Identifying North American freshwater invertebrates using DNA barcodes: are existing COI sequence libraries fit for purpose?

Colin J. Curry[1,4], **Joel F. Gibson**[1,2,5], **Shadi Shokralla**[3,6], **Mehrdad Hajibabaei**[3,7], **and Donald J. Baird**[1,8]

[1]Environment and Climate Change Canada at Canadian Rivers Institute, P.O. Box 4400 Fredericton, New Brunswick E3B 5A3 Canada
[2]Royal British Columbia Museum, 675 Belleville Street, Victoria, British Columbia V8W 9W2 Canada
[3]Biodiversity Institute of Ontario and Department of Integrative Biology, University of Guelph, 50 Stone Road East Guelph, Ontario N1G 2W1 Canada

Abstract: We reviewed the availability of cytochrome *c* oxidase subunit I (COI) sequences for 2534 North American freshwater invertebrate genera in public databases (GenBank and Barcode of Life Data Systems) and assessed representation of genera commonly encountered in the Canadian Aquatic Biomonitoring Network (CABIN) database. COI sequence records were available for 61.2% of North American genera and 72.4% of Insecta genera in public databases. Mollusca (73.9%) and Nematoda (15.4%) were the best and worst represented groups, respectively. In CABIN, 85.4% of genera had COI sequence records, and 95.2% of genera occurring in >1% of samples were represented. Genera absent from CABIN tended to be uncommon or members of groups not routinely used for biomonitoring purposes. On average, 94.1% of genera in well-identified samples had associated sequence data. To leverage the full potential of genomics approaches, we must expand DNA-barcode reference libraries for poorly described components of freshwater food webs. Some genera appear to be well represented (e.g., *Eukiefferiella*), but deposited sequences represent few sampling localities or few species and lead to underestimation of sequence diversity at the genus level and reduced confidence in identifications. Public COI libraries are sufficiently populated to permit routine application of genomics tools in biomonitoring, and ongoing quality assurance/quality control should include re-evaluation as new COI reference sequences are added or taxonomic hierarchies change. Next, we must understand whether and how established biomonitoring approaches can capitalize on high-throughput sequencing tools. Biomonitoring approaches that use genomics data to facilitate structural and functional assessments are fertile ground for future investigation and will benefit from continued improvement of publicly available sequence libraries.

Key words: COI, invertebrates, biomonitoring, high-throughput sequencing, DNA metabarcoding, identification, genus, Biomonitoring 2.0

Rapid advances in our ability to obtain biodiversity information by sequencing genetic material from environmental samples, such as homogenized bulk tissue samples (Hajibabaei et al. 2011, Gibson et al. 2014) and water (e.g., Ficetola et al. 2008) or soil (e.g., Fahner et al. 2016) are transforming the way that we infer biodiversity and monitor environmental change (Hajibabaei et al. 2011, Chariton et al. 2015). Genomics technologies can provide taxonomic and functional data at the ecosystem level, thereby enabling large-scale environmental assessment (e.g., Gibson et al. 2015). In Fig. 1, we summarize the basic steps involved in the use of high-throughput sequencing (HTS) in biodiversity studies. HTS technologies can be used to process multiple samples in parallel during a single sequencing run with tagged primers, thereby yielding millions of sequence reads (Shokralla et al. 2012, 2014, 2015). When taxon identification is the goal, HTS can be targeted to specific DNA-barcode gene regions. These sequences can be clustered to generate molecular operational taxonomic units (MOTU) (Blaxter et al. 2005) that closely reflect species or compared to reference databases to identify the associated taxa present in a sample (Porter et al. 2014).

E-mail addresses: [4]curry.colin@gmail.com; [5]jgibson@royalbcmuseum.bc.ca; [6]sshokral@uoguelph.ca; [7]mhajibab@uoguelph.ca; [8]djbaird@unb.ca

DOI: 10.1086/696613. Received 9 January 2017; Accepted 15 September 2017; Published online 24 January 2018.
Freshwater Science. 2018. 37(1):178–189. © 2018 by The Society for Freshwater Science.

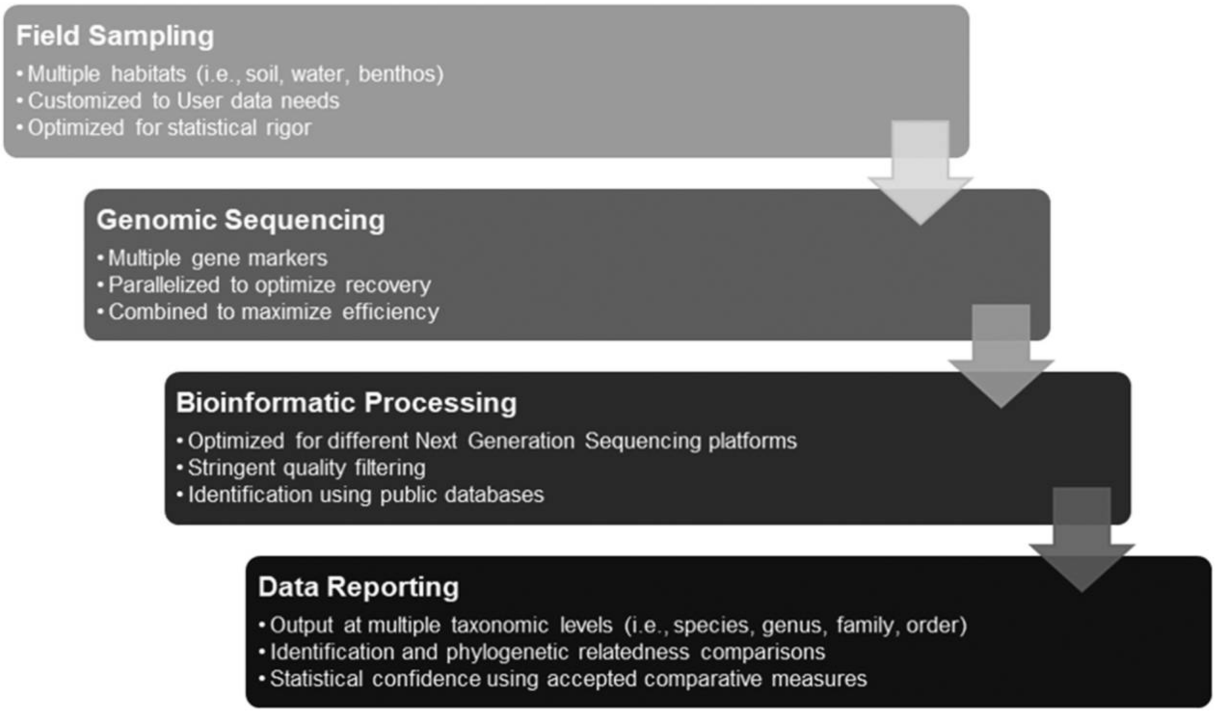
Field Sampling
• Multiple habitats (i.e., soil, water, benthos)
• Customized to User data needs
• Optimized for statistical rigor
Genomic Sequencing
• Multiple gene markers
• Parallelized to optimize recovery
• Combined to maximize efficiency
Bioinformatic Processing
• Optimized for different Next Generation Sequencing platforms
• Stringent quality filtering
• Identification using public databases
Data Reporting
• Output at multiple taxonomic levels (i.e., species, genus, family, order)
• Identification and phylogenetic relatedness comparisons
• Statistical confidence using accepted comparative measures

Figure 1. A typical high-throughput sequencing for biomonitoring workflow.

