Deep sequencing of environmental DNA isolated from the Cuyahoga River highlights the utility of river water samples to query surrounding aquatic and terrestrial biodiversity.

Cannon MV[1], Hester J[1,2], Shalkhauser A[1], Chan ER[1,3], Logue K[1], Small ST[1,4] and Serre D[1]

[1]Genomic Medicine Institute, Cleveland Clinic, 9500 Euclid Ave., Cleveland OH 44195, USA

[2]current address: BioConductor, Fred Hutchinson Cancer Center, 1100 Fairview Ave., Seattle WA 98109, USA

[3]current address: Institute for Computational Biology, Case Western Reserve University, 10900 Euclid Ave., Cleveland OH 44106, USA

[4]Department of Global Health and Diseases, Case Western Reserve University, 10900 Euclid Ave., Cleveland OH 44106, USA

Correspondence:

David Serre
Genomic Medicine Institute
Cleveland Clinic
9500 Euclid Ave.
Cleveland OH 44195, USA
Fax +1 216 636-0009, E-mail: serred@ccf.org

Running Title: Deep sequencing of environmental DNA

Keywords: DNA barcoding, environmental DNA, high-throughput sequencing, biodiversity.

Abstract

Analysis of environmental DNA (eDNA) enables detection of specific species from water and soil samples. Typically, these analyses are performed by amplifying a short DNA sequence using species-specific primers. Alternatively, primers amplifying many species within a taxonomic group are used and amplicons are subsequently sequenced. Here, we describe a method to quickly characterize the biodiversity of a given environment by amplification of eDNA using a combination of primer pairs targeting a wide range of taxa and species identification by high-throughput sequencing. We tested this approach by analyzing 91 water samples of 40mL collected along the Cuyahoga River (Ohio, USA). We amplified eDNA extracted from each water sample using 12 primer pairs targeting mammals, fish, amphibians, birds, bryophytes, arthropods, copepods, plants and several microorganism taxa and simultaneously sequenced all PCR products by high-throughput sequencing. Despite the small sample volumes analyzed, we identified DNA sequences from 15 species of fish, 17 species of mammals, 8 species of birds, 15 species of arthropods, one turtle and one salamander. Interestingly, in addition to aquatic and semi-aquatic animal species, we identified DNA from many terrestrial species that live near the banks of the Cuyahoga River. We also identified DNA from one Asian carp species invasive to the Great Lakes but that had not been reported in the Cuyahoga River. Our study shows that analysis of eDNA extracted from a small sample of water using wide-range PCR amplification combined with massively parallel sequencing can provide a broad perspective on the biological diversity of a given environment.

Introduction

Environmental samples, such as river and pond water or soil, contain a complex mixture of fragmented DNA molecules originating from the feces, mucous, gametes, shed tissues or decaying parts from organisms living in or near the sampling site (Taberlet et al. 2012). DNA extracted from these samples (often referred to as environmental DNA or eDNA) can be amplified to test for the presence of a particular species and this approach has been widely applied to ecological studies (Goldberg et al. 2011; Minamoto et al. 2012; Biggs et al. 2015), conservation (Pilliod et al. 2014; Treguier et al. 2014) or to identify the presence of invasive species (Mahon et al. 2013). eDNA studies typically focus on the analysis of a single macro-organism species and therefore rely on species-specific amplification of DNA. By contrast, characterization of microbial communities from environmental samples (Tringe et al. 2005) are often conducted using "universal" primers amplifying the 16S ribosomal RNA gene (rRNA) that yield, after DNA sequencing, enough sequence information to identify the species carrying each DNA molecule.

Here, we present an approach that combines these two methods and provides a quick and cost-efficient assessment of the macroorganism diversity in small environmental samples using deep sequencing of PCR products amplified using taxon-specific primers. First, we describe a novel R package that enables *in silico* evaluation of the specificity and information content of "universal" primer pairs. Next, we describe the application of this approach to the analysis of eDNA extracted from 91 samples of 40 mL of surface water collected along the Cuyahoga River, Ohio, USA. We amplified each sample using 12 primer pairs targeting mammals, fish, amphibians, birds, bryophytes, arthropods, copepods and plants (as well as several microorganism taxa) and, after indexing, sequenced simultaneously the PCR products by massively parallel sequencing. Our analyses show that this methodology can provide a broad

72 perspective on the aquatic and terrestrial biodiversity at the sampled sites in a simple, rapid and cost-
73 effective manner.
74
75 **Results**
76 *In silico assessment of the specificity and information content of different primer pairs*
77 PCR primers are typically designed to amplify one specific locus in a given organism. Even "universal"
78 primers, designed to amplify many species within the same taxon, are usually used only to amplify DNA
79 extracted from a single organism (of this taxon). For studies amplifying eDNA or DNA from unknown
80 taxa, we require that the amplification works on all members of a given taxon while avoiding off-target
81 amplification (that could reduce the number of sequences from the desired taxon). In addition, the
82 amplified region should contain enough sequence information to identify the species carrying the DNA
83 sequence. PrimerTree enables a rapid and visual assessment of these parameters for any primer pair by
84 displaying the results of *in silico* PCR as a taxonomically-annotated phylogenetic tree. For example,
85 **Figure 1** shows a subset of the PrimerTree results for a primer pair targeting the mammalian
86 mitochondrial 16S ribosomal RNA genes (**Figure 1A**) and primers designed to amplify the chloroplast
87 trnL gene of non-vascular plants (**Figure 1B**). The tree display enables rapid evaluation of the specificity
88 of the primer pairs (e.g., off-target amplification of amphibians and ray-finned fishes on **Figure 1A**). In
89 addition, the information content can be easily assessed by the length of the branches leading to
90 different sequences (scaled in number of nucleotide differences). For example, PrimerTree reveals much
91 longer branch lengths on **Figure 1A** than in **Figure 1B** suggesting a better discriminating power for the
92 mammalian sequences than the bryophyte sequences (see also below). By default, PrimerTree displays
93 phylogenetic trees annotated at all taxonomic levels enabling the user to determine the level of
94 specificity of each primer set (**Supplemental Figure 1**).

Identification of aquatic and terrestrial organisms from river water samples

We analyzed DNA extracted from 40 mL of surface water collected in 91 sites along the Cuyahoga River (**Figure 2**). For each sample, we performed 12 PCR amplifications targeting mammals, amphibians, birds, fish, arthropods, copepods, bryophytes and vascular plants (as well as several microorganism taxa not analyzed here). We then individually indexed the PCR products of each sample and sequenced them on an Illumina MiSeq (**Figure 3**) to generate a total of 10,507,986 paired-end reads (**Table 1**). After stringent quality filtering we retained between 1,645,452 and 6,213 reads (14-95%) for the analysis of each taxon (**Table 1**).

