ASSESSING MACROINVERTEBRATE BIODIVERSITY IN FRESHWATER ECOSYSTEMS: ADVANCES AND CHALLENGES IN DNA-BASED APPROACHES

MICHAEL E. PFRENDER*
Department of Biology, Utah State University, Logan, Utah 84322-5305 USA
E-MAIL: MICHAEL.PFRENDER.1@ND.EDU

CHARLES P. HAWKINS
Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Watershed Science, Utah State University, Logan, Utah 84322-5210 USA
E-MAIL: CHUCK.HAWKINS@USU.EDU

MARK BAGLEY
National Exposure Research Laboratory, U.S. EPA, Cincinnati, Ohio 45268 USA
E-MAIL: BAGLEY.MARK@EPAMAIL.EPA.GOV

GREGORY W. COURTNEY
Department of Entomology, Iowa State University, Ames, Iowa 50011-3222 USA
E-MAIL: GWCOURT@IASTATE.EDU

BRIAN R. CREUTZBURG
Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Watershed Science, Utah State University, Logan, Utah 84322-5210 USA
E-MAIL: BRIAN.CREUTZBURG@AGGIEMAIL.USU.EDU

JOHN H. EPLER
Crawfordville, Florida 323 27 USA
E-MAIL: JOHNEPLER3@COMCAST.NET

STEVE FEND
U.S. Geological Survey, Menlo Park, California 94025 USA
E-MAIL: SVFEND@USGS.GOV

*Current address: Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana 46556 USA

LEONARD C. FERRINGTON, JR.
Department of Entomology, University of Minnesota, Saint Paul, Minnesota 55108-6125 USA
E-MAIL: FERRIO16@UMN.EDU

PAULA L. HARTZELL
Pacific Cooperative Studies Unit, University of Hawaii at Manoa, and the Hawaii Department of Land and Natural Resources, Honolulu, Hawaii USA
E-MAIL: PAULAHARTZELL@HOTMAIL.COM

SUZANNE JACKSON
National Exposure Research Laboratory, U.S. EPA, Cincinnati, Ohio 45268 USA
E-MAIL: JACKSON.SUZANNE@EPAMAIL.EPA.GOV

DAVID P. LARSEN
Pacific States Marine Fisheries Commission, Corvallis, Oregon 97333 USA
E-MAIL: LARSEN.PHIL@EPAMAIL.EPA.GOV

C. ANDRÉ LÉVESQUE
Agriculture and Agri-Food Canada, Ottawa, Ontario, K1A 0C6 Canada
E-MAIL: ANDRE.LEVESQUE@AGR.GC.CA

JOHN C. MORSE
Department of Entomology, Soils, & Plant Sciences, Clemson University, Clemson, South Carolina 29634-0315 USA
E-MAIL: JMORSE@CLEMSON.EDU

MATTHEW J. PETERSEN
Department of Entomology, Iowa State University, Ames, Iowa 50011-3222 USA
E-MAIL: MJP266@CORNELL.EDU

DAVE RUITER
Centennial, Colorado 80121 USA
E-MAIL: DRUITER@MSN.COM

The Quarterly Review of Biology, September 2010, Vol. 85, No. 3

0033-5770/2010/8503-0003$15.00

DAVID SCHINDEL
Consortium for the Barcode of Life, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012 USA
E-MAIL: SCHINDELD@SI.EDU

MICHAEL WHITING
Department of Biology, Brigham Young University, Provo, Utah 84602 USA
E-MAIL: MICHAEL_WHITING@BYU.EDU

KEYWORDS
barcoding, invertebrates, bioassessment, biodiversity, freshwater, next-generation sequencing

ABSTRACT

Assessing the biodiversity of macroinvertebrate fauna in freshwater ecosystems is an essential component of both basic ecological inquiry and applied ecological assessments. Aspects of taxonomic diversity and composition in freshwater communities are widely used to quantify water quality and measure the efficacy of remediation and restoration efforts. The accuracy and precision of biodiversity assessments based on standard morphological identifications are often limited by taxonomic resolution and sample size. Morphologically based identifications are laborious and costly, significantly constraining the sample sizes that can be processed. We suggest that the development of an assay platform based on DNA signatures will increase the precision and ease of quantifying biodiversity in freshwater ecosystems. Advances in this area will be particularly relevant for benthic and planktonic invertebrates, which are often monitored by regulatory agencies. Adopting a genetic assessment platform will alleviate some of the current limitations to biodiversity assessment strategies. We discuss the benefits and challenges associated with DNA-based assessments and the methods that are currently available. As recent advances in microarray and next-generation sequencing technologies will facilitate a transition to DNA-based assessment approaches, future research efforts should focus on methods for data collection, assay platform development, establishing linkages between DNA signatures and well-resolved taxonomies, and bioinformatics.

FRESHWATER BIODIVERSITY ASSESSMENT: BACKGROUND AND SIGNIFICANCE

QUANTIFYING SPECIES composition and richness is fundamental to the study of freshwater ecosystems, but obtaining accurate and precise estimates of these biodiversity metrics is both difficult and costly. We need accurate, precise, rapid, and cost-effective methods to assess the status of local and regional biodiversity and to predict responses to changes in climate, invasive species, and land and water alterations (Sharley et al. 2004; Carew et al. 2003, 2005, 2007a,b; Ball et al. 2005; Pfenninger et al. 2007; Sinclair and Greens 2008). Resource scientists and managers need such information to understand the richness and variability of natural ecosystems, as well as how biological communities respond to both stress (e.g., natural and anthropogenic disturbances) and management (e.g., restoration practices). The implementation of a high-throughput, DNA-based identification system for biodiversity assessment could greatly improve data quality, while reducing both the costs of obtaining data and the time between sample collection and data compilation.

The availability of high-quality biodiversity data is especially critical for effective ecological assessment. Local, state, and federal environmental management agencies throughout the United States and elsewhere use biodiversity data derived from samples of benthic invertebrate assemblages to quantify the ecological condition of aquatic ecosystems (e.g., Rosenberg and Resh 1993; USEPA 2002). These assessments generally use community-level indices based on aspects of taxonomic composition to measure the degree to which biological communities differ from those that would be expected to occur under reference or baseline conditions (Hughes et al. 1986; Reynoldson and Wright 2000; Stoddard et al. 2006; Hawkins et al. 2010). Assessments based on these indices are key to quantifying both the biolog-

ical impacts of pollution and the degree to which management practices are effective in restoring or rehabilitating damaged ecosystems. However, the utility of these indices is strongly influenced by their accuracy and precision—two statistical properties that are markedly affected by how thoroughly a site is sampled and how accurately the biota is identified.

For biological indices to realize their potential, they need to be as accurate and precise as possible—that is, they need to characterize the targeted biological assemblage without bias. Sampling error associated with estimates of the biota present at a site should be small enough to allow unambiguous detection of ecologically significant changes in condition. Recent empirical studies show how inadequate sample counts and coarse taxonomic resolution can independently hinder detection of real biological impacts. First, Cao et al. (2002a,b, 2007) and Cao and Hawkins (2005) showed how the use of small (i.e., 100–300 count) samples produces both imprecision and bias when estimating relative differences in taxa richness and composition among locations, regardless of the taxonomic resolution used in assessments. Second, the use of more highly resolved identifications can reveal the effects of landscape and waterway alteration on freshwater assemblages that are not detected when coarse taxonomy is used (review by Jones 2008). For example, Hawkins et al. (2000) showed that a measure of taxa completeness based on genus/species level identifications detected the effects of watershed alternation on stream invertebrate assemblages in the Sierra Nevada of California, USA, whereas an otherwise similar, family-based measure detected no difference between streams in reference and managed watersheds. Similar results based on a variety of assemblage-level indices and analyses for freshwater invertebrate assemblages have been reported in northeastern France (Guerold 2000); Florida, USA (King and Richardson 2002); New York, USA (Arscott et al. 2006); North Carolina, USA (Hawkins 2006); Northern Territories, Australia (Lamche and Fukuda 2008); and West Virginia, USA (Pond et al. 2008).

The potential costs of drawing incorrect inferences, as either false-negatives or false-positives, from inaccurate or imprecise indices can be staggering. The inability to detect ecological degradation when it actually exists condemns freshwater ecosystems to continued degradation. Incorrect assessments that systems are degraded when they really are not can trigger expensive but unwarranted restoration/remediation, as well as litigation and reduction in public support. Considering that approximately $110 million are spent each year in the U.S. on water quality assessments and that $260 million is viewed as the amount that is actually needed (ASIWPCA 2002), the development of indices that are as accurate and precise as possible, as well as those that are cost-effective and able to be rapidly implemented, should be a national priority.

Recent and rapidly emerging developments in the analysis of genetic material (i.e., diagnostic DNA markers) should provide a fast, cost-effective means of addressing these needs. These DNA-based assay tools have the potential to greatly improve the quality of data collected from freshwater ecosystems, therefore allowing us to better assess and predict the consequences of landscape and waterway alteration on these systems. In this paper, we review the opportunities and challenges associated with moving toward the routine use of DNA markers for the identification of the taxonomic composition of bulk samples of freshwater invertebrates.