Biodiversity measurement and biomonitoring traditionally have relied on the collection of samples from local biotic assemblages. Invertebrates are often the target taxon for biomonitoring studies in freshwater habitats because of their high species richness, ubiquity, and range of responses to anthropogenic stressors (Rosenberg and Resh 1993). After collection and preservation, samples routinely are subsampled before sorting and identification. For instance, the Canadian Aquatic Biomonitoring Network (CABIN) uses a Marchant Box (Marchant 1989) for subsampling (Environment Canada 2012). Nevertheless, sample processing is often time consuming, particularly for studies involving large numbers of samples. Given the costs of training and the lack of taxonomic expertise, identification is usually carried out to family level, with genus-level identification in special circumstances. Disagreements in taxonomic assignment based on morphology can occur between taxonomists, often at high rates for difficult-to-identify groups (Stribling et al. 2008), so identification usually is limited to a taxonomic level at which quality can be controlled. The CABIN sample database contains >16,000 samples, but the presence of early life stages and specimen damage often preclude identification below order or family. Approximately 58% of specimens are identified to genus, but this number reflects a relatively small number of samples that were not subsampled and were subjected to higher-than-usual taxonomic effort. Only 3279 samples had >70% of individuals identified to genus level (https://www.ec.gc.ca/rcba-cabin/; data extracted 3 September 2014). Some CABIN taxa also are listed as a couplet at the genus level (e.g., Diptera *Bezzia/Probezzia*). Moreover, many invertebrate taxa are not included in final counts or are included at coarse taxonomic resolution. For instance, most crustacean taxa are excluded from CABIN counts (Environment Canada 2012), and members of phyla, such as Annelida, are rarely identified below class level. Even for insect groups for which larval taxonomy is relatively well understood, late instars often are required for confident identification.

HTS can provide taxonomic information at greater resolution, depth, and consistency, and at lower cost than morphologically identified samples (Gibson et al. 2015). HTS ultimately provides molecular operational taxonomic units (MOTUs) that closely approximate species, but these taxon lists cannot be related to existing trait databases or other knowledge about organism biology. Therefore, application of HTS for taxonomic identification of samples in a biomonitoring context is limited by the availability of cytochrome *c* oxidase subunit 1 (COI) sequence records in reference databases. Numerous investigators have shown that without adequate representation in a reference library, obtaining accurate taxonomic identification for a given sequence can be very difficult (e.g., Ekrem et al. 2007, Wilson et al. 2011).

Numerous campaigns exist to complete barcode libraries for specific groups (e.g., Trichoptera; www.trichopterabol.org), or regions (e.g., Arctic; Zhou et al. 2009), but little documentation exists for the completeness of barcode libraries and their ability to provide a sufficient description of genetic diversity of the wider freshwater invertebrate community. Published lists of barcoded taxa exist for certain key groups (e.g., Ephemeroptera; Webb et al. 2012), and regions (e.g., fauna of the Great Lakes, Trebitz et al. 2015; macroinvertebrates of Australia, Carew et al. 2017), but how comprehensively HTS can be applied in the context of freshwater biomonitoring or where library-building activities must be focused to fully realize the potential of HTS tools is not clear at present.

The primary objectives of our study were to assess the state of publicly available COI reference sequences for North American freshwater invertebrate genera and to estimate COI library completeness for the most commonly encountered genera in CABIN samples. Other genetic markers are available for species identification (e.g., 16S ribosomal RNA), but COI has been used extensively for metazoan barcoding because it consistently discriminates among closely related species (Sweeney et al. 2011) and established campaigns exist to populate COI barcode libraries (Ratnasingham and Hebert 2007). We chose to focus on the completeness of genus-level libraries for 3 reasons. First, as mentioned earlier, processing of biomonitoring samples focuses primarily on family-level identification, so genus-level identification represents a significant increase in taxonomic resolution over standard practice. Second, species lists for North American invertebrate taxa are incomplete, particularly for poorly studied groups, such as Nematoda and Rotifera (Traunspurger 2000, Segers 2008). A similar statement could be made for genus lists, but we think it likely that the proportion of unknown genera is much lower than the proportion of unknown species. Moreover, the availability of functional trait information in databases, such as the US Environmental Protection Agency Freshwater Biological Traits Database (https://www.epa.gov/risk/freshwater-biological-traits-database-traits), is lacking or incomplete at the species level and is often summarized at genus level for trait-based biomonitoring.

A full survey of the genetic diversity of North American freshwater invertebrates represented at the genus level in COI libraries was beyond the scope of our study. However, we saw value in illustrating some potential pitfalls created by incomplete reference libraries by looking in detail at sequence representation for contrasting genera: 1 with many sequences, broad geographic coverage, and numerous species (Ephemeroptera:*Baetis*) and 1 with many sequences but restricted geographic coverage (Diptera:*Eukiefferiella*). Genera with few or no associated sequences can be assumed to have poor species and geographic coverage in COI libraries, reflecting incomplete knowledge of COI sequence diversity. The purpose of our analysis is to inform discussion of COI sequence library-building activities moving forward.

METHODS

List of North American freshwater invertebrate genera

We compiled lists of freshwater invertebrate genera from a wide variety of sources. We compiled arthropod lists primarily from taxonomic keys, specifically those by Merritt et al. (2008) and Thorp and Covich (2010). We compiled additional information for Canadian lists for the key insect groups Ephemeroptera, Plecoptera, Trichoptera, and Odonata (EPTO) and for bivalve mollusks from existing and forthcoming General Status Assessment reports (CESCC 2011). We compiled lists for other invertebrate phyla primarily from information published by Thorp and Covich (2010) and lists maintained by Daniel Graf (http://winvertebrates.uwsp.edu). In total, we included 2534 North American invertebrate genera in 16 phyla/subphyla, 29 classes, and 532 families in the list for our study. Our list is likely to be incomplete but is comparable to the 2468 Nearctic invertebrate genera recorded in the Freshwater Animal Diversity Assessment by Balian et al. (2008). Nomenclature is a potential issue when querying databases for sequence records because sequences associated with outdated scientific names may be missed. Where possible, we used valid genus names recognized by the Integrated Taxonomic Information System (ITIS). A full list of genera queried in our search is included in Table S1.

Database search

We searched GenBank (Benson et al. 2015) and Barcode of Life Data Systems (BOLD) (Ratnasingham and Hebert 2007) for publicly available COI sequence records. Considerable overlap exists between these databases. All BOLD sequences ultimately are deposited in GenBank, and the BOLD public-record search tool mines GenBank for sequence records. However, instances can exist where data publicly available in BOLD have not yet been released to GenBank, and some data available through GenBank may not be captured by the BOLD public-data search tool. For our GenBank search query, we included multiple names for the COI region (COI, cox1, coxI, CO1, cytochrome *c* oxidase subunit I, cytochrome *c* oxidase subunit 1) and returned confirmed COI sequences that were between 300 and 5000 base pairs (bp) in length and excluded sequences labeled as pseudogenes, sequences containing multiple Ns, unverified specimens, and whole genome sequences. We chose the sequence length minimum because even partial COI sequence records (e.g. mini-barcodes) can be used for identification, particularly at the genus level (Meusnier et al. 2008, Gibson et al. 2014). We did not exclude partial sequences on the 3′ end of the COI gene. These sequences are not considered part of the COI barcode region, they are uncommon in sequence libraries for the same reason, and are unlikely to significantly affect our results. The BOLD public-data search tool does not permit specification of sequence length minima, but returns only verified sequences >500 bp in length. Our searches could have captured sequences with inappropriate base calls resulting from weak reads, but verification of this problem would be difficult for such a large data set. Nevertheless, this situation could lead to overestimation of the number of genera represented in libraries.