In contrast to microorganisms, where several hundred species are likely present in a given sample, we only expect to amplify DNA from a few different macro-organism species per sample. Therefore, even if the number of initial DNA molecules from a species is low, as long as the template can be amplified efficiently, many reads will be generated from these few DNA sequences. For example, if a sample had DNA from ten mammals and one organism only accounted for 1% of the total DNA, its DNA sequence would be represented, on average, by 142 reads (mammal amplifications averaged 14,211 reads per sample). We therefore considered that DNA sequences represented by less than 10 reads total across all samples were caused by sequencing errors and discarded them, removing, depending on the taxon considered, between 1% and 25% of all reads generated (**Table 1**). We blasted the remaining DNA sequences to identify the closest sequences in NCBI and to assign a species label to each DNA sequence. In agreement with our *in silico* analyses, we observed large variations among primers in the specificity of the taxa identified (**Supplemental Table 1**). For some primers, the sequence information was insufficient to differentiate the organisms down to the species level: for example, each DNA sequence amplified from the trnL gene of green plants matched sequences from 34.89 different species on average. In fact,

118 DNA sequences amplified from this primer pair matched a single taxon only when considering families
119 or higher taxonomic levels. This contrasted with more informative sequences such as the mammalian
120 16S rRNA for which each DNA sequence generated matched, on average, 1.27 species (**Supplemental**
121 **Table 1**). Note that the observed specificity of the bird primers differed from that expected from *in silico*
122 analyses (12.39 species per DNA sequence generated compared to 1.57 expected based on sequences
123 available in NCBI). This apparent low specificity in our data was caused by the presence of many DNA
124 sequences from thrushes (a family of passerine birds) for which many species with the exact same DNA
125 sequences at the 12S rRNA gene have been sequenced. In addition to differences in their information
126 content (that influences the ability to identify a given sequence), primer pairs also differed in the
127 amplification specificity (as predicted by PrimerTree analyses). For example, while initially designed to
128 amplify mammals (Taylor 1996), the mammalian 16S rRNA primers also amplified *Actinopterygii* (ray-
129 finned fishes). On the other hand, the primers targeting cytochrome B of fish (Thomsen et al. 2012) only
130 amplified carp and creek chub (*Semotilus atromaculatus*) under our PCR conditions (**Supplemental Table**
131 **2**). The samples positive for carp by the 16S primers were usually also positive by the cytochrome B
132 primers. Of the 15 samples positive for carp by the cytochrome B primers, 11 were also positive by the
133 16S primers (**Supplemental Table 3**). The 16S primers were, however, positive in a total of 31 samples
134 for carp. Only two samples were positive for creek chub. Consequently, for all subsequent analyses we
135 used fish sequences amplified by the mammalian 16S RNA primers rather than those amplified with the
136 cytochrome B fish primers.
137 The extraction and PCR controls almost exclusively yielded human DNA sequences (>99 % of the reads)
138 with one extraction control also displaying pig DNA in 0.6 % of the reads. The human contamination
139 likely originated in the laboratory and is difficult to prevent under standard laboratory conditions.
140 Overall, these controls indicated that most DNA sequences retrieved from the water samples must be

141 genuine and that cross-contamination in the laboratory was minimal. Overall, across 91 water samples
142 collected along the Cuyahoga River, we identified 54 samples positive for fish DNA (representing 15
143 species), 77 samples positive for mammalian DNA (17 species excluding human), 12 samples positives
144 for bird DNA (from at least eight species), 18 samples positive for arthropod DNA (15 species), 16
145 samples positive for copepod DNA (two species) while the "amphibian" primers amplified turtle and
146 two-lined salamander DNAs in two samples (**Table 2, Supplemental Table 3**). In addition to many
147 organisms living in the river (e.g., fish, aquatic insects) or semi-aquatic animals (beaver, mink, muskrat),
148 we also amplified DNA from many terrestrial species that live near the banks of the Cuyahoga River such
149 as raccoon, groundhog, squirrel or mouse. Similarly, we identified DNA from many birds that live on
150 (swan, duck, sea gulls) as well as near the river (sparrow, wild turkey). The species identified also often
151 corresponded with the local environment where the samples were collected: for example, beaver DNA
152 was amplified from samples collected in protected forested areas, gull DNA near Lake Erie. In this
153 regard, it is interesting to note that fish DNA showed significant differences in their geographical
154 distribution. DNA from fish of the *Moxostoma* genus (probably Silver Redhorse) was commonly detected
155 in the Upper Cuyahoga River but rare elsewhere ($p=0.02$, **Figure 2**). The central stoneroller was detected
156 only in the middle and lower Cuyahoga ($p=3.1\times10^{-3}$). On the other hand, common carp were found
157 throughout the entire Cuyahoga River ($p=0.23$). Surprisingly, we also identified DNA from one invasive
158 Asian Carp species in the Cuyahoga River near Lake Erie (**Figure 2**).
159
160 *Temporal variations in diversity*
161 We performed the collection of water samples at two time points separated by an episode of heavy rain
162 falls that dramatically altered the water level and flow of the Cuyahoga River: the discharge at Hiram
163 Rapids, in the Upper Cuyahoga River, increased from 2.46 cubic meters per second (close to the median

daily statistics) at the time of the first sampling to 20.87 cubic meters per second and was still as high as high as 5.95 cubic meters per second at the time of the second sampling 12 days later (**Supplemental Figure 2**). For most macro-organisms, we did not detect any statistical difference between the samples collected at the two time points. One notable exception concerns fish DNA that was more often detected in samples collected before rain than after (p=0.05, Fisher's exact test).

Discussion

Since the first reports that DNA could be retrieved from environmental samples (Ogram et al. 1987), studies of environmental DNA have broadened in scope from studies of bacterial communities (e.g., (Tringe et al. 2005; Rusch et al. 2007)), to identification and monitoring of a given species (e.g., (Ficetola et al. 2008; Goldberg et al. 2011; Thomsen et al. 2012; Mahon et al. 2013)) and the characterization of microorganism populations (e.g., (Bik et al. 2012; Hirai et al. 2015)) and recently to the identification of macroinvertebrates or vertebrates (e.g., (Andersen et al. 2012; Deiner and Altermatt 2014; Kelly et al. 2014; Mächler et al. 2014)). However, several factors have limited a broader implementation of these approaches for ecological studies. These limitations include the lack of tools enabling a wide range of species to be studied simultaneously and the amount of starting material (often liters of water for aquatic environments). Additionally, the careful evaluation of potential primer pairs prior to laboratory work is critical as most published primers have only been tested on DNA directly extracted from the target organism and, while efficient, might not be specific or could even better amplify other organisms.