Limitations to Current Morphologically-Based Bioassessments

data quality and cost

The quality of biological surveys depends on the degree to which field samples accurately and precisely characterize the biota at sites. Data quality largely depends on two sample properties: (1) how well the collected sample represents the biota inhabiting the targeted site (i.e., the quality of the site scale design), and (2) how well the collected sample(s) are evaluated (i.e., quality of sample

processing). The first property is a function of both the mix of habitats sampled and the number of individuals collected. The second property is dependent on the taxonomic resolution and the specimen misidentification rate. Selecting a coarse taxonomic resolution can obscure responses of one or more finer-level taxa (e.g., species) to environmental alteration (Hawkins et al. 2000; Jones 2008). Because processing samples of benthic invertebrates is time consuming, typically less than 1 m^2 of stream or lake bottom is sampled, and only 100–500 individuals from the sample are usually identified. This small number of individuals is then used to characterize the entire benthic assemblage—i.e., millions of individuals—at the site. Both assessment accuracy and precision improve when the area sampled or the number of individuals included in such a sample can be increased (Cao et al. 2002a, b; Lorenz et al. 2004; Ostermiller and Hawkins 2004; Cao and Hawkins 2005; Clarke et al. 2006; Nichols et al. 2006) (Figure 1), yet small subsamples continue to be used because of the unacceptable costs associated with processing larger samples (Carter and Resh 2001). Variability in descriptions of freshwater benthic invertebrate assemblages is considerably influenced by the type of water body as well as with the specific biological metric examined. For example, the percentage of total variance (i.e., across sites and among replicate samples) in metric values that was associated with sampling error (within-stream replicates) ranged from 0–99% across all combinations of 27 metrics and 19 types of European streams, and averaged 3–28% among the 27 different biological metrics (Clarke et al. 2006). This source of error could nearly be eliminated for many metrics if more extensive areas of stream could be sampled, thus resulting in the identification of a larger number of invertebrates.

While significant limitations can be overcome through increased sample size, data quality is further compromised by the fact that identifications are generally made at a level of taxonomic resolution above the species level (e.g., genus, family, or higher taxonomic levels), and both the consistency and the accuracy of identifications can vary greatly across laboratories. The use of coarse taxonomic resolution in bioassessments can blur species-specific signals and the sensitivity of assessments. This lack of sensitivity ultimately limits our ability to detect effects of either adding or removing stressors (Lenat and Resh 2001; Schmidt-Kloiber and Nijboer 2004; Arscott et al. 2006; Hawkins 2006). Unfortunately, one reason we are forced to use these coarse levels of taxonomic resolution is that most of the individuals in benthic samples are juveniles that cannot be identified to species based on their morphological traits. Juveniles, and specimens that have been damaged beyond recognition, may make up the bulk of a sample, resulting in poor and potentially misleading interpretations of species composition.

Inconsistencies among labs and individuals in identification skills will produce data of variable quality. For example, in the recent national assessment of wadeable streams and rivers in the USA (USEPA 2006), identification errors ranged between 8% and 30% (mean = 21%) across eight labs (Stribling et al. 2008). Also, some taxa are more cryptic to species identification than others. For instance, larvae of the ubiquitous Chironomidae (midges) are notoriously difficult to identify; Epler (2001) reported a 6–60% misidentification rate among these organisms. Such variability in data quality will cause variation in assessments as well as in assessment quality at specific sites. This variability compromises our ability to combine data sets for regional assessments, and, in order to use data from multiple labs or individuals for such assessments, it would first be necessary to post-process all samples to a common level of taxonomic resolution, which typically translates to the lowest quality data in the data sets of interest.

TURN-AROUND TIME

The time required to conduct a biological assessment is largely a function of the time it takes to identify the taxa collected in a sample. For assessments based on benthic macroinvertebrates, the set of taxa used most widely across the U.S. and elsewhere (USEPA 2002), the turn-around time can

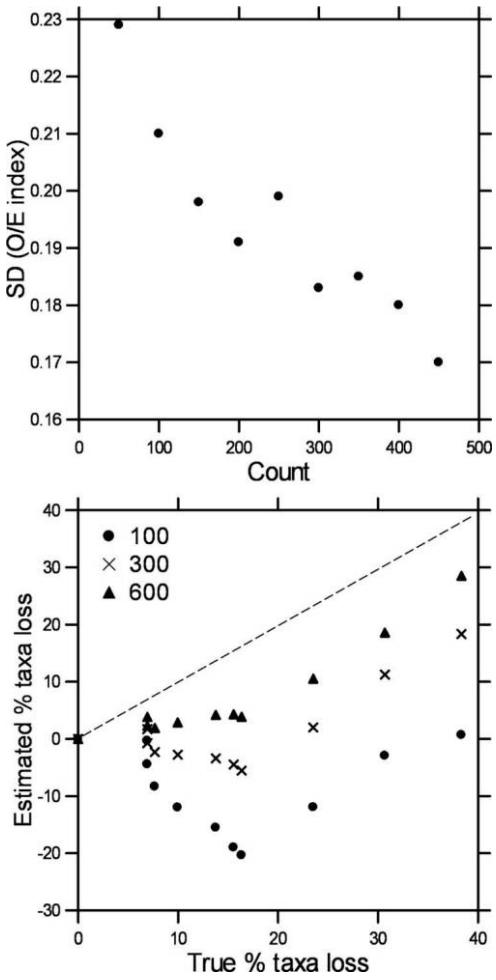

Figure 1. Effect of Sample Size on Bioassessment

Examples of how both assessment precision (top panel) and accuracy (bottom panel) are affected by the size of fixed-count samples. The graph in the top panel is modified from data presented in Ostermiller and Hawkins (2004) and illustrates how the precision of an O/E index used in bioassessment is affected by the sample counts used to calibrate models that predict the number of taxa expected to occur at a site (E). There is a strong relationship (r^2=0.84) between the precision of the O/E index (i.e., the standard deviation of O/E values observed at reference quality sites) and the sample count. This relationship implies that samples would need to contain about 2,000 individuals (SD = 0.224 – 0.000119*count) in order to achieve perfect precision. In reality, the linear relationship observed here would likely become asymptotic (concave up) at higher sample counts, thus requiring even higher counts to achieve high precision. The graph in the bottom panel is modified from data presented by Cao and Hawkins (2005) and illustrates how the numbers of taxa lost with increasing simulated stress are underestimated when small fixed-count samples are used in biological assessments. Each data point is a mean value derived from 11 fixed-count resamplings of the assemblage that resulted following the application of 9 increasingly severe levels of stress. The dashed line represents a 1:1 correspondence between estimated and true taxa loss. The magnitude of difference between estimated and true taxa loss increases with decreasing sample size, and estimates can often imply that assemblages are either not losing or are even gaining taxa (negative values of estimated taxa loss) when taxa loss is actually occurring.

often be several months. Although there are many labs around the world that specialize in the identification of freshwater benthic invertebrates, their throughput is affected by the timing of sample delivery, the time it takes to process an individual sample, and by economies of scale. For example, samples are typically collected in the field from late spring to early autumn, and accumulated batches of samples are sent to contracted labs. These samples are then typically processed in the order that they are received. Because individual samples can take 2–8 hours to process, many samples will sit for several weeks or months before they are processed. Reduction in sample processing time would improve our capability to act promptly in response to environmental degradation.

Improving the reliability and defensibility of general survey data, as well as those data used in bioassessments, will require larger samples and more accurate and refined identification of the taxa contained therein. More rapid turnaround in the time it takes to convert field samples into usable data would greatly enhance the use of biological information for management purposes. The present constraints could be minimized if we could rapidly assess the identities of all species in a large, bulk sample of invertebrates, and the use of diagnostic DNA markers, as we discuss below, has great potential to provide us with that ability.

Genetic Approaches to Bioassessment

A number of DNA-based assay platforms are currently available and are being used in increasingly diverse ecological contexts (Thomas and Klaper 2004). Most notably, genetic characterization of prokaryotic communities through the application of next-generation DNA sequencing and DNA microarrays is becoming commonplace (DeSantis et al. 2005; Brodie et al. 2007; He et al. 2007; Dinsdale et al. 2008; Zhou et al. 2008). A similar genetic characterization and measurement of diversity in eukaryotic communities is now emerging (e.g., Creer et al. 2010). The same basic approaches used in microbial communities are applicable to eukaryotic communities and should prove to be particularly useful in benthic invertebrate biodiversity assessment. These technical approaches differ substantially in their data requirements, application, and utility for high-throughput assays. The first set of these approaches, including PCR-based fragment analysis and the use of micro- and macroarrays, requires considerable up-front DNA data infrastructure to design assay tools that target a predefined and well-characterized biota. These techniques rely on *a priori* knowledge to design effective PCR primers for DNA amplification and/or hybridization probes on arrays.

An alternative approach, using next-generation sequencing strategies (Marguilis et al. 2005), requires minimal initial development, but relies heavily on computationally intensive data processing of DNA fragment (sequence) data to detect the occurrence of unique sequences of DNA that are assumed to represent different taxa. These inferred taxa form operational taxonomic units, which may or may not be linked to an established taxonomic framework. Ultimately, as emphasized below, establishing the link between DNA-level data and ecological assessments based on traditional taxonomy is a necessary component of any of these efforts.

In the following sections, we provide a brief overview of genetic techniques that have been used in biodiversity assessment as well as emerging techniques amenable to high-throughput assessment, and we discuss the data infrastructure required to employ these techniques. Our goal is to highlight some of the currently used and most promising avenues for high-throughput systems, not to provide a comprehensive explanation of all possible techniques. In particular, we emphasize the limitations and *a priori* data requirements for these diverse approaches. For a brief description of a wide variety of genetic techniques suitable for freshwater bioassessment, see Box 1.