We used the Integrated Taxonomic Information System (ITIS) as a taxonomic standard when preparing our genus lists. ITIS provides lists of synonyms where possible, but these lists cannot be considered exhaustive. Where searches based on ITIS taxonomy did not yield any sequence records, we conducted searches based on known synonyms. Searches were conducted between 12 April and 4 May 2016.

Data analysis

We recorded the number of sequences returned by each query separately for both databases. Given the overlap in results between BOLD and GenBank, we included the maximum number of sequences returned from either query when combining results. We calculated the percentage of genera with >0, 1–10, 11–25, and >25 COI sequences for each database and the combined data set, and we performed these calculations at phylum/subphylum and class levels.

Relatively uncommon taxa often are downweighted in biomonitoring studies because their occurrence is difficult to predict in most monitoring and reference-condition models (Cao et al. 2001, Reynoldson et al. 2001). Therefore, we were interested in gauging the availability of COI DNA sequence records for the most commonly encountered genera in freshwater biomonitoring studies. To create this list, we summarized information from 16,671 benthic invertebrate samples collected between 1987 and 2013 in the CABIN sample database. The samples were distributed across Canada, but significant gaps existed in spatial coverage (Fig. 2). This summary yielded a list of 804 genera for which the % genera represented were calculated using the previously described categories. This list was further separated into 'common' (occurring in >1% of samples) and less commonly encountered genera. For a subset of 3279 samples in which >70% of individuals had been identified to genus and ≥100 individuals were identified, we also calculated the proportion of genera for which COI sequence records were available.

COI sequence diversity analysis

We downloaded COI sequences for *Baetis* and *Eukiefferiella* from BOLD on 18 May 2016. We included only sequences >350 bp long, and we excluded flagged sequences, misidentifications, or sequences containing stop codons. We aligned sequences for each genus (*Baetis* and *Eukiefferiella*) manually with MEGA5 software (Tamura et al. 2011). We calculated minimum, maximum, and average % nucleotide sequence differences based on pairwise comparisons of all sequences. We mapped sequences with associated geographic

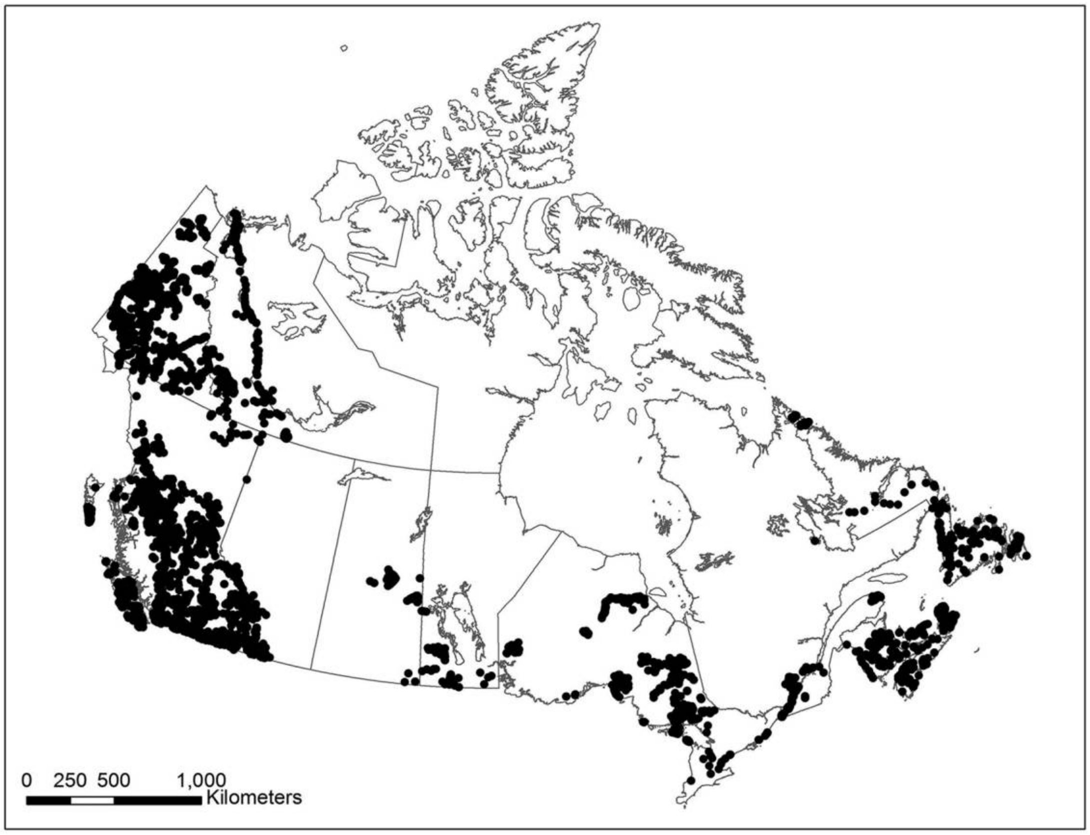

Figure 2. Distribution of the 16,671 Canadian Aquatic Biomonitoring Network (CABIN) samples collected between 1993 and 2013 that were considered for this analysis.

data with ArcGIS® (version 10.2; Environmental Systems Research Institute, Redland, California) software to illustrate geographic coverage in North America and generated bubble plots to illustrate the relative number of sequences for species within each genus.

RESULTS

GenBank and BOLD survey

Across both databases, 61.2% of genera were associated with COI sequence records (Fig. 3), but either database on its own contained COI sequences for ~51% of genera. Noncrustacean arthropods were the most diverse group of genera, with 1500 genera recorded in North America. Across both BOLD and GenBank, 69.2% of these genera had COI reference sequences recorded, but only 53.9% of genera were represented in GenBank and 56.1% in BOLD. This result included particularly low representation for class Arachnida (principally aquatic mites, 171 genera), where only 43.4% of genera were represented in either database. Henceforth, we report only combined results for GenBank and BOLD.

For subphylum Crustacea (409 genera), the 2nd-most diverse, 48.4% of genera were associated with COI sequence records in either database, though this ranged from 22.5% for class Ostracoda (89 genera) to 63.6% for class Branchiopoda (110 genera) (Fig. 4). Phylum Mollusca (161 genera) was the most complete, with 73.9% of genera with COI sequence records in either reference database. Phylum Nematoda (117 genera) was the least complete, with 15.3% of genera with COI sequence records in either reference library.

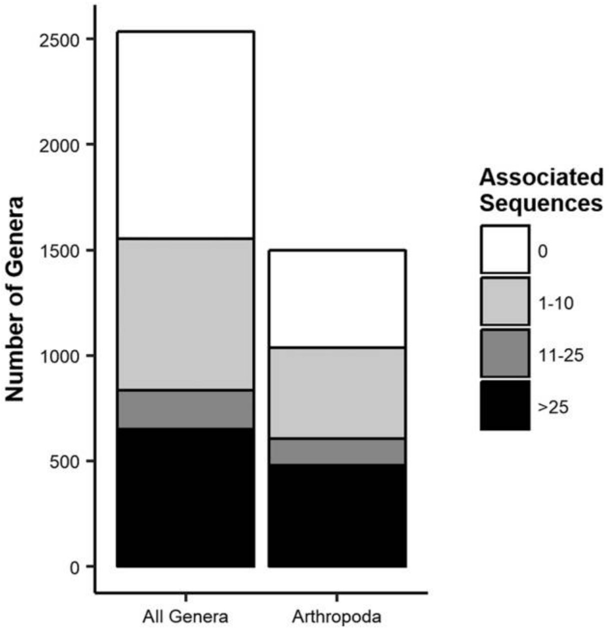

Figure 3. Number of genera with associated cytochrome *c* oxidase subunit I (COI) reference sequences and distribution of COI sequence abundance for freshwater invertebrate genera across all phyla and for noncrustacean arthropods.