In silico assessment of primers

We present here a simple R package that enables rapid screening of primers suitable for differentiating species from a chosen taxon. PrimerTree allows evaluating the specificity of a given primer pair (to avoid off-target amplification) and whether the amplified DNA sequences would provide enough information

187 to identify the organisms carrying the DNA sequences. Since the amplified regions are typically short
188 (100-300 bp), the resulting phylogenetic trees (**Figure 1** and **Supplemental Figure 1**) do not necessarily
189 reflect the true species relationships, but they enable an easy and rapid assessment of the primer
190 suitability. First, the automatic taxonomic annotation of the branches allows the user to quickly see
191 which taxa are amplifiable by the primers. This enables identifying that undesirable (off-target) taxa may
192 be amplified or that the primers might not amplify specific genera within the taxon of interest. Second,
193 the branch lengths show how many nucleotides separate DNA sequences from different species. Short
194 branch lengths might results in a low resolution in taxon identification (or possibly false-positive results
195 due to sequencing or PCR errors) while long branches will lead to high confidence species identification
196 (e.g., **Figure 1A** vs. **1B**).
197 An attractive feature of PrimerTree is its ease of use. As an R package, it is easily installed or updated
198 and only requires that clustal be installed in the user's path. The entire process of getting blast hits for a
199 primer pair is done through R using a single command and plots are generated using a second
200 command. Additionally, information on the blast results, sequences obtained, taxonomy of amplifiable
201 sequences, neighbor joining distance matrix and phylogenetic tree are preserved within the PrimerTree
202 object if the user wants to further analyze or summarize the results. Finally, PrimerTree runs fast: for a
203 primer pair with no degenerate bases, PrimerTree results are usually obtained in less than five minutes.
204 One limitation of PrimerTree is that, by default, it only retrieves up to 500 random amplifiable
205 sequences from the NCBI database and this random subset might not accurately represent the
206 specificity of a given primer pair for a particular environment. For example, the bird primers used in our
207 study display a high species-specificity for most birds but were not able to differentiate among thrushes.
208 Note however that this issue can be easily circumvented by selecting specific targets in PrimerTree if
209 one is interested in a particular taxon (e.g., one could run PrimerTree for querying only *Turdidae*

sequences in NCBI to evaluate if a given primer pair is informative within this taxon). Additionally, the problem of returning a random subset of all sequences can be minimized by requesting more amplifiable sequences using the num_aligns argument in the command. Another important limitation is that PrimerTree does not highlight taxonomic groups that are not retrieved. If a particular taxonomic group is not amplifiable by a primer pair, the user must recognize the absence of that taxon. For example, if one primer pair amplifies all mammals except monotremes, the user must note the absence of the order *Monotremata* from the resulting plot or taxonomic information in the PrimerTree object. Along these same lines, the absence of a species or larger taxonomic group from PrimerTree results may not mean that the assayed primers cannot amplify those taxa. As PrimerTree uses the blast nr/nt database as a reference, any species without DNA sequence for the targeted locus in the database cannot be retrieved. This is one reason why PrimerTree cannot indicate missing species/taxa. To do this, we would need to first accurately distinguish 1) species with sequences within the blast database that are homologous to the targeted loci but where the primers cannot bind efficiently from 2) species not sequenced at this locus.

EcoPCR (Ficetola et al. 2010) also allows evaluation of primer specificity but by utilizing a user-provided database to identify amplifiable sequences and providing a summary of sequences and taxonomic information for amplifiable sequences. The ecoPCR program retrieves all amplifiable sequences in the provided database, compared to only a subset of the BLAST nr/nt database for PrimerTree. However, one advantage of PrimerTree is that it always queries the most current available version of the entire BLAST nr/nt database, while the user would need to update the database provided to ecoPCR and re-generate the ecoPCR database files to include new sequences in the analysis. Additionally PrimerTree benefits from the graphical output, which is automatically generated using the plot() command in R, to summarize the results in a user-friendly manner.

233 *Characterization of eDNA extracted from the Cuyahoga River*

234 We analyzed eDNA extracted from 91 samples of water collected along the Cuyahoga River. We showed
235 that, despite a small amount of starting material, we were able to retrieve eDNA from a wide range of
236 organisms, including many vertebrate species: on average, each water sample contained eDNA from 3.2
237 vertebrates (fish, mammal or bird), 0.3 arthropods and numerous plants. We noted that several species
238 were unevenly distributed along the Cuyahoga, such as the Silver Redhorse that was mostly detected in
239 the Upper Cuyahoga, which demonstrates the potential of our approach to identify variations in species
240 distribution (**Figure 2**). Among the 48 water samples positive for fish DNA, we also identified one sample
241 containing DNA from one Asian carp (either *Ctenopharyngodon idella* or *Mylopharyngodon piceus*),
242 invasive species of the Great Lakes that were not known to be present in the Cuyahoga River (though
243 they had been previously reported in western sections of Lake Erie) (**Figure 2**).
244 Our two samplings of the Cuyahoga River were separated by a major rain event. Since we sampled the
245 upper portion of the Cuyahoga both before and after the rain, we compared species detection rates to
246 determine if the influx of water into the river would stir up sediment and wash genetic material into the
247 river to increase species detection rates or dilute the river water and decrease species detection rates.
248 We found that the number of samples positive for mammals, birds, arthropods and copepods did not
249 change after the rain, whereas the number of samples positive for fish species decreased after the rain.
250 This may suggest that, for fish, the rainwater dilutes the genetic material of fish, but for other taxa this
251 dilution effect might be balanced by the influx of genetic material in water washing into the river.
252 Replication of these results using other sample sets is necessary to confirm that this is a general
253 phenomenon, but given only our data, sampling during a period with minimal rain may maximize the
254 detection rates of fish species.

255 Certain taxa were identified less frequently than we would have expected (e.g., amphibians, arthropods
256 and copepods, **Table 1**). This could be due to a number of factors. Our DNA extraction protocol (relying
257 on the DNeasy extraction kit after centrifugation) may not have been optimal to retrieve DNA from such
258 organisms as the choice of DNA isolation protocol from water samples can strongly influence the
259 proportion of DNA isolated from different taxa (Rees et al. 2014). In particular, it is important to note
260 that our isolation method was unlikely to capture free DNA in the river water sample, but rather only
261 DNA within, or adhered to, particulate matter which could lead to disproportional representation of the
262 eDNA (see e.g. Turner et al. (2014); Deiner et al. (2015). It is also possible that the PCR conditions we
263 used were inadequate to efficiently amplify some of the targeted templates (Elbrecht and Leese 2015;
264 Pinol et al. 2015). Alternatively, the amount of genetic material present in the river may be lower for
265 these taxa than for others such as fish or mammals (possibly due to differences in the size of the dead
266 animals or the amount of feces or shed tissues). Additional studies, including experimental validation of
267 the primers for these taxa, will be necessary to differentiate these possibilities.

268 *A fast, high-throughput and cost-efficient method to characterize biodiversity from environmental*
269 *samples*

270 We described here a customizable approach that builds on existing methods (Hajibabaei et al. 2012;
271 Gibson et al. 2014; Elbrecht and Leese 2015; Pinol et al. 2015) to enable simultaneous analyses of eDNA
272 extracted from many samples for a wide range of taxa. This method provides several advantages over
273 classical approaches. First, the most exciting feature of our findings is the amplification of DNA from
274 many terrestrial organisms from the river samples. We notably amplified sequences from deer, squirrel,
275 raccoon, groundhog, vole, mink and skunk in addition to semi-aquatic species such as beaver and
276 muskrat. We also identified DNA sequences for agricultural species (cow and swine) and companion
277 animals (dog and cat). DNA from these species in the river sample likely originated from fecal matter

