where read counts were rarefied to the 15th percentile. Abbreviations: dissolved oxygen (DO); pressure (mmHg), temperature (Temp).

Figure 3. Food webs from ‘fair’ sites had higher trophic height and more macroinvertebrate predators. Vertical food webs with the lowest trophic level, producers (diatoms), at the bottom. Estimated trophic position for each node in the food web was determined using chain averaged trophic level (trophic height).

Figure 4. Freshwater benthic food webs show low to medium levels of modularity. Though these are directed graphs with interactions that point from resource to consumer, arrow heads were removed from the plot to improve readability. Trophic links (edges) are shown in grey. The small grey loops indicate cannibals, taxa known to consume members of the same taxon. Nodes (taxa) were arranged in a circle

and colored according to trophic position (green-producers (diatoms); blue-macroinvertebrates). Clusters of taxa, as well as isolated taxa, are circled in black.

Figure 5. Kleinburg hub scores reflect the importance of diatoms in linking macroinvertebrate consumers together in trophic networks. Two measures used to detect potential keystone taxa are shown: 1) degree centrality and Kleinberg's hub centrality scores. Degree shows the number of connections into and out of each taxon node. Hub scores reflect the number of connections out of each taxon node (hubs), from resource to consumer, weighted more havily when the consumers are themselves linked to by many other hubs. Sites are shown as follows: A) Beaver18 (good), B) Clair15 (good), C) Clair12 (fair), and D) Laurel7 (fair).

Table 1. Diatom (n=15) and macroinvertebrate (n=17) habitat status indicator taxa. Based on rarefied data using ESV richness. Only shows species with a p-value < 0.05.

Site Status	Phylum	Species / Genus + ESV ID	Indicator Value	Point biserial correlation coefficient	BCG score	HBI score
Fair	Annelida	*Bothrioneurum vejdovskyanum*	0.913	0.680		7.0
Fair	Annelida	*Nais stolci*	0.913	0.596		8.0
Fair	Annelida	*Tubifex tubifex*	0.911	0.309		10.0
Fair	Arthropoda	*Candona candida*	0.913	0.508		8.0
Fair	Arthropoda	*Cricotopus triannulatus*	1.000	0.751		7.0
Fair	Arthropoda	*Cypridopsis vidua*	0.913	0.553		8.0
Fair	Arthropoda	*Libellula* ml-jg_Zotu156	1.000	0.560		9.0
Fair	Arthropoda	*Microtendipes pedellus*	0.913	0.524		6.0
Fair	Arthropoda	*Orthocladius oliveri*	0.949	-		6.0
Fair	Arthropoda	*Polypedilum convictum*	1.000	0.741		6.0
Fair	Arthropoda	*Simulium vittatum*	0.959	0.605		7.0
Fair	Arthropoda	*Tanytarsus guerlus*	0.987	0.547		6.0
Fair	Bacillariophyta	*Amphora* rbcL_Zotu149	0.836	-	4.0	6.0
Fair	Bacillariophyta	*Amphora* rbcL_Zotu201	0.913	0.477	4.0	
Fair	Bacillariophyta	*Amphora* rbcL_Zotu71	1.000	0.633	4.0	
Fair	Bacillariophyta	*Cyclotella* rbcL_Zotu109	0.913	0.581	5.0	
Fair	Bacillariophyta	*Cyclotella* rbcL_Zotu76	1.000	0.720	5.0	
Fair	Bacillariophyta	*Discostella* rbcL_Zotu32	0.913	0.583	3.0	
Fair	Bacillariophyta	*Fallacia* rbcL_Zotu128	1.000	0.598	4.0	
Fair	Bacillariophyta	*Gomphonema pumilum* rbcL_Zotu146	1.000	0.566	2.0	
Fair	Bacillariophyta	*Gomphonema* rbcL_Zotu9	0.992	0.477	2.0	
Fair	Bacillariophyta	*Nitzschia* rbcL_Zotu206	1.000	0.793	4.0	
Fair	Bacillariophyta	*Nitzschia* rbcL_Zotu25	0.962	0.559	4.0	
Fair	Bacillariophyta	*Nitzschia* rbcL_Zotu340	0.913	0.767	4.0	
Fair	Bacillariophyta	*Nitzschia* rbcL_Zotu91	1.000	0.701	4.0	
Fair	Bacillariophyta	*Sellaphora* rbcL_Zotu69	1.000	0.924	4.0	
Fair	Bacillariophyta	*Tabularia* rbcL_Zotu150	1.000	0.544	4.0	
Fair	Mollusca	*Pisidium casertanum*	0.896	-		6.0
Good	Arthropoda	*Optioservus ovalis*	0.984	0.534		4.0