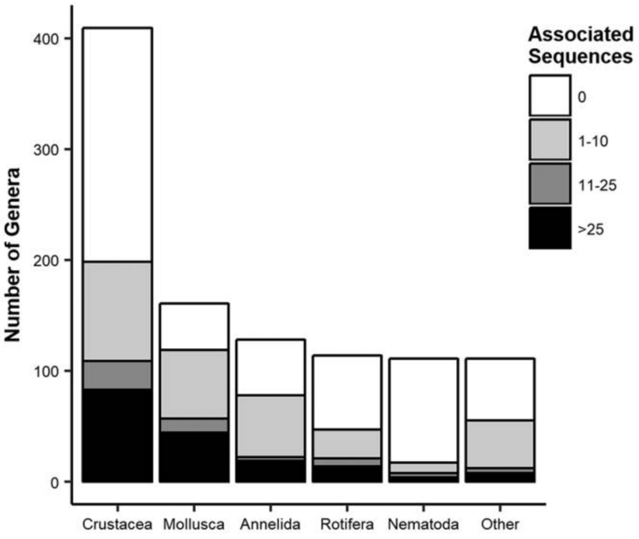

Figure 4. Number of genera with associated cytochrome *c* oxidase subunit I (COI) reference sequences and distribution of COI sequence abundance for freshwater invertebrate genera across remaining phyla/subphyla.

Insects were by far the most diverse class of invertebrates, represented by 1305 genera in our study (Fig. 5). COI sequences were associated with 72.4% of aquatic insect genera. This representation was not evenly distributed among insect orders, but ≥50% of genera within each order except Orthoptera (4/9 genera) had associated COI sequence records. Within the orders Diptera, Ephemeroptera, Hymenoptera, and Lepidoptera, ≥50% of genera with COI sequence records had >10 associated sequences. Trichoptera, a key group for freshwater biomonitoring, had the greatest sequence representation, with associated COI sequence records for 87.9% of genera. However, only 41.4% of genera had >25 associated sequence records. Plecoptera and Odonata had the lowest genus-level representation (60.0 and 62.2%, respectively) among other orders routinely scrutinized in freshwater biomonitoring.

Representation of genera in the CABIN database

Taxonomic resolution varied greatly across samples and phyla. The CABIN assessment protocol is focused on running waters in wadeable rivers, so certain groups (e.g., Crustacea) usually are excluded from totals, whereas others (e.g., Annelida) typically are not identified below a coarse level of resolution. Across the entire data set, only 57.7% of individuals were identified to genus level.

Of the 804 unique genera identified in this data set, 206 occurred in >1% of samples. Figure 6 illustrates that more frequently encountered genera are more likely to be associated with COI sequence records. Across all genera 688 (85.4%) had associated COI sequence records in either database, and

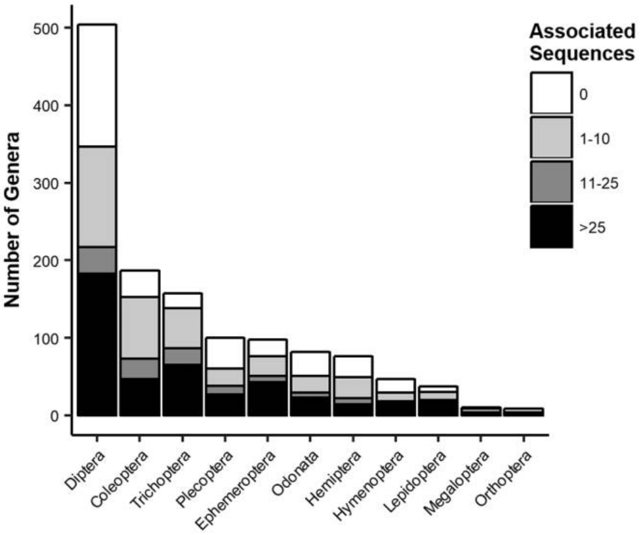

Figure 5. Number of genera with associated cytochrome *c* oxidase subunit I (COI) reference sequences associated with genera and distribution of COI sequence abundance for freshwater insect orders.

395 of these had >25 associated sequences. For genera occurring in >1% of samples, 197 genera (95.2%) had records and 146 of these had >25 associated sequences. In the subset of 3279 samples in which >70% of individuals were identified to genus based on morphology, an average of 94.1% of genera/sample had associated COI sequence data. The most frequently encountered genus lacking associated COI sequence records was the balloon fly genus *Metachela* (Diptera:Empididae), the 58th most common genus, occurring in 6.3% of samples. Of the 100 most frequently encountered genera, 3 lacked associated COI sequence records: *Metachela*, *Visoka* (Plecoptera:Nemouridae), and *Testudacarus* (Trombidiformes: Torrenticolidae).

COI sequence diversity analysis

The genus *Baetis* was represented by 2750 sequences. Of these sequences, 2262 represented 29 unique species, whereas the remaining species were identified only to genus. The average genetic difference between sequence pairs was 16.2%, and the minimum difference was 0% (115,257 pairs, 3.06% of all comparisons). The maximum distance between pairs was 28.2%. Figure 7A illustrates the broad geographic distribution of *Baetis* COI sequences in North America, and Fig. 8 demonstrates the high variability in the number of sequences across species within the genus.

The genus *Eukiefferiella* was represented by 1008 sequences and 1 species in BOLD. Most sequences were from specimens identified only at the genus level. The average genetic difference between sequence pairs was 8.6%, and the minimum difference was 0% (56,753 sequence pairs, 11.2% of all comparisons). The maximum difference between pairs was 19.6%. Figure 7B illustrates the comparatively restricted geographic distribution of *Eukiefferiella* sequences in North America.

DISCUSSION

As freshwater biomonitoring moves toward a diagnostic, trait-based framework (Poff et al. 2006, Baird and Hajibabaei 2012), demand is increasing for higher taxonomic resolution and better representation of the full benthic community than is routinely obtained in most biomonitoring programs. For example, >12.5 million individuals were identified in samples collected by CABIN personnel across Canada from 1993 to 2013, but only 57.7% were identified to genus level based on morphology. Similarly, Orlofske and Baird (2014) found that only 49% of the Ephemeroptera, Plecoptera, Trichoptera, Odonata (EPTO) in a typical benthic sample were identifiable to genus using the best available taxonomic keys. In comparison, publicly accessible COI DNA sequence records exist for 61.2% of North American aquatic invertebrate genera (74.4% when considering EPTO). More than 85% of genera in the CABIN data set have associated COI sequence records, and that number increases to >90% when considering the most frequently encountered genera. These numbers reflect the taxonomic effort required in the CABIN protocols, and many taxa routinely identified at a coarse level or not included in count totals are not included in CABIN genus lists. Others may not be identified consistently to genus across all samples from all collection sites. Hence, assignment of genera as common or rare must be considered in

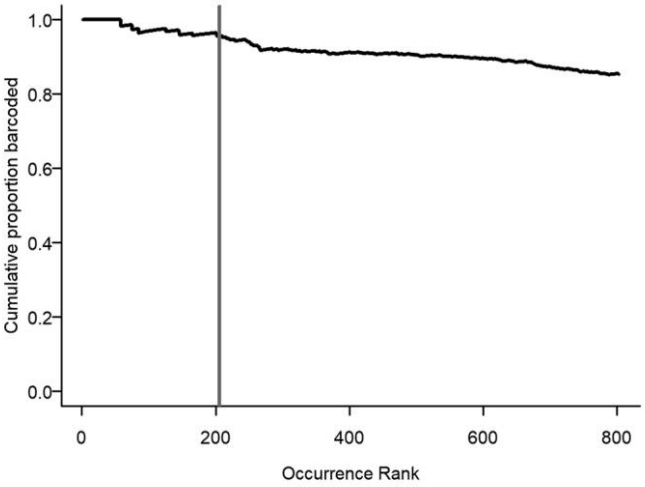

Figure 6. Cumulative proportion of Canadian Aquatic Biomonitoring Network (CABIN) genera present in public cytochrome *c* oxidase subunit I (COI) sequence libraries. Genera were ranked from most (1) to least (804) common based on the proportion of samples in which they occurred. The cumulative proportion barcoded was recalculated with the inclusion of successively less common genera. Genera to the left of the gray line are considered "common", occurring in >1% of samples.