278 washing into the river, deceased animals in or around the river or animals drinking or wading in the
279 river. Similarly, we detected many avian sequences, including species that do not live in the river (e.g.,
280 chickens, turkeys, blackbirds and sparrows). These findings illustrate the power of eDNA studies to
281 survey, not only organisms living in a river, but also the overall biodiversity of the environment
282 neighboring the river. This feature is particularly exciting as it may provide a unique opportunity to
283 better characterize the biodiversity of environments that are difficult to access and sample, such as
284 dense tropical forests where most of the unexplored diversity of many taxa still reside.
285 Second, we showed that this methodology can be successfully applied to analyze small water samples.
286 This aspect, while it has its own limitations (see below), is essential for any study for which sampling
287 and storing of large volumes is logistically challenging or impossible. In particular, small volume
288 collection enables sampling in sites that can only be reached by hiking or paddling and will therefore
289 make this approach suitable for many ecological studies. Additionally, sample processing (such as
290 filtration) in the field can be slow and requires specialized equipment, particularly for large sample
291 volumes. By minimizing in-field handling, more samples can be collected across an aquatic system in a
292 given timeframe, minimizing temporal sampling bias. For instance, we were able to sample roughly half
293 the length of the Cuyahoga River in a single day with a single sampling team using kayaks. The speed of
294 sampling is also an important parameter as it enables studies investigating the consequences of specific
295 events such as rainfall, ecological disasters (e.g., oil, chemical or pollution spills) or even transient
296 ecological events such as fish spawning.
297 Finally, this approach is cost-efficient: in our study of the Cuyahoga River, we characterized eDNA
298 amplified from mammals, fish, birds, amphibians, arthropods, vascular plants, bryophytes and many
299 microorganisms in 91 water samples for approximately US $2000 (including DNA extraction, PCR
300 reagents and sequencing costs, see **Supplemental Table 4** for details). This high level of multiplexing

(across samples and across taxa) decreases the burden of the next-generation sequencing price, but, thanks to the tremendous sequencing output, still leaves enough reads to rigorously characterize the composition of each sample. The use of an inexpensive PCR to add the sequencing and barcoding adapters also dramatically reduces the cost compared to generation of typical next-generation sequencing libraries (Ficetola et al. 2008; Zimmermann et al. 2011; Kelly et al. 2014; Cowart et al. 2015). Note that the cost and efficiency could potentially be further improved by multiplexing primer pairs in the PCR reaction (Hajibabaei et al. 2012).

However, there are also several limitations to our study. While detection of eDNA is by nature stochastic, the small sample size used in our study likely increases this randomness and increases the chance of false negatives. By optimizing PCR conditions and increasing the number of cycles, one can limit the chance of having a DNA template present in a given sample failing to amplify. Nonetheless, a small sample of water may not entirely capture the diversity present in the environment, especially for species represented by few DNA molecules in the environment. While the amplification of a species' DNA sequence is a clear evidence of the presence of this species (assuming a low level of cross-contamination), the failure to detect an organism is not a proof of its absence in the environment as the false negative rate is likely to be important. For example, while 20 out of 38 samples in the upper portion contained carp DNA, we would not be able to rule out the presence of carp at the other sampled locations in this part of the river. The use of universal primer pairs also contributes to the likely high false negative rates: since our assay is based on primers that amplify multiple taxa rather than being species-specific, it is possible that DNA sequences from one species completely overwhelm the signal originating from rarer eDNA templates amplified simultaneously. This effect could be magnified if there are differences in amplification efficiencies (Elbrecht and Leese 2015; Pinol et al. 2015). For example, it is possible that some samples that did not yield carp sequences actually contained carp eDNA but that

these molecules remained undetected due to the abundance of other fish eDNA (e.g., Silver Redhorse) in these samples. The presence of multiple amplification targets is also one reason why this assay is non-quantitative: the number of reads obtained for one species in a given sample is not only determined by how much DNA from that species is present in the sample, but also by how many other amplifiable species are present. Therefore, one cannot directly compare species read counts between samples to determine relative abundance (Kelly et al. 2014). In these regards, it is important to emphasize that the present approach does not replace classical eDNA studies targeting a single species but is designed for enabling broader ecological survey (that may guide further in-depth investigations).

Another limitation regards the interpretation of the results from this type of experiment. First, because we performed the species identification of sequences using blast, it is affected by the content of the NCBI database and the reliability might vary for different taxa. For instance, our analyses revealed, in one sample, the presence of DNA sequences most similar to a Taiwanese vole, which is not present near the Cuyahoga River, but with only 93.4% identity. These sequences likely originate from a local species of vole (closely related to the Taiwanese vole) that has not been sequenced for the 16S rRNA gene. This illustrates that matches with low percent identity blast hits need to be cautiously interpreted. On the other hand, this example also shows that this method can be used to identify organisms even if they have not been previously sequenced for the locus of interest, as long as a closely related species is present in the database, but that rigorously identifying the actual species present will require further analyses. In addition, when considering best match for a given sequence, the identification should be considered in the context of the other sequences amplified from the sample. For instance, in our analyses we identified sequences that best matched unexpected fowl species (e.g., *Gallus lafayetii* and *Gallus sonneratii*). However, these DNA sequences were only observed in small numbers and in samples that had a much larger numbers of *Gallus gallus* sequences. The most likely explanation for these

347 sequences is that these reads derived from PCR or sequencing errors from *Gallus gallus* templates: the
348 chicken sequences were so abundant in these samples that even rare errors led to a number of reads
349 sufficient to pass through our stringent filtering criteria. The same phenomenon generated a handful of
350 sequences most similar to monkey or ape sequences in samples with many human reads.
351 The study of environmental DNA using massively-parallel sequencing technologies enables
352 characterization of species biodiversity in a simple, high-throughput and cost-effective manner. Our
353 results reveal that a sampling protocol relying on small sample volume enables the preliminary
354 evaluation of complex environments. The ability to characterize a very diverse range of taxa from a
355 single sample in a high-throughput manner will allow future studies to expand the scope of the
356 biodiversity studied and to explore complex ecological interactions among species. Additionally, the
357 identification of eDNA of local terrestrial flora and fauna in river samples provides a simple way to assess
358 the local diversity of environments adjacent to rivers or other water bodies. Overall, our findings
359 illustrate the sensitivity and utility of broad surveys of eDNA by deep sequencing and shows how this
360 approach can constitute an excellent foundation for ecological and environmental studies.
361
362 **Methods**
363 *In-silico evaluation of universal primer pairs*
364 To evaluate the amplification breadth and informativity of "universal" primers we developed
365 PrimerTree, an R package that performs the following functions for each primer pair provided by the
366 user:
367 (1) *In silico* PCR against a selected NCBI database
368 (2) Retrieval of DNA sequences predicted to be amplified
369 (3) Taxonomic identification of these sequences

(4) Multiple DNA sequence alignment

(5) Reconstruction of a phylogenetic tree

(6) Visualization of the tree with taxonomic annotation

PrimerTree utilizes the *in silico* primer search implemented in Primer-BLAST (Ye et al. 2012) by directly querying the NCBI Primer-BLAST search page. This allows access to all options available on the NCBI website. By default, PrimerTree searches the NCBI (nt) nucleotide database but alternative NCBI databases, such as only assembled genomes, or Refseq mRNA, can be queried. Note that when the proposed primers are degenerate, PrimerTree automatically tests up to 25 possible combinations (by default, with more possible) of primer sequences in Primer-BLAST and merges the results. The primer alignment results are then processed using the NCBI E-utilities (Sayers 2013) to i) retrieve DNA sequences located between the primers (i.e., the "amplified" sequences) and ii) obtain taxonomic information related to each DNA sequence using the NCBI taxonomy database (Ostell and McEntyre 2003). PrimerTree next aligns all "amplified" DNA sequences using Clustal Omega (Sievers et al. 2011) with a user configurable substitution matrix and reconstructs a Neighbor-Joining tree using the ape package (Paradis et al. 2004). Finally, PrimerTree displays the resulting phylogenetic tree using the ggplot2 package, labeling each taxon in a different color and adding the names of the main taxa using the directlabels package (Wickham 2009; Hocking 2013).