PCR-based fragment and sequence analysis

Initial DNA-based applications for biodiversity assessment implemented PCR-based techniques targeting defined DNA regions in the mitochondrial or nuclear genomes in specific taxa. One widely used approach is

Box 1

Techniques and approaches to genetic assessment of biodiversity. This box provides some short definitions of the genetic techniques referred to in the text, as well as a brief overview of the techniques that have been applied to bioassessment and of techniques that may become useful in this rapidly growing field.

Polymerase Chain Reaction (PCR): A variety of genetic assay techniques rely on the use of PCR amplification to target genetic variation in defined fragments of nuclear or mitochondrial genomic DNA.

PCR-RFLP: In this widely used approach, a predefined gene region is amplified with DNA primers, and the amplified DNA segment is fragmented with restriction enzymes (REs). REs have specific nucleotide sequence recognition sites and cut the DNA at these sites. Variation in the sequence at these recognition sites in the amplified DNA fragment yields different size fragments among taxa, hence the name restriction fragment length polymorphism (RFLP). This technique involves sets of DNA primers that bind to and amplify the target fragments in all the study taxa. Products are visualized with an agarose gel, or the amplified PCR products are labeled with fluorescent dyes and visualized in a capillary DNA sequencing machine. Banding patterns will differ if communities are different. Sequencing of the bands—something that is quite technically challenging—is required in order to identify the species affected.

Real-Time PCR: A modification of standard PCR amplification that monitors the amount of product formed in the reaction at each cycle. The rate of increase is related to the initial concentration of the template DNA in the PCR. This technique is useful for detecting low abundance DNA and can be used to quantify relative abundance at various taxonomic levels, depending on the specificity of the primers and fluorescent probes. Multiplexing, i.e., the ability to detect multiple species, is limited with the current technology.

Barcoding: The largest coordinated effort of the CBOL is devoted to characterizing the DNA sequence at predefined, highly informative gene regions. The most commonly used barcoding region is the cytochrome oxidase I (COI) gene in the mitochondrial genome. Unique nucleotide sequences are a "barcode" that can be linked to traditional taxonomic designations and used for identification. This approach uses PCR and DNA sequencing.

Universal and specific PCR primers: The utility of all the PCR based techniques is contingent on the performance of the PCR primers. In some applications (e.g., Barcoding), it is highly advantageous to have universal primers that amplify any clean specimen, but, in order to detect an important species directly in the environment, primers need to be specific at the species level (see Real-time PCR above).

PCR restriction fragment length polymorphism (PCR-RFLP). PCR-RFLP uses a combination of the PCR amplification of a predefined polymorphic DNA region and the subsequent fragmentation of the amplified DNA by restriction enzyme digestion. Characteristic patterns of fragment lengths produced by the restriction enzymes occur because of nucleotide variation in the DNA sequences that alter the number of restriction enzyme recognition sites. This technique is relatively inexpensive, requires only basic and widely available molecular laboratory equipment, and has the potential for high-throughput. Start-up information is minimal, simply requiring sets of PCR primers that will amplify the target DNA region in all the relevant taxa, as well as polymorphic

BOX 1
Continued

Micro- and macroarrays: Unique single-stranded DNA oligonucleotide fragments are bound to a substrate (e.g., nylon membrane, glass slides or computer chip) in sets of distinct spots. Many small spots make microarrays, while fewer large spots produce macroarrays. Sample DNA is labeled (e.g., with fluorescent dyes) and hybridized to the arrays. Spots with hybridized DNA fluoresce and indicate the presence of complementary DNA in the sample. This technique is commonly used for expression profiling to survey the activity of genes, and has been used in a number of studies to detect the presence of species DNA markers in an environmental sample. Often arrays are coupled with PCR to enrich a sample for specific DNA targets prior to hybridization. The major advantage is the ability to survey many different DNA variants simultaneously. The disadvantages include the relatively high cost when compared to RT PCR and the inability to detect unknown sequences in a sample when compared to sequencing techniques.

Illumina Inc., bead-based arrays: This platform utilizes a combination of a marker-specific nested PCR and a novel hybridization approach to survey polymorphism at a large number of DNA sites. Originally, this technique was designed to assay single nucleotide polymorphisms (SNPs) in human genomes. Bead-based arrays can assay up to 1500 polymorphic sites in a single assay and can be scaled up to run 96-well plates, thus giving it high-throughput potential.

Next-generation DNA sequencing: Unlike the PCR and hybridization approaches that require DNA data for design, direct next-generation sequencing can be done anonymously on virtually any sample. The sequence data are collected in Mb quantities, and then unique variants are filtered out bioinformatically. There are three leading platforms, each with a different sequencing methodology and different strengths and advantages. The Roche GS FLX generates ~500 Mb per run, with read lengths of 400–500 bp. Illumina Inc. has partnered with Solexa to provide a Genome Analyzer System producing ~2.3 Gb per run, with read lengths of 70 bp. Applied Biosystems Inc. provides a relatively new addition to next-generation sequencing with their SOLiD System 2.0. This platform produces an impressive 5 Gb of data per run, with read lengths of 35 bp. The limitation to all these systems is the current high cost per run. However, it is becoming practical to combine multiple uniquely tagged samples in a single run. Given the rapid acceleration of these technologies and the increasingly lower costs, next-generation sequencing holds tremendous promise for future application in biodiversity assessment.

Flow cytometry: In a clever combination of PCR, DNA hybridization, and flow cytometry, Diaz et al. (2006) developed a fungal identification system. This approach could be tailored to high-throughput. The remaining challenge is to establish the upper bounds of unique variants that can be detected in a sample.

nucleotides in DNA sequences that fall within the recognition sites of the restriction enzymes. Given a set of known sequences, it is possible to define *a priori* the number of unique sequence variants that can be identified by an optimized set of restriction enzymes. A number of taxonomic groups, including benthic invertebrate assemblages, have been characterized through PCR-RFLP (e.g., Carew et al. 2003). The major disadvantage to the PCR-RFLP approach is its limited ability to screen polymorphic sites in the target DNA region. Because variation in DNA fragment size is the result of nucleotide

variation in restriction enzyme recognition sites, this approach is, in practice, only assaying variation at a small fraction of the nucleotide sites in a DNA sequence. Much of the potentially informative variation in DNA sequences is cryptic to PCR-RFLP, thus reducing the information content and making it difficult to identify novel sequence variants. In particular, distinguishing among closely related taxa can be problematic, as diagnostic nucleotide variation may not coincide with restriction sites, thereby limiting this approach in fine-scale taxonomic resolution.

One approach to circumventing these data limitations is PCR amplification and DNA sequencing of the target region. Direct DNA sequencing reveals all polymorphic sites and, thus, yields greater information. More data result in greater power to discriminate among taxa, but this gain in resolution comes at a greater cost in both time and expense. Samples need to be processed one individual at a time, and the costs of processing samples and sequencing DNA greatly limit the goal of a gain in sample size. An additional complication shared by these approaches arises when dealing with communities comprised of highly divergent taxa. As taxonomic distance increases, it becomes progressively more difficult to design "universal" PCR primers. For example, capturing the taxonomic diversity typically present in assemblages of benthic invertebrates would likely require the development of multiple sets of PCR primers, each targeting the same DNA region in different sets of species.

A direct application of a PCR-DNA sequencing approach is illustrated by the substantial and coordinated efforts of the Barcode of Life Initiative (http://www.dnabarcodes.org). The ambitious aim of this group is to characterize DNA sequence variation in all the major eukaryotic groups based on predefined gene region(s), and to link that variation to traditional taxonomic identity. Sequence variation in a diagnostic ~650 base pair portion of the mitochondrial gene cytochrome c oxidase I (COI) is the target for eukaryotic identification. The barcoding approach based on COI as well as other DNA regions, such as intergenic spacers, has been highly successful in a wide variety of taxonomic groups, including terrestrial (Hebert et al. 2004; Barrett and Hebert 2005) and aquatic taxa (Neigel et al. 2007), and has been applied to a number of pressing ecological issues (e.g., invasive species [Armstrong and Ball 2005; Harvey et al. 2009]). Sequence variation in the COI region has been used effectively in a number of studies to characterize benthic invertebrate diversity (Sharley et al. 2004; Carew et al. 2005, 2007a,b).

DNA hybridization-based approaches

Given a comprehensive database of diagnostic DNA markers, a spectrum of currently available technologies is amenable to the development of genetic assay tools; we will expand on the development of such a database and the relationship between DNA markers and taxonomy below. These technologies range from oligonucleotide microarray platforms to single nucleotide polymorphism (SNP) bead-based arrays. Ultimately, the choice of platforms will depend on the practical consideration of the number of taxa included, as well as per sample costs. A well-characterized library of diagnostic DNA markers for species allows the development of genetic approaches based on DNA-DNA hybridization techniques as a reasonable alternative to PCR and DNA sequencing. DNA hybridization takes advantage of the complementary base-paired structure of double-stranded DNA. In these hybridization techniques, a unique single strand of DNA is bound to a membrane, glass slide, or bead. These platforms of bound DNA can contain a few unique sequences in a macroarray, or can be constructed with a high density of many sequence variants into microarrays. DNA extracted from an environmental sample is labeled with fluorescent dye and washed over the DNA array. Complementary sequences in the sample hybridize to the DNA captured on the array, and the presence or absence of any particular sequence can be distinguished by the intensity of the fluorescence. The major potential gain of an array-based platform is the ability to batch-process whole community-level samples. This capability alleviates the laborious one-by-one approach currently used for morphological identifi-

cation and required for standard PCR and DNA sequencing.