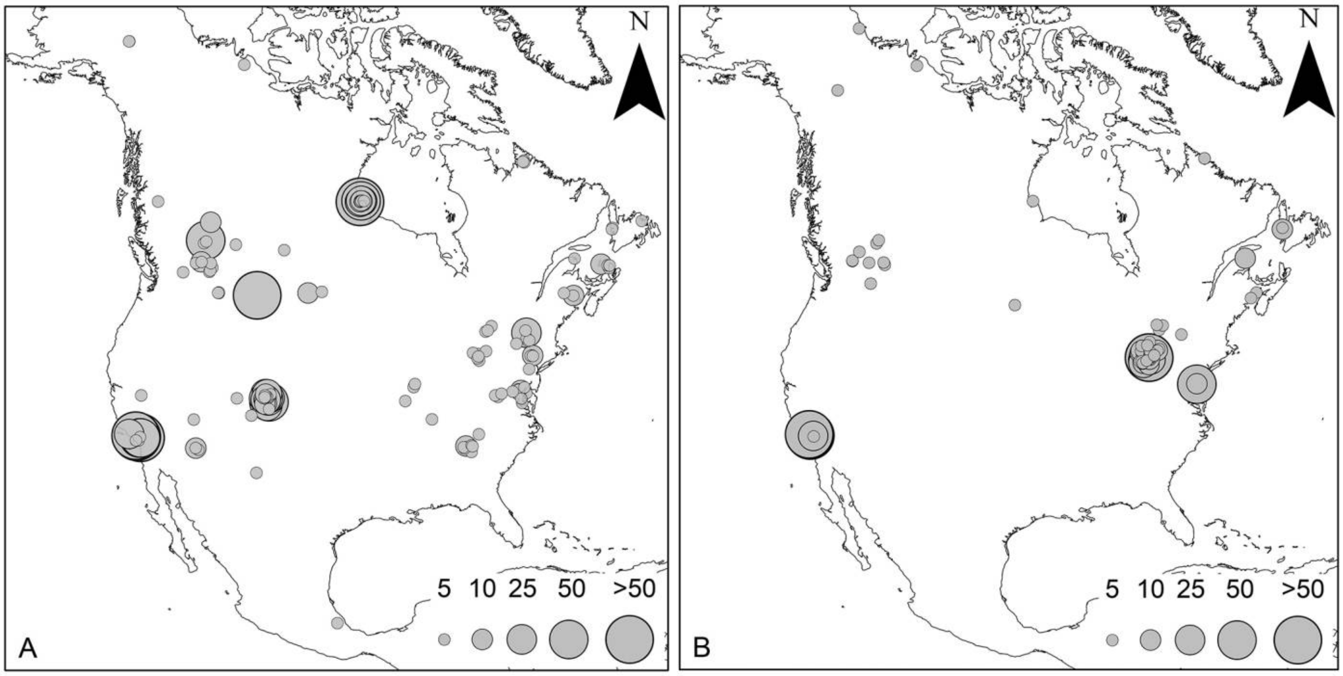

Figure 7. Distribution and relative abundance of cytochrome *c* oxidase subunit I (COI) sequence records extracted from the Barcode of Life Data systems for specimens within the genera (A) *Baetis* and (B) *Eukiefferiella*. The size of circles reflects the number of sequence records from a locality.

that context. For this reason alone, expecting genomics tools to return lists of taxa that perfectly match lists derived from morphology is unreasonable. Results emerging from recent studies indicate that increased taxonomic resolution provided by DNA-based identification leads to increased sensitivity and discriminatory power for biodiversity metrics (Pilgrim et al. 2011, Stein et al. 2014, Gibson et al. 2015).

Representation for certain invertebrate groups in public DNA COI-sequence libraries is poor and may limit the applicability of DNA barcode analysis in certain habitats or contexts. For example, Maxillopoda and Ostracoda are important arthropod components of the food web in lentic systems, but <<50% of the genera in these groups are included in public DNA libraries. Furthermore, many of the DNA COI-sequence records available in public databases are not identified to the species or genus level, but merely family or order (Kwong et al. 2012). For some orders, such as harpacticoid copepods, the paucity of COI sequences may simply reflect a lack of attention from researchers (Watson et al. 2015). These factors may limit the utility of DNA-based approaches in lentic habitats over the short term. However, Gibson et al. (2015) found that the number of invertebrate families and genera identified in a set of wetland samples using an HTS approach was far greater than that obtained from morphological approaches.

Some groups may have relatively high representation at the genus level, but these genera may be represented by few total associated sequences. Having a greater number of sequences for a given genus, especially from geographically distinct populations, will improve confidence in identifications, particularly for speciose genera (Lou and Golding 2012). The effect of an insufficient DNA reference library will differ among taxonomic groups. For example, an interspecies COI sequence difference of 2% is consistent for most invertebrate groups (Jackson et al. 2014, White et al. 2014), but can be as high as 10% for groups, such as Oligochaeta (Vivien et al. 2015). This variability may affect taxonomic identification and recovery. For example, we found that >50% of oligochaete genera have COI sequence records, but most of these are associated with <10 sequences. In this instance, more oligochaete genera are needed in libraries and at greater total numbers to identify any oligochaete DNA in an environmental sample accurately.

The robustness of COI libraries for identification purposes depends on several factors. Many genera contain large numbers of species. For instance, the Trichoptera genus *Limnephilus* is represented by ≥64 species in Canada alone (CJC, unpublished data), so one might expect to see greater sequence variation at the genus level for *Limnephilus* than for a genus containing 1 or just a few species. Our analysis of COI sequence diversity for a well-studied genus (*Baetis*) highlighted a large degree of COI sequence diversity (average genetic distance: 16.2%). Therefore, one can be confident that an unknown sequence that matches at, e.g., 90 to 95% similarity is a true *Baetis* sequence. The high level of ge-

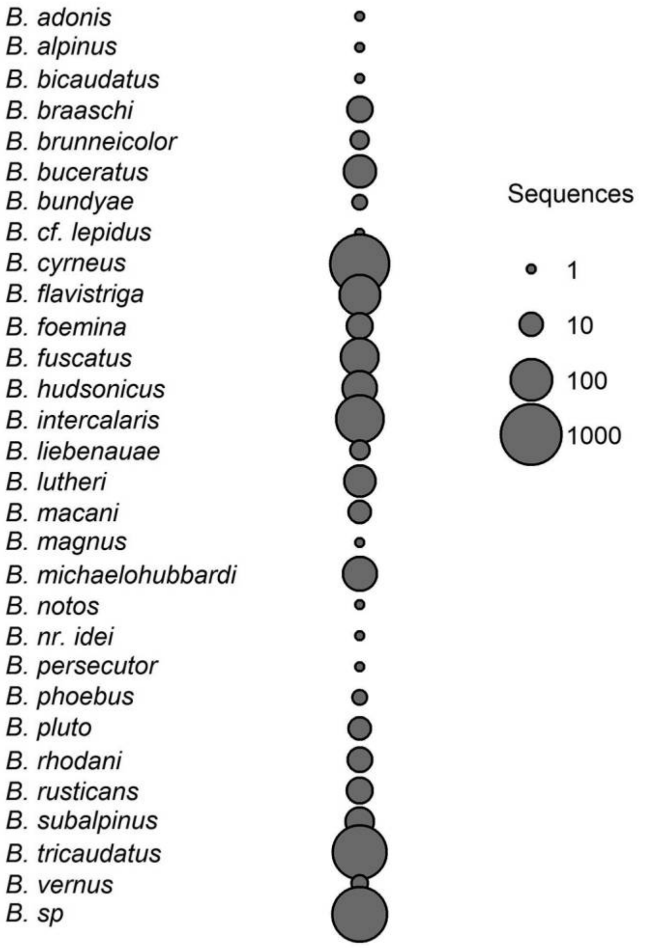

Figure 8. Relative abundance of cytochrome *c* oxidase subunit I (COI) sequence records among different *Baetis* species from North America. An equivalent plot is not shown for *Eukiefferiella* because only 1 species was represented among the sequences.