PrimerTree usually runs in less than five minutes, but the runtime varies greatly depending on the primer specificity (i.e., how many DNA sequences are "amplified"), the search parameters chosen, the current load on the NCBI servers and internet connection. In particular, each degenerate position in a primer will result in up to four times as many primers to be tested, which can considerably increase the runtime. To limit maximum runtime in this situation, PrimerTree randomly samples only a portion of the total primer permutations (25 by default). Changing the number of sampled permutations or including

393 all variants is possible. PrimerTree uses the plyr package extensively and has full support for any of the
394 parallel backends compatible with the foreach package (Wickham, 2011; Analytics, 2012). In particular,
395 parallel retrieval of the primer sequences from NCBI speeds up the total runtime considerably. Note that
396 parallel queries to Primer-BLAST are queued by NCBI's servers and are only processed once there is free
397 compute time.
398
399 *Sampling and DNA extraction*
400 We collected water samples from the upper (n=39), middle (n=16) and lower (n=24) Cuyahoga River
401 (**Figure 2** and **Supplemental Table 5** for details). These sections correspond to, respectively, an area with
402 lower population density; a section with more human presence, dams and water treatment plants; and
403 a heavily industrialized area. Each sample consisted of ~50 mL of surface water collected roughly one
404 meter from the bank of the river. In addition, we also collected additional samples from water sources
405 entering the river (n=12, **Supplemental Table 5**). The collections were performed during two sampling
406 times separated by a major rain episode that dramatically increased the river discharge (**Supplemental**
407 **Figure 2**).
408 We isolated DNA from each water sample using the following procedure adapted from previous studies.
409 (Chan et al. 2011; Deiner et al. 2015). We first mixed each water sample by inversion and transferred 40
410 mL to a new tube for centrifugation at 8,000 x *g* for 30 minutes at 4°C. We discarded the supernatant
411 and resuspended the pellet in 1 mL of ATL lysis buffer (DNeasy kit, Qiagen) supplemented with 0.47 %
412 Triton-X (Ricca Chemical Company), 7.88 mg of lysozyme (Fisher Scientific) and 19.2 units of lysostaphin
413 (Sigma Aldrich). We then incubated the samples at 37°C for 1 hour while shaking them. We digested
414 further by incubating 350 µl of each sample with 50 µl of proteinase K and 350 µl buffer AL (Qiagen) for
415 60 minutes at 56°C. Finally, we extracted DNA using Qiagen DNeasy columns according to the

manufacturer's instructions. We included two extraction controls and processed them identically and at the same time as the rest of the samples to monitor cross- or laboratory contamination. In addition, all experiments were performed in a laboratory where no eDNA or vertebrate DNA (aside from human and mouse) had been previously extracted or amplified.

DNA amplification and sequencing

We selected primers to amplify different taxa from the literature and evaluated *in silico* their specificity and information content using PrimerTree. We amplified DNA extracted from each sample (with two extraction controls and one PCR water control) using the following conditions: initial denaturation of 95°C for 15 minutes followed by 50 cycles of 95°C for 30 sec., 55°C for 30 sec. 72°C for 30 sec in 1X Quantitect mastermix (Qiagen) with 0.4 µM of each primer. Overall, we performed, on each sample, 12 independent DNA amplifications targeting Archaea (Baker et al. 2003), Mammals (Taylor 1996), Algae (Sherwood and Presting 2007), Amphibians (Thomsen et al. 2012), Birds (Epp et al. 2012), Fish (Thomsen et al. 2012), Bryophytes (Epp et al. 2012), Arthropods (Zeale et al. 2011), Copepods (Bissett et al. 2005), Diatoms (Zimmermann et al. 2011), Fungi (Epp et al. 2012) and vascular plants (Taberlet et al. 2007) (**Supplemental Table 6**). Each primer included a 5' tail for barcoding and Illumina sequencing (see below). We then pooled all 12 amplification products obtained from each water sample. We added Illumina adapter sequences and labeled each sample with an individual six nucleotide index (with each index distinct from all other index by at least two nucleotides) using primers targeting the 5' oligonucleotide tail in 10 cycles of PCR (initial denaturation of 94°C for 3 minutes followed by 10 cycles of 94°C for 45 sec., 56°C for 45 sec. 72°C for 45 sec in 1X buffer, 1.25U GoTaq (Promega), 2mM $MgCl_2$ and 2µM of each primer). We then pooled all indexed samples together and we sequenced the resulting library on an Illumina MiSeq to generate 10,507,986 paired-end reads of 250 bp.

439

440 *DNA sequence analysis pipeline*

441 We used custom PERL scripts to retrieve the index information and identify and trim the amplification
442 primer sequences. We discarded any DNA sequence shorter than 50 bp, with the exception of
443 sequences amplified with plant and bryophyte primers for which short amplification products were
444 expected (see **Supplemental Table 1**). We also discarded any trimmed read pair for which the difference
445 in sequence length was greater than 5 bp between the two paired-end reads to eliminate any reads
446 where primers were not found in both reads. We then merged the paired reads into a single consensus
447 DNA sequence using PANDAseq with default parameters (Masella et al. 2012). We used Mothur (Schloss
448 et al. 2009) to identify unique DNA sequences and counted how many reads carried each unique DNA
449 sequence.

450 While all raw sequences are freely available online (accession number SRP058316), we only describe
451 here the analyses of macro-organism DNA sequences for sake of simplicity (microorganism sequences
452 can be easily analyzed using standard packages such as those implemented in QIIME, (Caporaso et al.
453 2010)). For macro-organisms such as mammals, many species have been sequenced for the locus of
454 interest and if not, a closely related species is likely present in the NCBI database (but see also below).
455 Therefore, to analyze DNA sequences from macro-organisms – mammals, amphibians, birds,
456 bryophytes, arthropods, copepods and plants – we used Blast (Altschul et al. 1990) to directly identify
457 the closest DNA sequences in the database and the likely species of origin. Briefly, we removed from our
458 analyses any DNA sequence observed in less than 10 reads (across all samples), as these likely represent
459 sequencing errors. We then compared each DNA sequence to all sequences deposited in the NCBI nt
460 database using Blastn (excluding uncultured samples) and only considered matches with greater than
461 90% identity over the entire sequence length. We then retrieved taxonomic data of all best match(es)

462 for each sequence from NCBI. If multiple species/genera/etc. matched a single sequence, all species
463 names were assigned to the sequence. We conducted further analyses at the species level for all taxa.
464
465 **Data Accessibility**
466 All sequencing data generated in this study are available through the SRA at NCBI (Accession
467 #SRP058316 (data to be released upon publication)). The R package PrimerTree is freely available at
468 http://github.com/jimhester/primerTree.
469
470 **Acknowledgments**
471 This work was supported by Cleveland Clinic funds to DS.
472
473 **Author Contributions**
474 MVC, STS and DS conceived the study. All authors contributed to the sample collection. AS performed
475 the DNA extraction and PCR amplification. MVC analyzed the data. JH implemented PrimerTree. MVC
476 and DS drafted the manuscript. All authors have read and approved the final version.
477
478 **Disclosure Declaration**
479 The authors have no conflicts of interest to declare.
480
481 **Tables**
482
483 **Table 1: Sequencing and Sequence analysis summary**