Microarrays have been commonly used as a tool for functional studies of gene expression (Gracey and Cossins 2003; Stoughton 2005) and for detection of single nucleotide polymorphisms (SNPs) in single taxa (Comai et al. 2004; Gilchrist et al. 2006; Gresham et al. 2006), but applications to biodiversity assessment are just beginning to appear in the literature. The taxonomic groups that have thus far received the most attention are prokaryotic communities (Gentry et al. 2006; Wagner et al. 2007). Microarrays based on sequence variation in the 16S ribosomal RNA provide broad coverage of prokaryotic lineages and have been used to categorize bacterial diversity in environmental samples (Castiglioni et al. 2002; Call et al. 2003; Loy et al. 2005; Lozupone and Knight 2007). In these studies, wide taxonomic coverage is achieved by using sets of "universal" PCR primers to amplify the 16S rRNA, and the amplification products are then hybridized to arrays containing taxon-specific DNA fragments. Through the incorporation of DNA sequences of genes found within common metabolic pathways, the design of prokaryotic arrays and the objectives of these studies have recently shifted to the quantification of functional diversity in prokaryotic communities (Dinsdale et al. 2008). Interestingly, this functional view of prokaryotic diversity is, in turn, causing a shift in the emphasis of studies on biodiversity in prokaryotic communities away from a description of the abundance and distribution of unique lineages defined by characteristic 16S rRNA sequences, and instead toward the description of the abundance and distribution of genes and metabolic pathways in communities (Dinsdale et al. 2008). In eukaryotic taxa, array-based approaches have been applied to diversity studies of fungal communities (Lévesque at al. 1998; Siefert and Lévesque 2004; Tambong et al. 2006) and mammalian taxa (Pfunder et al. 2004), and have been used as forensic tools for detecting endangered vertebrates (Teletchea et al. 2008). A barrier to the wider application of arrays to eukaryotic biodiversity assessment is the lack of DNA sequence data available to design arrays in target taxonomic groups.

The greatest advantages to a microarray approach will be gained in contexts where a large number of species are present within or among the target communities, and when there is a need to process a high volume of samples. The initial investment in array design and the relatively high per-array cost make this approach an impractical option for assessments targeting a relatively small number of taxa, or in situations where the number of samples is small enough that PCR or sequencing strategies can more easily be employed. However, large increases in the number of individuals in an environmental sample require a substantial increase in labor and cost when using PCR-based strategies such as PCR-RFLP or DNA sequencing. Arrays, in contrast, may contain unique DNA signatures from a large variety of taxa that can be assayed simultaneously. Currently available high-density arrays may contain several hundred thousand unique DNA fragments. Also, strategies to overcome the high per sample cost of array processing are now available. An example is the bead-based array, produced by Illumina, Inc., which uses a nested PCR approach. This array can survey up to 1500 unique polymorphisms and can be scaled to a 96-well format, thereby allowing the simultaneous processing of multiple samples. Membrane-based arrays can also be stripped and reused multiple times, (Fessehaie et al 2003; Tambong et al. 2006) and are amenable to spotting with microarrayers at near microarray density (Chen et al. 2009).

A number of technical challenges in the design of arrays for biodiversity assessment need to be addressed. What is the optimal length of DNA probes needed to reduce nonspecific hybridization that may generate a false positive signal? How much redundancy in the number of probes should be incorporated into a microarray for the assessment of highly diverse and complex communities of eukaryotes? How many gene regions will be needed to identify a given number of target taxa? It is possible that a single gene or few genes (e.g., the COI barcoding region) and a small number of unique probes may be suitable for array design. In a modeling

exercise, Hajibabaei et al. (2007) used the design specification for an oligonucleotide-based array consisting of short 25mer probes, as well as the information content in the COI and cytochrome b (cytb) genes, to gauge the feasibility of constructing a mammalian identification array *in silico*. By designing three unique probes for each species, they could unambiguously identify more than 90% of the species in the original data set, based upon the level of sequence variation observed in either of the two DNA fragments.

The work of Hajibabaei et al. (2007) and others (e.g., Zahariev et al. 2009) are important first steps in examining these critical design issues. Both bioinformatics and empirical testing are still required to determine the potential accuracy of such a single gene array and the extent to which multiple gene regions would be required in order to avoid cross hybridization of closely related taxa, as well as false positives. Inevitably, as the number of taxa incorporated in an array increases, the scope of these problems becomes increasingly more complex. In addition to the refinement of array design for whole communities, more powerful analytic approaches are needed that maximize the information content from multiple DNA probes in order to identify closely related species (Engelmann et al. 2009). In future applications, specifically designed arrays could be used as tools for developing characteristic DNA signatures (Cannon et al. 2006), similar to the way in which community typing is currently used for microbial communities through RFLP techniques.

NEXT-GENERATION SEQUENCING STRATEGIES

The approaches that we have discussed thus far require that substantial DNA sequence information be linked to the particular taxa that would be sampled in a survey. This requirement is true for both PCR-based and hybridization platforms, as well as for post-data collection processing to generate measures of biodiversity. An alternative strategy is to apply next-generation DNA sequencing to whole community DNA extractions, or to sequence amplified DNA from the products of universal primers applied to whole community DNA. Prokaryotic metagenomic projects based on next-generation DNA sequencing generally take this approach (Angly et al. 2006; Dinsdale et al. 2008). The two major advantages to a next-generation sequencing approach are that it is possible to generate large amounts of DNA sequence information from environmental samples, and that little upfront development is required. For example, a single run of a 454-Roche Inc. machine generates in excess of 500 million bases of sequence. Other platforms (Illumina Inc. and ABIs SOLiD) generate substantially more bases in total, but shorter individual sequences. Applying a tagging strategy to uniquely identify samples would allow the combination of multiple whole community assays in a single run, thereby reducing the per sample cost (Meyer et al. 2007, 2008; Parameswaran et al. 2007).

Since the use of next-generation sequencing for whole community diversity assessment in eukaryotic biotas is still in its earliest stages, there are relevant issues to be addressed. One important factor will be to establish the lower detection limit with regard to the numerical abundance of individuals in rare taxa (i.e., the relative contribution of DNA to a sample). There are also pitfalls in the commonly used strategy of PCR amplification and in the sequencing of target genes. The lack of truly universal PCR primers allows for potential taxon-specific amplification bias due to primer binding inefficiencies. It has also been suggested that PCR amplification can introduce sequence variants that are the result of errors in the PCR and sequencing process and that are not reflective of true variation in the original sample. For example, as many as 16% of the sequences in a benthic sample showing chimeric sequences have been noted (Porazinska et al. 2009). These errors, which could inflate the estimates of sequence diversity in a sample and make detection of unknown taxa challenging, can in large part be overcome with a comprehensive reference database. A remaining challenge is to develop approaches that extend next-generation sequencing beyond presence-absence deter-

mination to a quantitative assessment of taxonomic diversity. Along these lines, controlled and replicated experimental tests using nematode communities with differing abundance among taxa have been promising. The high level of repeatability in sequence coverage among replicates suggests that a qualitative assessment may eventually be achievable (Porazinska et al. 2010); however, achieving this goal in a benthic invertebrate community will certainly be a complicated endeavor, given the dramatic size differences among the most commonly observed species (see below).

A bioinformatics approach can be used on these data to bin or cluster the sequences into unique variants at particular genes—for instance, using the partial COI locus as in standard barcoding. This approach, in which the variation in the sequences informs the assessment of taxonomic diversity, has been referred to as reverse taxonomy (Markmann and Tautz 2005). In principle, biodiversity metrics could be based completely upon the clustering of taxonomically anonymous DNA sequences, with sequence divergence criteria used to assign groups of similar sequences to molecular operational taxonomic units (MOTUs) (Floyd et al. 2002; Blaxter et al. 2005). However, caution should be used in inferring the validity of taxonomic assessments based on DNA markers without first making a significant effort to establish the relationship between the markers and taxonomy. Phenetic clustering of DNA sequences into MOTUs ignores the detailed taxonomic frameworks available for many taxa that incorporate an evolutionary phylogenetic perspective that cannot be readily matched through a single-gene molecular genetic data set. A more powerful approach would be to compare DNA data from next-generation sequencing to a well-vetted reference database that relates sequence variants to formally described taxa. Here again, we emphasize that it is necessary for a comprehensive DNA sequence library to be a prominent component of a mature, DNA-based biodiversity assessment tool. Linking DNA sequence data to established taxonomic classifications remains a significant but essential challenge in maximizing the utility of DNA-based assessment strategies. The current lack of a fully linked DNA sequence database and taxonomy highlights the value of applying next-generation sequencing approaches for the development of a comprehensive characterization of the genetic diversity found in target biotas. Aquatic invertebrate biodiversity assessment efforts can provide the essential data with which to synergistically focus taxonomists in their efforts to resolve areas of ambiguity.

CHALLENGES

The practical realization of a rapid DNA-based method for assessing the biodiversity of freshwater and other ecosystems will require a focused and coordinated research effort if it is to overcome a number of technical and conceptual challenges (Figure 2). To realize the potential of this methodology we must address three primary issues:

1. The development and management of the primary data on which DNA-based surveys will depend, including establishing the relationship between a well-resolved taxonomic framework and diagnostic DNA signatures
2. The development and validation of the technical methods that can most efficiently, accurately, and cost-effectively produce the DNA-sequences used to identify taxa
3. The development of the bioinformatics necessary to store and efficiently translate DNA data into useable information, and provide public access to these data.