netic diversity in *Baetis* probably is caused by its high species diversity (19 species recognized by ITIS; most of these included in BOLD) and its ancient origins. Fossils from the family Baetidae are known from the lower Cretaceous (120–135 mya; McCafferty 1997). The *Baetis* sequences in the BOLD database have broad geographic distribution, but most sequences come from a few localities (Fig. 7A).

Variability in reference sequences at the genus level may be underestimated if sequences represent a single species or specimens collected from a restricted location. *Eukiefferiella* is one such example. Although represented by >1000 sequences in BOLD, records for this genus include just 1 definitively identified species. ITIS recognizes 15 species. Many unique sites are represented, but >60% of the records come from 4 localities (Fig. 7B). Thus, COI diversity within the genus may be underestimated, although sequences listed as *Eukiefferiella* sp. may represent additional species. Were additional library building to be pursued for this genus (particularly increasing the number of species represented within the genus), the diversity of COI sequences in the library probably would increase and further our ability to assign unknown sequences to it with confidence.

Recent work on invertebrates suggests that taxa can be assigned correctly to genus or higher taxonomic levels with a high degree of confidence (Wilson et al. 2011, Gibson et al. 2014, Shokralla et al. 2014). Ensuring that reference material has broad geographic coverage, that multiple reference sequences exist for each genus, and that numerous species are included for the most speciose genera will greatly reduce the likelihood of incorrect assignment at the genus level and will help to flag potentially misidentified reference sequences (Ekrem et al. 2007, Lou and Golding 2012, Porter et al. 2014).

Both BOLD and GenBank take measures to ensure that voucher material is correctly identified and associated with accurate reference sequences. However, clerical and identification mistakes can still occur. Moreover, HTS is likely to reveal a large amount of cryptic diversity at both the species and genus level. As the state of phylogenetic and taxonomic knowledge changes (e.g., species within 1 genus are split into different genera), samples should be revisited to ensure that identifications reflect the current state of knowledge. In practice this rarely happens because retrieving and re-identifying many archived specimens is prohibitively time consuming. Sequence information is stored digitally and easily searched, so a further advantage of a DNA-based approach is that identifications can be updated rapidly to reflect current knowledge, provided changes to taxonomic knowledge are accompanied by barcode sequence information. Revisiting identifications based on genomics tools (and any resulting data analysis) should be considered a part of quality assurance/quality control workflows for biomonitoring studies.

In the short term, groups that are routinely used for biomonitoring activities and considered responsive to environmental stress should be the focus of library-building activities. For riverine ecosystems, efforts could be directed toward expanding and completing the reference library for EPTO orders and Diptera. EPTO are widely considered to be sensitive to pollution and hydrological alteration (Compin and Céréghino 2003), whereas Diptera demonstrate a wide range of responses (Nicacio and Juen 2015), from high tolerance to sensitivity. More than ⅔ of North American EPTO genera already have associated COI sequence records, whereas the Diptera genus library is >60% complete. Many of the genera missing from these lists probably are uncommon at the continental scale or difficult to identify. The continued support and assistance of research taxonomists will be necessary to increase completeness of COI sequence libraries, particularly for habitats, such as wetlands, where COI libraries for numerically dominant taxa (e.g., crustacean zooplankton) are far from complete. This task is urgent given the globally rapid rate of wetland loss (Davidson 2014). The focus of our review is the genus level because considerable trait information has been collated for freshwater taxa at this level of taxonomic resolution. However, a genus-level approach does not preclude identification of specimens to species-level where reference-sequence information exists or molecular OTUs where it does not.

A further consequence of higher-resolution biodiversity information obtained through DNA-based approaches is that descriptions of aquatic invertebrate assemblages will be radically altered. Insects are numerically dominant and speciose and often constitute the bulk of taxa identified in biomonitoring samples, particularly in lotic systems (Cushing and Allan 2001). However, larger-bodied predatory taxa (e.g., Odonata) are less abundant, even if they were encountered regularly in fully identified samples. Their relative abundance in a sample is difficult to estimate accurately because only a small subsample usually is identified (Courtemanch 1996, Vinson and Hawkins 1996, Walsh 1997). For instance, Odonata are widely distributed across Canada, including the areas covered by CABIN samples. However, odonate taxa are recorded in only 16.5% of CABIN samples. Some odonate taxa specialize in edge habitats that are poorly sampled with CABIN field protocols (Curry et al. 2012), but we find it somewhat surprising that odonates would be encountered so infrequently. Subsampling might lead to underrepresentation of this insect order. Large predatory invertebrates probably will be better represented in biomonitoring data sets assembled based on HTS tools because their inclusion is less prone to errors from count-based subsampling. Likewise, rare genera present at only a small number of sites can greatly affect our ability to assess degrees of stressor effects (Stein et al. 2014).

The models and analytical approaches currently used in biomonitoring and bioassessment studies must be adapted to use the information provided by genomics tools fully. Traditional approaches based on taxonomic composition have been used to provide pass/fail assessments of sites or to position sites within deviance envelopes from expected conditions, but users struggle to provide mechanistic explanations for observed effects. Ecosystem response to stress is mediated through species interactions, so investigating how networks of ecological interactions and their properties change along stressor gradients may allow more robust and mechanistically grounded assessments (Gray et al. 2014). An ecological network approach is feasible only if the nodes (i.e., taxa) and interactions within the network can be resolved adequately. HTS technologies are crucial to the broader application of such approaches because they can provide sufficiently detailed taxonomic information across the entire food web. Taxa that play a large role in organic-matter processing and energy transfer or parasitic interactions, but that often are discounted during morphological identification, will be better described. For instance, nematodes are thought to constitute a large proportion of meiobenthic biomass and to mediate the transfer of nutrients between microbes and higher consumers (Majdi and Traunspurger 2015), and mermithid nematodes are common parasites of freshwater invertebrates (Thorp and Covich 2010). The use of nematodes as a biomonitoring tool has existed for some time (Bongers and Ferris 1999), but to our knowledge has not been combined with mainstream assessments of benthic macroinvertebrates. Understanding how anthropogenic stress and parasitism interact to affect the structure and function of aquatic ecosystems is fertile ground for future biomonitoring research (Marcogliese and Pietrock 2011), and field-based research in this area could be facilitated by the expansion of sequence libraries to include more parasitic taxa. Biomonitoring will move from describing the middle of aquatic benthic food webs (primary consumers and larger detritivores) to a more comprehensive picture that includes top predators, rare taxa, and smaller, non-insect components of the benthos. This shift is crucial for diagnostic assessments because the various components of benthic and pelagic food webs respond differently to anthropogenic and environmental stressors (Resh 2008).