Taxon targeted	Locus	Reads generated	Reads after QC	Unique sequences	Unique seqs. w/ >=10 reads	Reads w/ common seq.	Species or OTUs identified	Seq. length (min-max)
Mammals	mt16S rRNA	1,384,149	1,352,524	13,238	786	1,328,072 (98.2%)	65	95 bp (76 - 118)
Amphibian	mt-Cytb	6,517	6,213	225	20	5,724 (92.1%)	2	68 bp (68 - 68)
Birds	12S rRNA	79,549	64,144	873	136	62,433 (97.3%)	123	53 bp (50 - 211)
Arthropods	COI	299,714	42,987	5,043	238	34,855 (81.1%)	25	162 bp (156 - 166)
Copepods	28S rRNA	502,972	79,486	13,252	368	59,753 (75.2%)	5	207 bp (53 - 368)
Fish	mt-Cytb	280,184	83,123	1,918	231	80,393 (96.7%)	5	90 bp (76 - 192)
Bryophytes	trnL	253,053	225,780	1,728	262	223,153 (98.8%)	32	52 bp (25 - 55)
Plants	trnL	1,954,381	1,610,089	33,584	3,468	1,554,086 (96.5%)	236	48 bp (17 - 80)

Table 2: Macro-organisms identified in the Cuyahoga River.

Common name	Scientific name	Percent Identity	#Samples positive
	Mammals		
Cow	Bos taurus/Bos indicus/Bos primigenius/Bos javanicus	100	22
Dog	Canis lupus/Canis aureus	100	22
Deer	Odocoileus virginianus/Odocoileus hemionus/Mazama americana	100	2
Sheep	Ovis aries/Ovis canadensis/Ovis vignei/Ovis dalli	98.91	1
Beaver	Castor canadensis	100	15
Cat	Felis catus/Felis silvestris	100	3
Groundhog	Marmota himalayana/Marmota monax	97.78	2
Skunk	Mephitis mephitis	98.89	1
Mouse	Mus musculus	100	33
Mink	Neovison vison	100	3
Muskrat	Ondatra zibethicus	100	17
Raccoon	Procyon lotor	100	5
Rat	Rattus norvegicus	100	2
Squirrel	Sciurus carolinensis	97.83	4
Pig	Sus scrofa/Sus barbatus/Sus philippensis	100	25
Chipmunk	Tamias striatus/Tamias sonomae	100	3
Taiwan vole	Microtus kikuchii	93.41	1
	Reptiles and amphibians		
Northern two-lined salamander	Eurycea bislineata	94.12	1
Box turtle	Terrapene carolina	100	1
	Birds		
Teal	Anas poecilorhyncha/Anas crecca/Anas platyrhynchos/Anas acuta/Anas clypeata/Mareca falcata/Cygnus melancoryphus/Histrionicus histrionicus/Ptaiochen pau/Thambetochen chauliodous/Anas hottentota/Anas querquedula/Anas cyanoptera	100	4
Swan/Goose	Cygnus columbianus/Anser fabalis/Anser anser/Cygnus olor/Branta bernicla/Callonetta leucophrys/Cygnus atratus/Anser indicus/Anser albifrons/Anser cygnoides/Cygnus cygnus/Branta sandvicensis/Anser rossii/Branta canadensis/Anser canagica/Cygnus buccinator	100	1
Chicken	Gallus gallus/Gallus sonneratii	100	8
Sparrow	Passer domesticus/Passer montanus/Chlorospingus canigularis	100	1
Ring-billed gull	Larus delawarensis	100	1
Wild turkey	Meleagris gallopavo	100	1
Great tit	Parus major	100	1
Thrush	Turdus sp.	100	1
	Fish		
Central stoneroller	Campostoma anomalum	96.55	10
White sucker	Catostomus macrocheilus/Catostomus commersonii	100	12
Grass/Black carp	Ctenopharyngodon idella/Mylopharyngodon piceus	100	1
Cyprinella spiloptera	Cyprinella spiloptera	100	9
Common carp	Cyprinus carpio/Carassius gibelio/Cyprinus multitaeniata/Cyprinus melanes	100	31
Cypress minnow	Hybognathus hayi	94.78	1
Northern hogsucker	Hypentelium nigricans	100	10
Shiner	Luxilus chrysocephalus/Luxilus cornutus	100	4
Spotted sucker	Minytrema melanops	100	4
Redhorse	Moxostoma cervinum/Moxostoma anisurum	95.65	16
Bluehead chub	Nocomis leptocephalus	95.65	1
Emerald shiner	Notropis atherinoides	99.13	2
Blacknose dace	Rhinichthys atratulus	100	4
Bluntnose minnow	Pimephales notatus	100	8
	Arthropods		
Non-biting midge/Fly	aff. Cyrtona/Tachinidae gen./Orthocladius sp./Orthocladiinae sp./Fannia serena	90.45	1
Cranefly	Antocha sp.	100	2
Fruitfly	Capparimyia aenigma/Scaptomyza frustulifera/Scaptomyza remota	90.45	2
Non-biting midge	Cricotopus bicinctus	100	1
Fruitfly	Drosophila medioimpressa	91.72	1
Whirligig beetle	Gyrinidae sp.	95.54	3
Non-biting midge	Microtendipes pedellus	99.36	2
Allegheny crayfish	Orconectes obscurus	100	1
Non-biting midge	Polypedilum convictum	100	1
Daphnia	Simocephalus cf.	99.36	3
Black fly	Simulium luggeri	100	1
Horsefly	Tabanus sp./Hybomitra zonalis	93.63	1
Cranefly	Tipula paludosa	96.82	1
Non-biting midge	Tribelos sp.	100	2
Non-biting midge	Xenochironomus xenolabis	98.09	1
	Copepods		
	Hemidiaptomus maroccanus/Hemidiaptomus amblyodon/Hemidiaptomus ingens/Hemidiaptomus roubaui	94.1	1
	Macrocyclops distinctus	93.73	2