DATA NEEDS

The success of DNA-based surveys will ultimately depend on the development of a database that relates DNA sequence information to an accepted and usable taxonomy—even if that taxonomy is, in the short term, partially based on MOTUs. This work has already been started via the Consortium for the Barcode of Life (CBOL). CBOL focuses on producing barcodes for species within different groups of taxa. The proposed work would build on the CBOL model but would focus on establishing sequences for multiple taxonomic groups within a single type of ecosystem. Given the increased ease of collect-

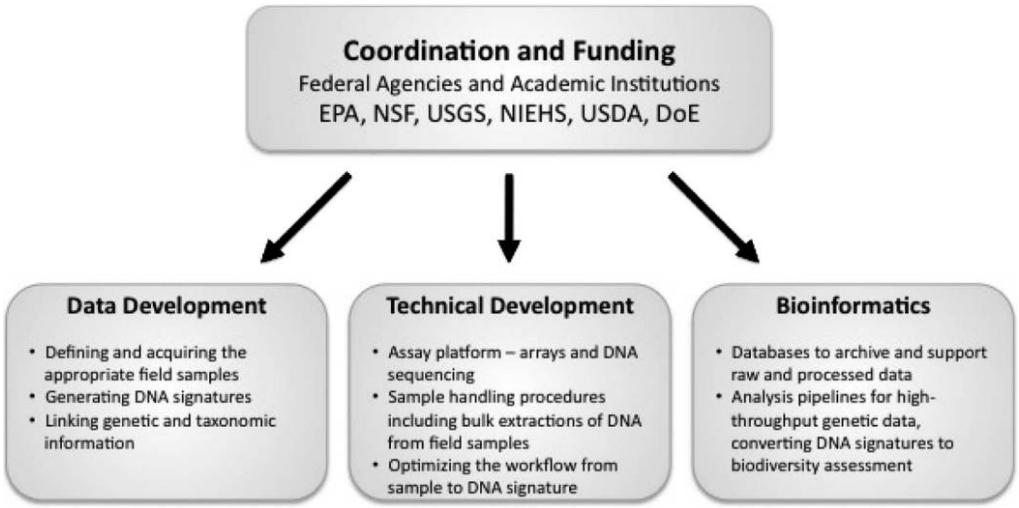

Figure 2. Coordinated Research to Achieve Effective Genetic-Based Bioassessment

Schematic overview of the components required for a coordinated research program to develop a genetic biodiversity assessment tool for North American benthic invertebrates. The critical elements are shown in boxes. Oversight and coordination among government and academic research labs is at the top, along with substantial funding from U. S. agencies whose mission includes direct involvement in the development and utilization of genetic tools. Three interdependent areas of research priority include: (1) sampling in applicable freshwater ecosystems and the generation of genetic DNA signature data, (2) development of protocols for DNA extraction and genetic assay platforms, and (3) bioinformatics infrastructure to process and archive data.

ing genetic data, taxonomically informative markers (e.g., COI gene) will inevitably be supplemented by a set of genes that have characteristic transcriptional patterns and that are bioindicators of environmental quality, similar to the functional approach of microbial community assays, thus allowing for a single assay or set of assays that will characterize both biodiversity and organismal function in a particular environmental context.

The efficacy of a DNA-based biodiversity assay depends on the identification of unique sequence variants for all of the target taxa; therefore, the development of a database of unique DNA sequence variants suitable for taxonomic identification is the primary challenge faced in advancing this research agenda. The scope of the task is defined by the invertebrate diversity in North American freshwaters and the requirements for the accurate biodiversity assessment of these fauna. The development of an assessment tool applicable to bodies of freshwater across North America could potentially require a database including all of the ~15,000 freshwater invertebrate species (Thorp and Covich 2001), but, in reality, only a fraction of this fauna needs to be characterized in order to construct an effective DNA-based assessment tool. The benthic invertebrates that are typically considered in current assessment programs and that are representative of the major focal taxonomic groups (e.g., arthropods, annelids, and mollusks) comprise 2000–3000 species. Given current next-generation sequencing approaches, generating the genetic data necessary to characterize these taxa and subsequently developing a DNA-based assay tool that would be applicable in the majority of North American aquatic ecosystems is a readily achievable goal. To gauge the scale of this effort, it is important to realize that population-level genetic variation must be considered. All target taxa must be characterized by multiple DNA sequences covering the range of genetic diversity found in the natural populations that will be biosurveyed. Significant inroads towards a genetic characterization of this freshwater

fauna have already been made. For example, the Barcode of Life Project (http://www.barcodinglife.org/) has started to develop this type of genetic database for species-level identifications, and the Barcoding approach is being applied with great success to a growing range of taxa, including benthic invertebrates (Ratnasingham et al. 2007).

Two important issues to address include: (1) how to prioritize the sequencing among various groups of freshwater taxa, and (2) what sequence information should be catalogued. Although, ultimately, sequence information on all freshwater species should be collected, the species-level taxonomy of some groups is much better understood than that of others. For example, most species of Ephemeroptera, Plecoptera, Trichoptera, and Coleoptera from North American freshwater ecosystems have been described (Balian et al. 2008), whereas Dipteran species—especially hyperdiverse groups such as the Chironomidae (Cranston 1995)—are poorly known. This gap in our knowledge base also extends to the association of life stages; for instance, within the Tipuloidea (Diptera), less than 4% of the 15,000 species described have immature life stages associated with the adult life stage. We believe it would be most prudent to establish sequences for species in the better known groups first, while still working to develop strategies to better characterize the taxonomic diversity in the less well-known groups. To date, a ~650 bp sequence of the mitochondrial COI gene has been viewed as a standard DNA barcode. However, a single gene region may not be sufficient for the identification of all taxa in a community sample. For example, in an assessment of nematode diversity using next-generation sequencing of a pool of known taxonomic composition, Porazinska et al. (2009) found that analysis based on a single gene sequence (small and large subunit rRNA sequences) underestimated the number of species. Using both sequences, the detection ability was increased from ~90% to 95%. Unambiguous identifications based on sequence data may well require information from more than one DNA region to completely resolve the diversity in a community sample.

INTEGRATING DNA INFORMATION AND TAXONOMY

The relationship between diagnostic DNA markers and systematics has been more than a little contentious over the past few years. This conflict is illustrated by a number of issues with barcoding. As with any biodiversity assessment approach, barcoding has acknowledged limitations. For example, the general utility of a barcoding strategy as a characterization of species level biodiversity has been questioned because, in some groups, the level of nucleotide polymorphism within species is comparable to the level of divergence among species (Meier et al. 2006; Skevington et al. 2007). Moreover, current DNA barcoding methods do not distinguish between the true mitochondrial target markers and nuclear mitochondrial pseudogenes (numts)—portions of the mitochondrial genome that have been incorporated into the nuclear genome over the evolutionary history of a group. The number of species can be over- estimated by 100% or more when numts are coamplified with the target barcode region (Song et al. 2008). The distribution of numts throughout arthropod and other invertebrate taxa has not yet been explored, although they appear to be widespread in grasshoppers and crayfish, and are likely present in other aquatic insects such as Plecoptera and Ephemeroptera as well (M. Whiting, unpublished data).

The relative level of within- and among-species variation has a potentially large impact on the link between DNA variation and taxonomic classification. Although the utility of barcoding across taxonomic groups is an ongoing empirical issue, the important outcome of these efforts is the rapid expansion of a DNA database of characteristic sequences associated with specific taxonomic groups, including knowledge of intra/interspecific variation among these groups. For many of the candidate genetic platforms for benthic invertebrate biodiversity assessment, the existence of these data is absolutely critical.

The relationship between DNA signatures and taxonomic description, as well as the dependence of one upon the other, is not

entirely clear. There is an ongoing and extensive dialogue in this area (Cognato and Caesar 2006); some advocates question the utility of DNA signatures without linkage to a formal taxonomic assessment (e.g., Will and Rubinoff 2004), while others point out the lack of established taxonomy for many groups of organisms and question the utility of groupings based solely on DNA sequences (e.g., Blaxter 2004). Advances in this area will require both an extensive research commitment to working through rigorous alpha-level taxonomy for benthic invertebrates, as well as an exploration of bioinformatic approaches to link DNA data to this taxonomic framework (Bertolazzi et al. 2009).

We take a pragmatic view of this issue, acknowledging that formal taxonomy plays a critical role in biodiversity assessment, but also realizing that major advances can be quickly gained in the absence of a well-resolved taxonomy linked to DNA sequence variants. In a real sense, these two endeavors are complementary, with DNA data informing both gaps and problematic areas in taxonomy, and should be applied in a synergistic and iterative fashion (Carew et al. 2005; Caesar et al. 2006). In practice, benthic invertebrate diversity estimates generally rely on the classification of juvenile individuals or of partial/damaged specimens in a sample that are notoriously difficult to assign to species-level taxonomy due to a lack of informative characters. To further complicate the issue, the taxonomy in many of these groups is largely based on adult morphology, often with no established link to juvenile forms. Here, diagnostic DNA markers serve the dual purpose of providing unambiguous identification of juvenile forms, as well as providing the critical data necessary to establish the link between immature and adult forms that is critical for the development of a completely resolved taxonomic framework.