The monumental task of distinguishing the effects of anthropogenic stress from natural variation in biodiversity in freshwater ecosystems is difficult because of inadequate sample size or limited capacity to gather information from remote areas and difficult-to-sample habitats. This difficulty is compounded by the scarcity of taxonomic expertise, which results in coarse-resolution taxonomy across a narrow section of the aquatic food web as standard practice. Genomics approaches can address many of these problems, but in the short term must rely upon publicly available reference sequence information for taxon identification. Kvist (2013) estimated that only 15% were represented in a general survey of all animal species represented in public COI databases. Our data show a much better representation at the genus level for a targeted subgroup of animals used in freshwater biomonitoring. Although imperfect and incomplete, the number of North American freshwater invertebrate genera for which COI sequences are publicly available suggests that genomics approaches are ready to play an increased role in freshwater biomonitoring. These approaches already produce levels of taxonomic resolution and breadth that are much better than those produced via routine morphological identification (Stein et al. 2014, Gibson et al. 2015). The next step is to understand how sequence data and their biodiversity inferences must be handled in biomonitoring and assessment models (e.g., Hajibabaei et al. 2016). This step includes how sequence reads relate to organism biomass and relative abundance, the appropriate transformations and dissimilarity measures for use in data analysis, and whether established approaches based on reference conditions (e.g., BEnthic Assessment of SedimenT

[BEAST], Reynoldson et al. 1995; River Invertebrate Prediction and Classification System [RIVPACS], Wright et al. 1998; Australian River Assessment System [AusRivAs], Smith et al. 1999) can capitalize on the large volume of taxonomic information produced by genomics methods. Surveys of library completeness for other barcode markers used for identifying invertebrates (e.g., 16S rRNA; Epp et al. 2012) are also necessary because biodiversity assessment based on genomics methods will ultimately consider markers beyond COI for identification of certain subgroups.

ACKNOWLEDGEMENTS

Author contributions: CJC, JFG, DJB, and MH conceived the study. SS provided data collection and analysis guidance. CJC and JFG collected and analyzed the data and prepared the manuscript.

This work was supported by Environment and Climate Change Canada-funded Natural Sciences and Engineering Research Council Visiting Fellowships to CJC and JFG.

LITERATURE CITED

Baird, D. J., and M. Hajibabaei. 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. Molecular Ecology 21:2039–2044.

Balian, E. V., H. Segers, C. Lévêque, and K. Martens. 2008. An introduction to the Freshwater Animal Diversity Assessment (FADA) project. Hydrobiologia 595:3–8.

Benson, D. A., K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2015. GenBank. Nucleic Acids Research 43:D30–D35.

Blaxter, M., J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, and E. Abebe. 2005. Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 360:1935–1943.

Bongers, T., and H. Ferris. 1999. Nematode community structure as a bioindicator in biomonitoring. Trends in Ecology and Evolution 14:224–228.

Cao, Y., D. P. Larsen, and R. St-J. Thorne. 2001. Rare species in multivariate analysis for bioassessment: some considerations. Journal of the North American Benthological Society 20:144–153.

Carew, M. E., S. J. Nichols, J. Batovska, R. St Clair, N. P. Murphy, M. J. Blacket, and M. E. Shackleton. 2017. A DNA barcode database of Australia's freshwater macroinvertebrate fauna. Marine and Freshwater Research 68:1788–1802.

CESCC (Canadian Endangered Species Conservation Council). 2011. Wild species 2010: the general status of species in Canada. National General Status Working Group, Ottawa, Ontario.

Chariton, A. A., M. Sun, J. Gibson, J. A. Webb, K. M. Y. Leung, C. W. Hickey, and G. C. Hose. 2015. Emergent technologies and analytical approaches for understanding the effects of multiple stressors in aquatic environments. Marine and Freshwater Research 67:414–428.

Compin, A., and R. Céréghino. 2003. Sensitivity of aquatic insect species richness to disturbance in the Adour–Garonne stream system (France). Ecological Indicators 3:135–142.

Courtemanch, D. L. 1996. Commentary on the subsampling procedures used for rapid bioassessments. Journal of the North American Benthological Society 15:381–385.

Curry, C. J., X. Zhou, and D. J. Baird. 2012. Congruence of biodiversity measures among larval dragonflies and caddisflies from three Canadian rivers. Freshwater Biology 57:628–639.

Cushing, C. E., and J. D. Allan. 2001. Streams: their ecology and life. Academic Press, San Diego, California.

Davidson, N. C. 2014. How much wetland has the world lost? Long-term and recent trends in global wetland area. Marine and Freshwater Research 65:935–941.

Ekrem, T., E. Willassen, and E. Stur. 2007. A comprehensive DNA sequence library is essential for identification with DNA barcodes. Molecular Phylogenetics and Evolution 43:530–542.

Environment Canada. 2012. Canadian Aquatic Biomonitoring Network laboratory methods: processing, taxonomy, and quality control of benthic macroinvertebrate samples. Environment Canada, Ottawa, Ontario.

Epp, L. S., S. Boessenkool, E. P. Bellemain, J. Haile, A. Esposito, T. Riaz, C. Erséus, V. I. Gusarov, M. E. Edwards, A. Johnsen, H. K. Stenøien, K. Hassel, H. Kauserud, N. G. Yoccoz, K. A. Brathen, E. Willerslev, P. Taberlet, E. Coissac, and C. Brochmann. 2012. New environmental barcodes for analysing soil DNA: potential for studying past and present ecosystems. Molecular Ecology 21:1821–1833.

Fahner, N. A., S. Shokralla, D. J. Baird, and M. Hajibabaei. 2016. Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA markers. PLoS ONE 11:e0157505.

Ficetola, G. F., C. Miaud, F. Pompanon, and P. Taberlet. 2008. Species detection using environmental DNA from water samples. Biology Letters 4:423–425.

Gibson, J., S. Shokralla, T. M. Porter, I. King, S. van Konynenburg, D. H. Janzen, W. Hallwachs, and M. Hajibabaei. 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proceedings of the National Academy of Sciences of the United States of America 111:8007–8012.

Gibson, J. F., S. Shokralla, C. Curry, D. J. Baird, W. A. Monk, I. King, and M. Hajibabaei. 2015. Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. PLoS ONE 10:e0138432.

Gray, C., D. J. Baird, S. Baumgartner, U. Jacob, G. B. Jenkins, E. J. O'Gorman, X. Lu, A. Ma, M. J. O. Pocock, N. Schuwirth, M. Thompson, and G. Woodward. 2014. Ecological networks: the missing links in biomonitoring science. Journal of Applied Ecology 51:1444–1449.

Hajibabaei, M., D. J. Baird, N. A. Fahner, R. Beiko and G. B. Golding. 2016. A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 371:20150330.

Hajibabaei, M., S. Shokralla, X. Zhou, G. A. C. Singer, and D. J. Baird. 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS ONE 6:e17497.

Jackson, J. K., J. M. Battle, B. P. White, E. M. Pilgrim, E. D. Stein, P. E. Miller, and B. W. Sweeney. 2014. Cryptic biodiversity in streams: a comparison of macroinvertebrate communities based on morphological and DNA barcode identifications. Freshwater Science 33:312–324.

Kvist, S. 2013. Barcoding in the dark? A critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. Molecular Phylogenetics and Evolution 69:39–45.

Kwong, S., A. Srivathsan, and R. Meier. 2012. An update on DNA barcoding: low species coverage and numerous unidentified sequences. Cladistics 28:639–644.

Lou, M., and G. B. Golding. 2012. The effect of sampling from subdivided populations on species identification with DNA barcodes using a Bayesian statistical approach. Molecular Phylogenetics and Evolution 65:765–773.