488

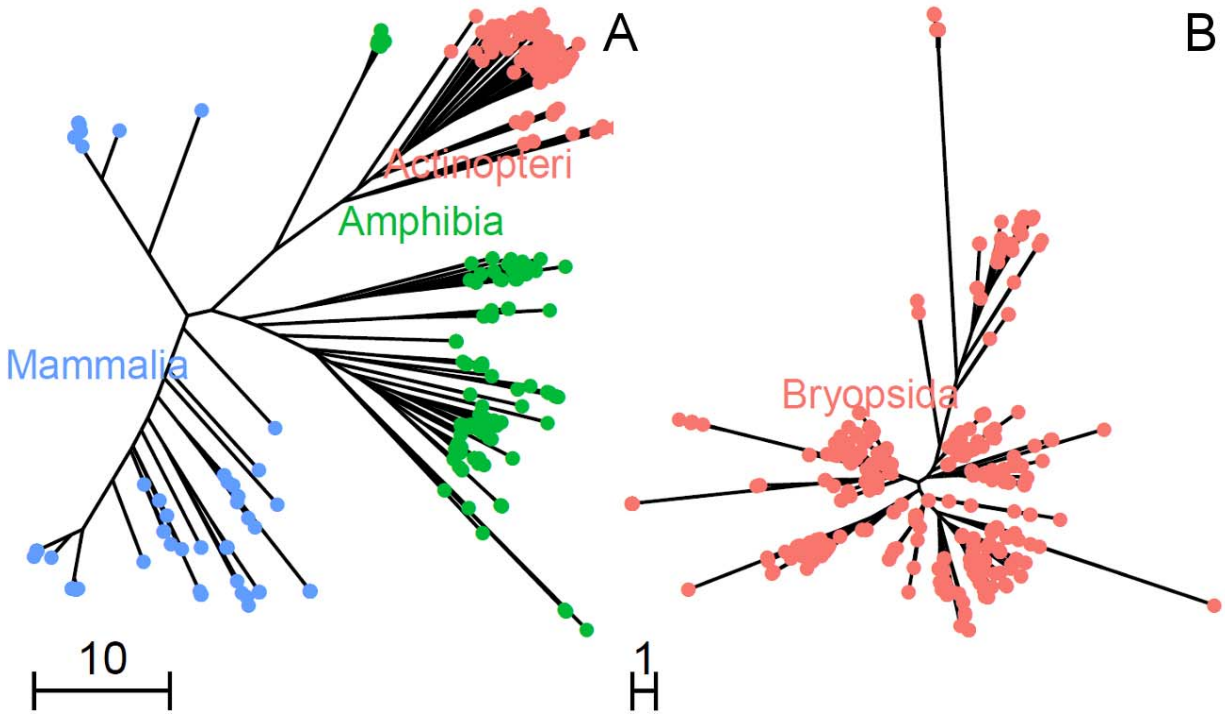
489 **Figure legends**

490

491

492

493



494

495

496 **Figure 1. Example of PrimerTree results.** The figure shows phylogenetic trees annotated at the class

497 level for (**A**) the mammalian 16S rRNA and (**B**) bryophyte trnL primer pairs. The complete PrimerTree

498 results for all 12 primers used in this study are presented in Supplemental Figure 1.

499

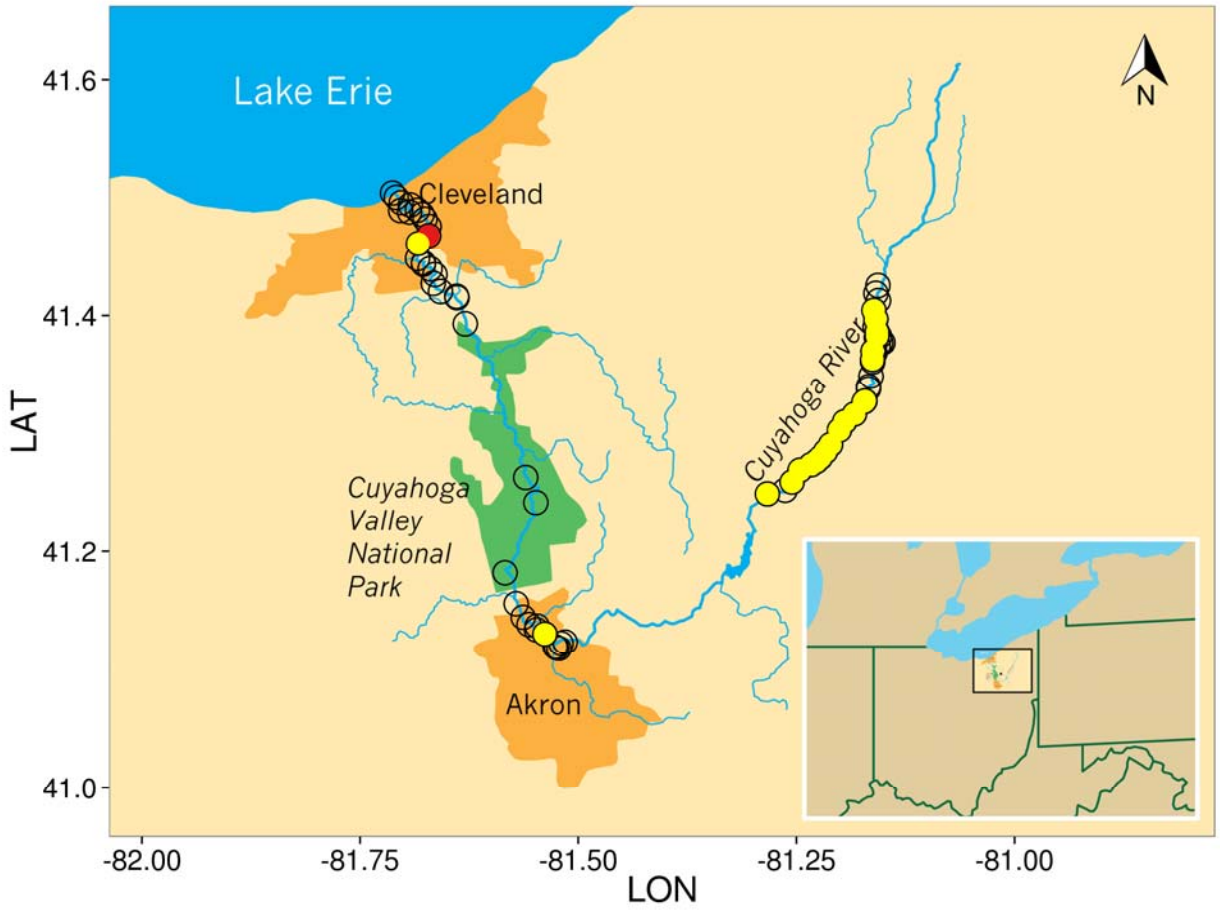
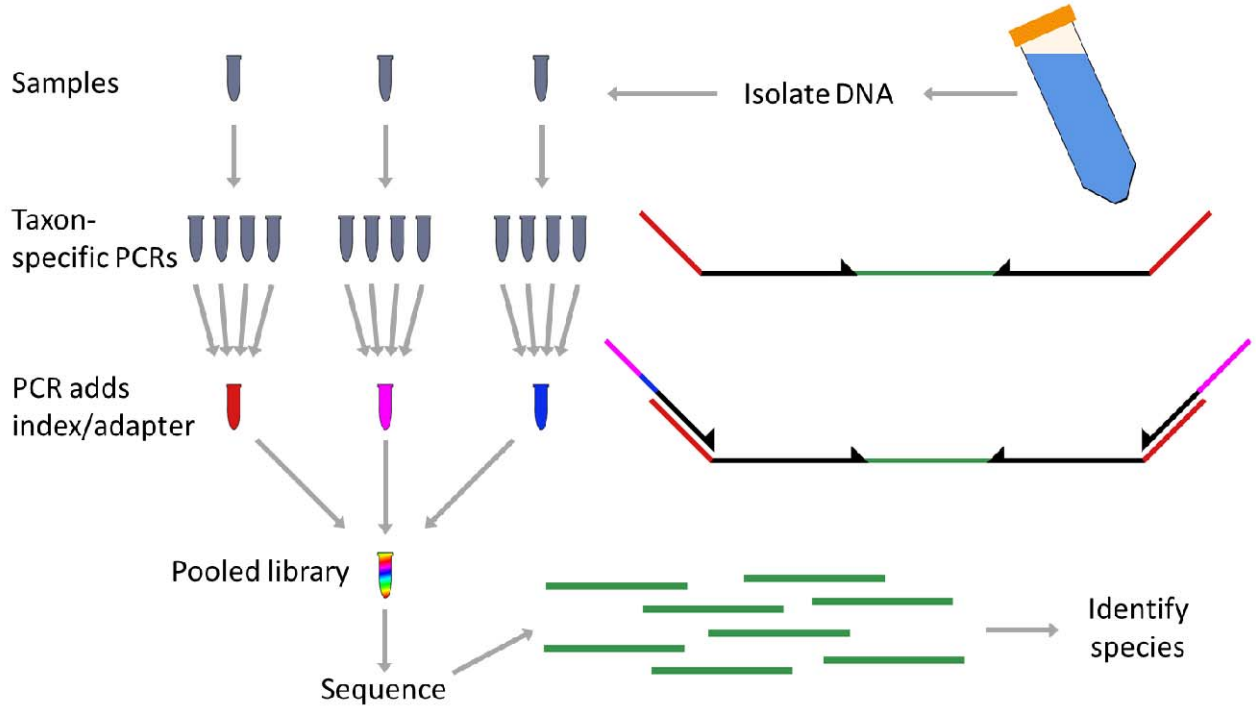