The integration of DNA information and taxonomy will require close collaboration among taxonomists, molecular biologists, bioinformatics specialists, freshwater ecologists, and resource managers/agencies. The primary focus of this work should be to establish a library of DNA signatures for recognized species. Also, a properly curated collection of voucher specimens that can be cross-referenced with the DNA signatures should be linked with this library, and revised according to changes in taxonomy. This should be viewed as a positive benefit to taxonomists, as the process of producing and analyzing sequences will provide them with data that can aid in the discovery of previously undescribed species and patterns of phylogeographic diversity, as well as in the linking of juvenile and adult forms.

Producing this sequence data requires access to physical specimens. Previously collected and archived material might provide a ready source for some species, assuming that there is intact DNA available for sequencing, but material for many other species will need to be collected, identified by experts, processed for sequence data, curated, and archived. In either case, the compilation and organization of material will require a nontrivial expenditure of time and funds, and must be viewed as a critical research activity. We suspect that material on the ~15,000 species of North American freshwater invertebrates could be collected within 5 years, given a successfully coordinated effort.

The first short-term goal is the large-scale sequencing of invertebrate samples over a geographically distributed set of samples. A parallel effort can be made by associating DNA signatures with described taxa, but this association is a longer-term objective, as it will be both time-consuming and labor intensive, and will require the direct involvement of expert taxonomists and, importantly for many taxa, the collection of adult life stages. During this effort, genetic data and taxonomy are synergistic. Novel genetic signatures will focus taxonomists on cryptic variation within species, and the linkages between described adults and larval forms will verify the identity of the DNA signatures. The process of collecting genetic data can begin immediately, and, given the requirement of 2000–3000 target species, can be completed within a reasonable time frame. The major constraint will be the collection of appropriate representative samples; making linkages with taxonomy will be an ongoing effort with a duration directly related to the intensity of that effort. Concurrently, an intense se-

quencing effort of invertebrate samples from a geographically limited set of stream samples would allow protocol development, provide specimens for a complete albeit geographically limited area, and provide proof-of-concept using genetic tools for freshwater assessment.

TECHNICAL ISSUES FOR DNA-BASED ASSESSMENT METHODS

The second major initiative is the modification and testing of sample collection protocols, as well as array and next-generation sequencing approaches, to develop a working, cost-effective biodiversity assay tool. This phase can begin once the emerging genetic database contains a sufficient amount of DNA sequences that represent target taxa. Because the appropriate technology is currently available and only requires modification to make it specific to freshwater bioassessment, this phase of development can rapidly follow the development of the genetic database. A final task will be to validate the efficacy of a genetic assay through a series of quality-control tests and pilot projects that directly compare the results of assessments based on standard morphological assay techniques.

Although the productive technical approaches to developing a DNA-based tool seem clear, significant challenges remain. For example, extracting high-quality DNA from an individual organism is routine, but efficiently doing the same on hundreds or even thousands of individuals in a single bulk sample may not be as straightforward. Advances in this area are a research priority, as the increase in processing time dictated by DNA extraction from individual samples could well offset the advantages of increased taxonomic resolution gained by a DNA-based approach. Methods for bulk DNA extraction need to be refined and rigorously tested, and these methods may vary depending on their particular application (e.g., microarray or next-generation sequencing) (Creer et al. 2010). Moving to a high-throughput bulk sample may also limit our ability to move beyond categorization of the presence/absence of taxa to a more quantitative assessment that includes relative abundance. In principle, it is possible to use arrays to quantify the abundance of DNA fragments; this is the underlying assumption of gene expression arrays that quantify the relative abundance of transcripts (DeSantis et al. 2005). However, estimating the relative abundance of DNA fragments in a pool of DNA extracted from a bulk sample of benthic invertebrates that vary by orders of magnitude in body size will likely be quite complicated. Some form of normalization, either in the DNA extraction or data processing phases, will be required.

Few studies have involved large quantities of extraction material and large numbers of samples. In a DNA-array study with large sample size, Robideau et al. (2008) examined 2000 fruit samples for evidence of fungal pathogens. They showed that as the sample size increased, the number of false negatives in the molecular assay increased up to 20% as compared with direct detection via plating, because calyx colonists grow readily on plates, but have low biomasses in fruit samples. The results of this study suggest that detection is a probability issue dependent on inclusion of DNA from low abundance colonists in the PCR reaction template. A similar issue would likely affect detection of small body size and/or rare invertebrates in benthic samples, and PCR instruments that work with increasingly smaller volumes to increase speed and reduce cost will compound this problem. The potential pitfalls of using a high-throughput DNA-based strategy for whole community diversity studies can be evaluated and resolved through controlled and replicated experimental studies. These studies should focus on issues of DNA isolation, detection limits, and error rates in a controlled laboratory setting. Finally, the efficacy of a particular platform will need validation in a field context by comparison to a detailed morphological assessment of species diversity.

An important point to consider is the distinct possibility that the "best" genetic assay platform may change rapidly in the next 5–10 years. Currently, microarray and next-generation sequencing approaches are the most promising avenues to pursue. However, given the rapid pace of technology in this

area, which is driven largely by the desire to characterize genomic level genetic variation in individuals and populations of humans, the most accurate and cost-effective assay platform will almost certainly change repeatedly in the near future. This realization reinforces the critical and central importance of a comprehensive DNA database with complementary vouchered collections that link genetic variation to taxonomy in benthic invertebrates. Given the cost and effort required to compile the necessary biological collections, the serious consideration of long-term maintenance of archived community DNA and RNA is warranted. As the technology becomes increasingly more cost-effective and comprehensive, it may be possible to mine these collections for additional taxonomic and functional diversity. Importantly, archived collections will form the basis for critical examinations of the effect of climate change on aquatic biotas over the next decades. This database and specimen collection are the lynchpins for assay tools that can be constructed now using current technology, and for those that will be developed with new technologies in the near future.

Managing and interpreting sequence data: the bioinformatics challenge

In order for it to be useful in an ecological context, we must be able to quickly translate the DNA data produced from assay instruments into a form used by ecologists (i.e., lists of species names or their codes) and into a file format that can be easily used by existing ecological software. The technology for handling such large arrays of data is generally available, but work will be needed to develop the most efficient ways of parsing the sequence information in order to unambiguously discriminate between taxonomic units and to then output that information in a format usable by ecologists. If they are to be truly useful, these sequence-taxa databases must be designed so that they will be able to communicate directly with existing taxa-ecology databases that house information regarding the ecological requirements and distributional records of species.

In summary, the establishment of a rapid genetic biodiversity assay will require a series of coordinated data collection efforts, as well as assay platform development. The primary requirement is a large data set of unique DNA signatures. These signatures can be coupled to described invertebrate taxa, and will form the basis for design of genetic assessment tools. A coordinated effort between taxonomists, molecular biologists, bioinformatics specialists, freshwater ecologists, and resource managers, focused by support from the appropriate state and federal agencies, should be able to effectively produce a viable toolset for DNA-based assessment of freshwater systems within the next 5–10 years.

Funding and research coordination

In our view, a priority for realizing the potential of DNA-based surveys for freshwater bioassessments is the establishment of a coordinated program of research support among those federal institutions that have some interest in either the development or application of biodiversity surveys. In the USA, the National Science Foundation (NSF) holds primary responsibility for funding the development of new science, while programs within the Environmental Protection Agency (EPA), the U.S. Geological Survey (USGS), U.S. Department of Agriculture (USDA), and the National Institute of Environmental Health Science (NIEHS) may be interested in supporting the application of this science to environmental and human health issues. Currently, no coordinated effort exists among these agencies to promote the development or refinement of the science supporting DNA-based surveys. Two critical issues need to be addressed. First, the appropriate agencies need to identify taxonomic and DNA-based bioassesment work as a high-priority research need when developing funding budgets. Second, these agencies must take primary responsibility for coordinating the multiple avenues of research that need to be pursued in parallel. This coordination could be achieved by tasking an appropriate federal research lab with this responsibility, or by funding a consortium of universities to oversee these efforts.

ACKNOWLEDGMENTS

The ideas and synthesis in this article were fostered by a workshop organized by MEP and CPH at Utah State University. The workshop was supported by funds from the USU Center for Integrated BioSystems, the USU Ecology Center, and the USU Office of Research. Although this work was reviewed by EPA and approved for publication, it may not necessarily reflect official Agency policy.

REFERENCES

Angly F. E., Felts B., Breibart M., Salamon P., Edwards R. A., Carlson C., Chan A. M., Haynes M., Kelley S., Lui H., Mahaffy J. M., Mueller J. E., Nulton J., Olson R., Parsons R., Rayhawk S., Suttle C. A., Rohwer F. 2006. The marine viromes of four oceanic regions. *PLoS Biology* 4(11):e368.

Armstrong K. F., Ball S. L. 2005. DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences* 360(1462):1813–1823.

Arscott D. B., Jackson J. K., Kratzer E. B. 2006. Role of rarity and taxonomic resolution in a regional and spatial analysis of stream macroinvertebrates. *Journal of the North American Benthological Society* 25(4): 977–997.

[ASIWPCA] Association of State and Interstate Pollution Control Administrators. 2002. *Water Quality Monitoring Programs Survey Report: Status and Future of State Ambient Water Quality Monitoring Programs.* Washington (DC): ASIWPCA.

Balian E. V., Lévesque C., Segers H., Martens K., editors. 2008. Freshwater animal diversity assessment. *Hydrobiologia* 595:1–637.

Ball S. L., Hebert P. D. N., Burian S. K., Webb J. M. 2005. Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *Journal of the North American Benthological Society* 24:508–524.

Bertolazzi P., Felici G., Weitschek E. 2009. Learning to classify species with barcodes. *BMC Bioinformatics* 10(supplement 14):S7.