Majdi, N., and W. Traunspurger. 2015. Free-living nematodes in the freshwater food web: a review. Journal of Nematology 47: 28–44.

Marchant, R. 1989. A subsampler for samples of benthic invertebrates. Bulletin of the Australian Society for Limnology 12:49–52.

Marcogliese, D. J., and M. Pietrock. 2011. Combined effects of parasites and contaminants on animal health: parasites do matter. Trends in Parasitology 27:123–130.

McCafferty, W. P. 1997. Discovery and analysis of the oldest mayflies (Insecta, Ephemeroptera) known from amber. Bulletin de la Société d'Histoire Naturelle de Toulouse 133:77–82.

Merritt, R. W., K. W. Cummins, and M. B. Berg (editors). 2008. An introduction to the aquatic insects of North America. 4th edition. Kendall/Hunt Publishing, Dubuque, Iowa.

Meusnier, I., G. A. C. Singer, J.-F. Landry, D. A. Hickey, P. D. N. Hebert, and M. Hajibabaei. 2008. A universal DNA minibarcode for biodiversity analysis. BMC Genomics 9:214.

Nicacio, G., and L. Juen. 2015. Chironomids as indicators in freshwater ecosystems: an assessment of the literature. Insect Conservation and Diversity 8:393–403.

Orlofske, J. M., and D. J. Baird 2014. The tiny mayfly in the room: implications of size-dependent invertebrate taxonomic identification for biomonitoring data properties. Aquatic Ecology 47:481–494.

Pilgrim, E. M., S. A. Jackson, S. Swenson, I. Turcsanyi, E. Friedman, L. Weigt, and M. J. Bagley. 2011. Incorporation of DNA barcoding into a large-scale biomonitoring program: opportunities and pitfalls. Journal of the North American Benthological Society 30:217–231.

Poff, N. L., J. D. Olden, N. K. M. Vieira, D. S. Finn, M. P. Simmons, and B. C. Kondratieff. 2006. Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships. Journal of the North American Benthological Society 25:730–755.

Porter, T. M., J. F. Gibson, S. Shokralla, D. J. Baird, G. B. Golding, and M. Hajibabaei. 2014. Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome *c* oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. Molecular Ecology Resources 14:929–942.

Ratnasingham, S., and P. D. N. Hebert. 2007. BOLD: the barcode of life data system (www.barcodinglife.org). Molecular Ecology Notes 7:355–364.

Resh, V. H. 2008. Which group is best? Attributes of different biological assemblages used in freshwater biomonitoring programs. Environmental Monitoring and Assessment 138:131–138.

Reynoldson, T. B., R. C. Bailey, K. E. Day, and R. H. Norris. 1995. Biological guidelines for freshwater sediment based on BEnthic Assessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state. Australian Journal of Ecology 20:198–219.

Reynoldson, T. B., D. M. Rosenberg, and V. H. Resh. 2001. Comparison of models predicting invertebrate assemblages for biomonitoring in the Fraser River catchment, British Columbia. Canadian Journal of Fisheries and Aquatic Sciences 58:1395–1410.

Rosenberg, D. M., and V. H. Resh (editors). 1993. Freshwater biomonitoring and benthic macroinvertebrates. Chapman and Hall, New York.

Segers, H. 2008. Global diversity of rotifers in freshwaters. Hydrobiologia 595:49–59.

Shokralla, S., J. F. Gibson, H. Nikbakht, D. H. Janzen, W. Hallwachs, and M. Hajibabaei. 2014. Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. Molecular Ecology Resources 14:892–901.

Shokralla, S., T. M. Porter, J. F. Gibson, R. Dobosz, D. H. Janzen, W. Hallwachs, G. B. Golding, and M. Hajibabaei. 2015. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific Reports 5:9687.

Shokralla, S., J. Spall, J. F. Gibson, and M. Hajibabaei. 2012. Next-generation DNA sequencing technologies for environmental DNA research. Molecular Ecology 21:1794–1805.

Smith, M. J., W. R. Kay, D. H. D. Edward, P. J. Papas, K. S. J. Richardson, J. C. Simpson, A. M. Pinder, D. J. Cale, P. H. J. Horwitz, J. A. Davis, F. H. Yung, R. H. Norris, and S. A. Halse. 1999. AusRivAS: using macroinvertebrates to assess ecological condition of rivers in Western Australia. Freshwater Biology 41:269–282.

Stein, E. D., B. P. White, R. D. Mazor, J. K. Jackson, J. M. Battle, P. E. Miller, E. M. Pilgrim, and B. W. Sweeney. 2014. Does DNA barcoding improve performance of traditional stream bioassessment metrics? Freshwater Science 33:302–311.

Stribling, J. B., K. L. Pavlik, S. M. Holdsworth, and E. W. Leppo. 2008. Data quality, performance, and uncertainty in taxonomic identification for biological assessments. Journal of the North American Benthological Society 27:906–919.

Sweeney, B. W., J. M. Battle, J. K. Jackson, and T. Dapkey. 2011. Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? Journal of the North American Benthological Society 30:195–216.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28:2731–2739.

Thorp, J. H., and A. P. Covich (editors). 2010. Ecology and classification of North American freshwater invertebrates. Academic Press, San Diego, California.

Traunspurger, W. 2000. The biology and ecology of lotic nematodes. Freshwater Biology 44:29–45.

Trebitz, A. S., J. C. Hoffman, G. W. Grant, T. M. Billehus, and E. M. Pilgrim. 2015. Potential for DNA-based identification of

Great Lakes fauna: match and mismatch between taxa inventories and DNA barcode libraries. Scientific Reports 5:12162.

Vinson, M. R., and C. P. Hawkins. 1996. Effects of sampling area and subsampling procedure on comparisons of taxa richness among streams. Journal of the North American Benthological Society 15:392–399.

Vivien, R., S. Wyler, M. Lafont, and J. Pawlowski. 2015. Molecular barcoding of aquatic oligochaetes: implications for biomonitoring. PLoS ONE 10:e0125485.

Walsh, C. J. 1997. A multivariate method for determining optimal subsample size in the analysis of macroinvertebrate samples. Marine and Freshwater Research 48:241–248.

Watson, N. T. N., I. C. Duggan, and I. D. Hogg. 2015. Assessing the diversity of New Zealand freshwater harpacticoid copepods (Crustacea: Copepoda) using mitochondrial DNA (COI) barcodes. New Zealand Journal of Zoology 42:57–67.

Webb, J. M., L. M. Jacobus, D. H. Funk, X. Zhou, B. Kondratieff, C. J. Geraci, R. E. DeWalt, D. J. Baird, R. Barton, I. Phillips, and P. D. N. Hebert. 2012. A DNA barcode library for North American mayflies: progress and prospects. PLoS ONE 7: e38063.

White, B. P., E. Pilgrim, L. M. Boykin, E. D. Stein, and R. D. Mazor. 2014. Comparison of four species-delimitation methods applied to a DNA barcode data set of insect larvae for use in routine bioassessment. Freshwater Science 33:338–348.

Wilson, J. J., R. Rougerie, J. Schonfeld, D. H. Janzen, W. Hallwachs, M. Hajibabaei, I. J. Kitching, J. Haxaire, and P. D. N. Hebert. 2011. When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. BMC Ecology 11:18.

Wright, J. F., M. T. Furse, and D. Moss. 1998. River classification using invertebrates: RIVPACS applications. Aquatic Conservation: Marine and Freshwater Ecosystems 8:617–631.

Zhou, X., S. J. Adamowicz, L. M. Jacobus, R. E. DeWalt, and P. D. N. Hebert. 2009. Towards a comprehensive barcode library for Arctic life—Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. Frontiers in Zoology 6:30.