Abbreviations: zero radius OTU (Zotu) also known as an exact sequence variant
+ = non-aquatic species (bird feather mite); * = non-aquatic species (hoverfly)
BCG: Biological Condition Gradient (averaged across species for genus); BCG attributes for each taxon: 2 'Highly sensitive taxa'; 3 'Intermediate sensitive taxa'; 4 'Intermediate tolerant, ubiquitous taxa'; 5 'Tolerant taxa' (Hausman et al., 2016)
HBI: Hilsenhoff Biotic Index (species-level); Tolerance values for each taxon range from 0 (very intolerant of organic wastes - 10 (very tolerant of organic wastes) (Mandaville, 2002)

Table 2. 'Good' sites tended to have higher directed connectance and modularity whereas 'fair' sites tended to have a greater number of nodes, trophic links, and trophic height.

Sites	Status	# Nodes	# Trophic links	Linkage density	Directed connectance	Characteristic path length	Trophic height	Modularity	Clusters* (total clusters)
Beaver 18	Good	83	316	3.8	0.05	2.4	3.7	0.08	5 (8)
Clair15	Good	45	106	2.4	0.05	2.2	2.9	0.16	4 (16)
Clair12	Fair	108	402	2.7	0.03	2.3	4.8	0.06	7 (10)
Laurel7	Fair	105	390	3.7	0.04	3	4.4	0.06	7 (15)

*Number of clusters with more than one taxon (total clusters including isolated taxa)

Table 3. More top predators that are tolerant to organic wastes are found in fair sites compared to good condition sites. Taxa with a trophic height greater than 3 are listed below, except for Clair 15 where the top macroinvertebrate predator is listed.

Site Condition	Site	Predators (Trophic Height)	Functional Feeding Group[1]	Tolerance[2,3]
Good	Beaver 18	*Oecetis* (3.00) *Hydropsyche* (3.48) *Conchapelopia* (3.66)	PR CG/CF/PR PR	5 1-6 6
Good	Clair 15	*Dicranota* (2.9375)	PR	3
Fair	Clair 12	*Demicryptochironomus* (3.00) *Parachironomus* (3.30) *Procladius* (3.58) *Hydropsyche* (3.59) *Conchapelopia* (3.77) *Enallagma* (4.18) *Libellula* (4.80)	CG/PR PR/CG/PA PR/CG CG/CF/PR PR PR PR	8 10 9 1-6 6 8 9 [3]
Fair	Laurel 7	*Conocephalus* (3.00), *Oecetis* (3.00), *Orconectes* (3.00), *Carabus* (3.33), *Libellula* (3.41), *Hydropsyche* (3.51), *Conchapelopia* (3.67), *Argiope* (4.38), *Graphoderus* (4.41)	PR PR CG/PR PR PR CG/CF/PR PR PR PR	- 5 6 - 9 1-6 6 - 5

[1] Functional feeding groups are from the EPA freshwater traits database or GloBI

[2] Ranges for tolerance values are from Mandaville (2002). Values range from 0 (intolerant to organic waste) to 10 (very tolerant to organic waste)

[3] For *Libellula* the tolerance for Libellulidae is given; '-' indicates a tolerance value is unavailable; For *Orconectes* the tolerance for Cambaridae given

Table 4. Diatom (n=6) and macroinvertebrate (n=5) putative keystone taxa.

Phylum	Genus*	Functional feeding guild	BCG Score	HBI score
Bacillariophyta	*Navicula*		4.0	
Bacillariophyta	*Fragilaria*		3.0	
Bacillariophyta	*Rhoicosphenia*		3.0	
Bacillariophyta	*Amphora*		4.0	
Bacillariophyta	*Diatoma*		2.0	
Bacillariophyta	*Gomphonema*		2.0	
Arthropoda	*Cricotopus* (Chironomidae)	Collector-Gatherer		7.0
Arthropoda	*Hydropsyche* (Trichoptera)	Collector-Filterer		4.0
Arthropoda	*Polypedilum* (Chironomidae)	Shredder		6.0
Arthropoda	*Baetis* (Ephemeroptera)	Collector-Gatherer		6.0
Platyhelminthes	*Dugesia*	N/A		6.0

*Keystone taxa selected from top three degree and top three hub scores from each site
N/A: Not applicable
BCG: Biological Condition Gradient (averaged across species for genus)
HBI: Hilsenhoff Biotic Index (genus-level)

Data Availability

Raw sequences will be available from NCBI SRA on acceptance. The MetaWorks-1.3.1 is available from https://github.com/terrimporter/MetaWorks, the rbcLdiatomClassifier v1 and COIClassifier v4 we used are available on GitHub at https://github.com/terrimporter/rbcLdiatomClassifier and https://github.com/terrimporter/CO1Classifier. Scripts and files used to generate outputs can be found at https://github.com/terrimporter/RobinsonEtAl2021_MacroinvertDiatom.