Barrett R. D. H., Hebert P. D. N. 2005. Identifying spiders through DNA barcodes. *Canadian Journal of Zoology* 83(3):481–491.

Blaxter M. 2004. The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences* 359:699–679.

Blaxter M., Mann J., Chapman T., Thomas F., Whitton C., Floyd R., Abebe E. 2005. Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences* 360(1462):1935–1943.

Brodie E. L., DeSantis T. Z., Moberg Parker J. P., Zubietta I. X., Piceno Y. M., Anderson G. L. 2007. Urban aerosols harbor diverse and dynamic bacterial populations. *Proceedings of the National Academy of Scienes USA* 104(1):299–304.

Caesar R. M., Sörensson M., Cognato A. I. 2006. Integrating DNA data and traditional taxonomy to streamline biodiversity assessment: an example from edaphic beetles in the Klamath ecoregion, California, USA. *Diversity and Distributions* 12(5): 483–489.

Call D. R., Borucki M. K., Loge F. J. 2003. Detection of bacterial pathogens in environmental samples using DNA microarrays. *Journal of Microbiological Methods* 53(2):235–243.

Cannon C. H., Kua C. S., Lobenhofer E. K., Hurban P. 2006. Capturing genomic signatures of DNA sequence variation using a standard anonymous microarray platform. *Nucleic Acids Research* 34(18): e121.

Cao Y., Hawkins C. P. 2005. Simulating biological impairment to evaluate the accuracy of ecological indicators. *Journal of Applied Ecology* 42:954–965.

Cao Y., Hawkins C. P., Larsen D. P., Van Sickle J. 2007. Effects of sample standardization on mean species detectabilities and estimates of relative differences in species richness among assemblages. *American Naturalist* 170(3):381–395.

Cao Y., Larsen D. P., Hughes R. M., Angermeier P. L., Patton T. M. 2002a. Sampling effort affects multivariate comparisons of stream assemblages. *Journal of the North American Benthological Society* 21(4):701–714.

Cao Y., Williams D. D., Larsen D. P. 2002b. Comparison of ecological communities: the problem of sample representativeness. *Ecological Monographs* 72:41–56.

Carew M. E., Pettigrove V., Cox R. L., Hoffmann A. A. 2007a. DNA identification of urban Tanytarsini chironomids (Diptera: Chironomidae). *Journal of the North American Benthological Society* 26(4):587–600.

Carew M. E., Pettigrove V., Cox R. L., Hoffmann A. A. 2007b. The response of Chironomidae to sediment pollution and other environmental characteristics in urban wetlands. *Freshwater Biology* 52(12):2444–2462.

Carew M. E., Pettigrove V., Hoffmann A. A. 2003. Identifying chironomids (Diptera: Chironomidae) for biological monitoring with PCR-RFLP. *Bulletin of Entomological Research* 93:483–490.

Carew M. E., Pettigrove V., Hoffmann A. A. 2005. The utility of DNA markers in classical taxonomy: using Cytochrome Oxidase I markers to differentiate Australian *Cladopelma* (Diptera: Chironomidae) midges. *Annals of the Entomological Society of America* 98(4): 587–594.

Carter J. L., Resh V. H. 2001. After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *Journal of the North American Benthological Society* 20(4):658–682.

Castiglioni B., Rizzi E., Frosini A., Mugnai M. A., Ventura S., Sivonen K., Rajaniemi P., Rantala A., Wilmotte A., Boutte C., Consolandi C., Borboni R., Mezzelani A., Busti E., Rossi Bernardi L., Battaglia C., De Bellis G. 2002. Application of an universal DNA microarray to cyanobacterial diversity assessment. *Minerva Biotech* 14(3–4):253–257.

Chen W., Seifert K. A., Lévesque C. A. 2009. A high density COX1 barcode oligonucleotide array for identification and detection of species of Penicillium subgenus Penicillium. *Molecular Ecology Resources* 9(supplement 1):114–129.

Clarke R. T., Lorenz A., Sandin L., Schmidt-Kloiber A., Strackbein J., Kneebone N. T., Haase P. 2006. Effects of sampling and sub-sampling variation using the STAR-AQEM sampling protocol on the precision of macroinvertebrate metrics. *Hydrobiologia* 566:441–459.

Cognato A. I., Caesar R. M. 2006. Will DNA Barcoding advance efforts to conserve biodiversity more efficiently than traditional taxonomic efforts. *Frontiers in Ecology and the Environment* 4(5):268–270.

Comai L., Young K., Till B. J., Reynolds S. H., Greene E. A., Codomo C. A., Enns L. C., Johnson J. E., Burtner C., Oden A. R., Henikof S. 2004. Efficient discovery of DNA polymorphisms in natural populations by Ecotiling. *Plant Journal* 37(5):778–786.

Cranston P. S. 1995. Introduction. Pages 1–7 in *The Chironomidae: The Biology and Ecology of Non-Biting Midges*, edited by P. D. Armitage et al. London (UK): Chapman and Hall.

Creer S., Fonseca V. G., Porazinska D. L., Giblin-Davis R. M., Sung W., Power D. M., Packer M., Carvalho G. R., Blaxter M. L., Lambshead P. J. D., Thomas W. K. 2010. Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology* 19(s1):4–20.

DeSantis T. Z., Stone C. E., Murray S. R., Moberg J. P., Andersen G. L. 2005. Rapid quantification and taxonomic classification of environmental DNA from prokaryotic and eukaryotic origins using a microarray. *FEMS Microbiology* Letters 245(2):271–278.

Diaz M. R., Boekhout T., Rheelen B., Bovers M., Cabañes F. J., Fell J. W. 2006. Microcoding and flow cytometry as a high-throughput fungal identification system for *Malassezia* species. *Journal of Medical Microbiology* 55:1197–1209.

Dinsdale E. A., Edwards R. A., Hall D., Angly F., Breitbart M., Brulc J. M., Furlan M., Desnues C., Haynes M., Li L., McDaniel L., Moran M. A., Nelson K. E., Nilsson C., Olson R., Paul J., Brito B. R., Ruan Y., Swan B. K., Stevens R., Valentine D. L., Thurber R. V., Wegley L., White B. A., Rohwer F. 2008. Functional metagenomic profiling of nine biomes. *Nature* 452:629–632.

Engelmann J. C., Rahmann S., Wolf M., Schultz J., Fritzilas E., Kneitz S., Dandekar T., Muller T. 2009. Modelling cross-hybridization on phylogenetic DNA microarrays increases the detection power of closely related species. Mol*ecular Ecology Resources* 9(1):83–93.

Epler J. H. 2001. *Identification Manual for the Larval Chironomidae (Diptera) of North and South Carolina: A Guide to the Taxonomy of the Midges of the Southeastern United States, including Florida.* Special Publication SJ2001-SP13. Raleigh (NC) and Palatka (FL): North Carolina Department of Environment and Natural Resources, and St. Johns River Water Management District.

Fessehaie A., De Boer S. H., Lévesque C. A. 2003. An oligonucleotide array for the identification and differentiation of bacteria pathogenic on potato. *Phytopathology* 93:262–269.

Floyd R., Abebe E., Papert A., Blaxter M. 2002. Molecular barcodes for soil nematode identification. *Molecular Ecology* 11:839–850.

Gentry T. J., Wickham G. S., Schadt C. W., He Z., Zhou J. 2006. Microarray applications in microbial ecology research. *Microbial Ecology* 52(2):159–175.

Gilchrist E. J., Haughn G. W., Ying C. C., Otto S. P., Zhuang J., Cheung D., Hamberger B., Aboutorabi F., Kalynyak T., Johnson L., Bohlmann J., Ellis B. E., Douglas C. J., Cronk Q. C. B. 2006. Use of ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. *Molecular Ecology* 15(5):1367–1378.

Gracey A. Y., Cossins A. R. 2003. Application of microarray technology in environmental and comparative physiology. *Annual Review of Physiology* 65: 231–259.

Gresham D., Ruderfer D. M., Pratt S. C., Schacherer J., Dunham M. J., Botstein D., Kruglyak L. 2006. Genome-wide detection of polymorphism at nucleotide resolution with a single DNA microarray. *Science* 311(5769):1932–1936.

Guerold F. 2000. Influence of taxonomic determination level on several community indices. *Water Research* 34:487–492.

Hajibabaei M., Singer G. A. C., Clare E. L., Hebert P. D. N. 2007. Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biology* 5:24.

Harvey J. B. J., Hoy M. S., Rodriguez R. J. 2009. Molecular detection of native and invasive marine invertebrate larvae present in ballast and open water environmental samples collected in Puget

Sound. *Journal of Experimental Marine Biology and Ecology* 369:93–99.

Hawkins C. P. 2006. Quantifying biological integrity by taxonomic completeness: its utility in regional and global assessments. *Ecological Applications* 16(4): 1277–1294.

Hawkins C. P., Norris R. H., Hogue J. N., Feminella J. W. 2000. Development and evaluation of predictive models for measuring biological integrity in streams. *Ecological Applications* 10:1456–1477.

Hawkins C. P., Olson J. R., Hill R. A. 2010. The reference condition: predicting benchmarks for ecological and water-quality assessments. *Journal of the North American Benthological Society* 29(1):312–343.

Hebert P. D. N., Penton E. H., Burns J. M., Janzen D. H., Hallwachs W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences USA* 101(41):14812–14817.

He Z., Gemntry T. J., Schadt C. W., Wu L., Liebich J., Chong S. C., Huang Z., Wu W., Gu B., Jardine P., Criddle C., Zhou J. 2007. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological, and environmental processes. *ISME Journal* 1:67–77.

Hughes R. M., Larsen D. P., Omernik J. M. 1986. Regional reference sites: a method for assessing stream potentials. *Environmental Management* 10: 629–635.

Jones F. C. 2008. Taxonomic sufficiency: the influence of taxonomic resolution on freshwater bioassessments using benthic macroinvertebrates. *Environmental Reviews* 16:45–69.

King R. S., Richardson C. J. 2002. Evaluating subsampling approaches and macroinvertebrate taxonomic resolution for wetland bioassessment. *Journal of the North American Benthological Society* 21(1):150–171.

Lamche G., Fukuda Y. 2008. *Comparison of Genus and Family Level AUSRIVAS Models for the Darwin-Daly Region in Relation to Land Use*. Report 01/20008D. Northern Territory (Australia): Aquatic Health Unit, Department of Natural Resources, Environment and the Arts, Northern Territory Government.

Lenat D. R., Resh V. H. 2001. Taxonomy and stream ecology—the benefits of genus and species level identifications. *Journal of the North American Benthological Society* 20(2):287–298.

Lévesque C. A., Harlton C. E., de Cock A. W. A. M. 1998. Identification of some oomycetes by reverse dot blot hybridization. *Phytopathology* 88(3):213–222.

Lorenz A., Kirchner L., Hering D. 2004. 'Electronic subsampling' of macrobenthic samples: how many individuals are needed for a valid assessment result? *Hydrobiologia* 516(1):299–312.

Loy A., Schultz C., Lücker S., Schöpfer-Wendels A., Stoecker K., Baranyi C., Lehner A., Wagner M. 2005. 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial Order "Rhodocyclales." *Applied Environmental Microbiology* 71(3):1373–1386.

Lozupone C. A., Knight R. 2007. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences USA* 104(27):11436–11440.

Margulies M., Egholm M., Altman W. E., Attiya S., Bader J. S., Bemben L. A., Berka J. et al. 2005. Genome sequencing in open microfabricated high-density picoliter reactors. *Nature* 437(7057): 376–380.

Markmann M., Tautz D. 2005. Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences* 360(1462):1917–1924.

Meier R., Shiyang K., Vaidya G., Ng P. K. L. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* 55(5):715–728.

Meyer M., Briggs A. W., Maricic T., Höber B., Höffner B., Krause J., Weihmann A., Pääbo S., Hofreiter M. 2008. From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Research* 36:e5.

Meyer M., Stenxel U., Myles S., Prüfer K., Hofreiter M. 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research* 35:e96.

Neigel J., Domingo A., Stake J. 2007. DNA barcoding as a tool for reef conservation. *Coral Reefs* 26:487–499.

Nichols S. J., Robinson W. A., Norris R. H. 2006. Sample variability influences on the precision of predictive bioassessment. *Hydrobiologia* 572(1): 215–233.

Ostermiller J. D., Hawkins C. P. 2004. Effects of sampling error on bioassessments of stream ecosystems: application to RIVPACS-type models. *Journal of the North American Benthological Society* 23:363–382.

Parameswaran P., Jalili R., Tao L., Shokralla S., Garizadeh B., Ronaghi M., Fire A. Z. 2007. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research* 35(19):e130.

Pfenninger M., Nowak C., Kley C., Steinke D., Streit B. 2007. Utility of DNA taxonomy and barcoding for the inference of larval community structure in

morphologically cryptic *Chironomus* (Diptera) species. *Molecular Ecology* 16:1957–1968.

Pfunder M., Holzgang O., Frey J. E. 2004. Development of a microarray-based diagnostics of voles and shrews for use in biodiversity monitoring studies, and evaluation of mitochondrial cytochrome oxidase I vs. cytochrome *b* as genetic markers. *Molecular Ecology* 13(5):1277–1286.

Pond G. J., Passmore M. E., Borsuk F. A., Reynolds L., Rose C. J. 2008. Downstream effects of mountaintop coal mining: comparing biological conditions using family- and genus-level macroinvertebrate bioassessment tools. *Journal of the North American Benthological Society* 27(3):717–737.

Porazinska D. L., Giblin-Davis R. M., Faller L., Farmerie W., Kanzaki N., Morris K., Powers T. O., Tucker A. E., Sung W., Thomas K. 2009. Evaluating high-throughput sequencing as a method for metagemonic analysis of nematode diversity. *Molecular Ecology Resources* 9(6):1439–1450.

Porazinska D. L., Sung W., Giblin-Davis R. M., Thomas W. K. 2010. Evaluating high-throughput sequencing as a method for metagemonic analysis of nematode diversity. *Molecular Ecology Resources* 9(6):1439–1450.

Ratnasingham S., Hebert P. D. N. 2007. BOLD: the barcode of life data system: barcoding. *Molecular Ecology Notes* 7:355–364.

Reynoldson T. B., Wright J. F. 2000. The reference condition: problems and solutions. Pages 293–303 in *Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques,* edited by J. F. Wright et al. Ambleside (UK): Freshwater Biological Association, The Ferry House.

Robideau G. P., Caruso F. L., Oudemans P. V., McManus P. S., Renaud M. A., Auclair M. E., Bilodeau G. J., Yee D., Desaulniers N. L., DeVerna J. W., Lévesque C. A. 2008. Detection of cranberry fruit rot fungi using DNA array hybridization. *Canadian Journal of Plant Pathology-Revue Canadienne de Phytopathologie* 30:226–240.

Rosenberg D. M., Resh V. H. 1993. *Freshwater Biomonitoring and Benthic Macroinvertebrates.* New York: Chapman & Hall.

Schmidt-Kloiber A., Nijboer R. C. 2004. The effect of taxonomic resolution on the assessment of ecological water quality classes. *Hydrobiology* 516(1):269–283.

Sharley D. J., Pettigrove V., Parsons Y. M. 2004. Molecular identification of *Chironomus* spp. (Diptera) for biomonitoring of aquatic ecosystems. *Australian Journal of Entomology* 43:359–365.

Siefert K. A., Lévesque C. A. 2004. Phylogeny and molecular diagnostics of mycotoxigenic fungi. *European Journal of Plant Pathology* 110:449–471.

Sinclair C. S., Gresens S. E. 2008. Discrimination of Cricotopus species (Diptera: Chironomidae) by DNA barcoding. *Bulletin of Entomological Research* 1:1–9.

Skevington J. H., Kehlmaier C., Stahls G. 2007. DNA barcoding: mixed results for big-headed flies (Diptera: Pipunculidae). *Zootaxa* 1423:1–26.

Song H., Buhay J. E., Whiting M. F., Crandall K. A. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences USA* 105(36):13486–13491.

Stoddard J. L., Larsen D. P., Hawkins C. P., Johnson R. K., Norris R. H. 2006. Setting expectations for the ecological condition of streams: the concept of reference condition. *Ecological Applications* 16(4):1267–1276.

Stoughton R. B. 2005. Applications of DNA microarrays in biology. *Annual Review of Biochemistry* 74:53–82.

Stribling J. B., Pavlik K. L., Holdsworth S. M., Leppo E. W. 2008. Data quality, performance, and uncertainty in taxonomic identification for biological assessments. *Journal of the North American Benthological Society* 27(4):906–919.

Tambong J. T., de Cock A. W. A. M., Tinker N. A., Lévesque C. A. 2006. Oligonucleotide array for identification and detection of *Pythium* species. *Applied and Environmental Microbiology* 72(4):2691–2706.

Teletchea F., Bernillon J., Duffraisse M., Laudet V., Hanni C. 2008. Molecular identification of vertebrate species by oligonucleotide microarray in food and forensic samples. *Journal of Applied Ecology* 45:967–975.

Thomas M. A., Klaper R. 2004. Genomics for the ecological toolbox. *Trends in Ecology and Evolution* 19(8):439–445.

Thorp J. H., Covich A. P., editors. 2001. *Ecology and Classification of North American Freshwater Invertebrates.* Second Edition. San Diego (CA): Academic Press.

[USEPA] United States Environmental Protection Agency. 2002. *Summary of Biological Assessment Programs and Biocriteria Development for States, Tribes, Territories, and Interstate Commissions: Streams and Wadeable Rivers.* EPA-822-R-02-048. Washington (DC): U.S. Environmental Protection Agency, Office of Environmental Information and Office of Water.

[USEPA] United States Environmental Protection Agency. 2006. *Wadeable Streams Assessment: A Coolaborative Survey of the Nation's Streamss.* EPA-841-B-06-002. Washington (DC): U.S. Environmental Protection Agency, Office of Research and Development and Office of Water.

Wagner M., Smidt H., Loy A., Zhou J. Z. 2007. Unravelling microbial communities with DNA-microar-

rays: challanges and future directions. *Microbial Ecology* 53(3):498–506.

Will K. W., Rubinoff D. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20(1):47–55.

Zahariev M., Dahl V., Chen W., Lévesque C. A. 2009. Efficient algorithms for the discovery of DNA oligonucleotide barcodes from sequence databases. *Molecular Ecology Notes* 9:58–64.

Zhou J., Kang S., Schadt C. W., Garten C. T., Jr. 2008. Spatial scaling of functional gene diversity across various microbial taxa. *Proceedings of the National Academy of Sciences USA* 105(22):7768–7773.

Associate Editor: Kent E. Holsinger