Figure 2. Geographic locations of water samples positive for Silver Redhorse and the Asian Carp eDNA. Each circle shows the location of a sampled site. The red circle indicates the location of the sample positive for Asian carp (*Ctenopharyngodon idella* or *Mylopharyngodon piceus*). Yellow circles are samples positive for Silver Redhorse (*Moxostoma cervinum* or *Moxostoma anisurum*).



506

507 **Figure 3. Experimental workflow.** We first isolated DNA from 40 ml of river water. We then amplified
508 each sample with 12 taxon-specific primer sets and pooled a portion of each PCR for each sample. Each
509 primer had a 5' tail to allow a second PCR which added Illumina adapter sequence and an individual
510 index. We then pooled all barcoded samples and sequenced the library on a MiSeq. We used the
511 sequence information to identify species of origin for DNA fragments isolated from the original samples.

512
513
514 **Supplemental Figure and Table Legends**
515
516
517

518 **Supplemental Figure 1: PrimerTree results for the 12 primer pairs**. Each primer set is denoted at the
519 top of the page. Each page is divided into eight sections that show different taxonomic levels, with the
520 points colored according to the taxa within that level. Levels with few taxa have colored text labels. The
521 guide below each tree shows the scale in number of nucleotide differences.

Supplemental Figure 2: USGS discharge at the days of sample collection. The figure shows the discharge of the Cuyahoga River at Hiram Rapids (Upper Cuyahoga, OH) in cubic feet per second (y-axis, in log scale) between July 9 and July 24 2013. The days of collection are indicated by the orange arrows.

Supplemental Table 1: Primer length and specificity based on observed and *in silico* data (in brackets)

Supplemental Table 2: Fish species identified using the mammalian 16S rRNA and fish mt-Cytb primers

Supplemental Table 3: Raw read counts and species absence/presence calls for macroorganisms

Supplemental Table 4: Experimental expense summary

Supplemental Table 5: Sample descriptions

Supplemental Table 6: Primer sequences

536 **References**

537

538 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*
539 **215**(3): 403-410.

540 Andersen K, Bird KL, Rasmussen M, Haile J, Breuning-Madsen H, Kjaer KH, Orlando L, Gilbert MT,
541 Willerslev E. 2012. Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Mol*
542 *Ecol* **21**(8): 1966-1979.

543 Baker GC, Smith JJ, Cowan DA. 2003. Review and re-analysis of domain-specific 16S primers. *J Microbiol*
544 *Methods* **55**(3): 541-555.

545 Biggs J, Ewald N, Valentini A, Gaboriaud C, Dejean T, Griffiths RA, Foster J, Wilkinson JW, Arnell A,
546 Brotherton P et al. 2015. Using eDNA to develop a national citizen science-based monitoring
547 programme for the great crested newt (Triturus cristatus). *Biol Conserv* **183**: 19-28.

548 Bik HM, Sung W, De Ley P, Baldwin JG, Sharma J, Rocha-Olivares A, Thomas WK. 2012. Metagenetic
549 community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and
550 shallow water sediments. *Mol Ecol* **21**(5): 1048-1059.

551 Bissett A, Gibson JAE, Jarman SN, Swadling KM, Cromer L. 2005. Isolation, amplification, and
552 identification of ancient copepod DNA from lake sediments. *Limnol Oceanogr-Meth* **3**: 533-542.

553 Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG,
554 Goodrich JK, Gordon JI et al. 2010. QIIME allows analysis of high-throughput community
555 sequencing data. *Nature Methods* **7**(5): 335-336.

556 Chan ER, Hester J, Kalady M, Xiao H, Li X, Serre D. 2011. A novel method for determining microflora
557 composition using dynamic phylogenetic analysis of 16S ribosomal RNA deep sequencing data.
558 *Genomics* **98**(4): 253-259.

559 Cowart DA, Pinheiro M, Mouchel O, Maguer M, Grall J, Mine J, Arnaud-Haond S. 2015. Metabarcoding is
560 powerful yet still blind: a comparative analysis of morphological and molecular surveys of
561 seagrass communities. *PloS one* **10**(2): e0117562.
562 Deiner K, Altermatt F. 2014. Transport distance of invertebrate environmental DNA in a natural river.
563 *PloS one* **9**(2): e88786.
564 Deiner K, Walser JC, Machler E, Altermatt F. 2015. Choice of capture and extraction methods affect
565 detection of freshwater biodiversity from environmental DNA. *Biol Conserv* **183**: 53-63.
566 Elbrecht V, Leese F. 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing
567 Primer Bias and Biomass-Sequence Relationships with an Innovative Metabarcoding Protocol.
568 *PloS one* **10**(7): e0130324.
569 Epp LS, Boessenkool S, Bellemain EP, Haile J, Esposito A, Riaz T, Erseus C, Gusarov VI, Edwards ME,
570 Johnsen A et al. 2012. New environmental metabarcodes for analysing soil DNA: potential for
571 studying past and present ecosystems. *Mol Ecol* **21**(8): 1821-1833.
572 Ficetola GF, Coissac E, Zundel S, Riaz T, Shehzad W, Bessiere J, Taberlet P, Pompanon F. 2010. An in silico
573 approach for the evaluation of DNA barcodes. *BMC Genomics* **11**: 434.
574 Ficetola GF, Miaud C, Pompanon F, Taberlet P. 2008. Species detection using environmental DNA from
575 water samples. *Biol Lett* **4**(4): 423-425.
576 Gibson J, Shokralla S, Porter TM, King I, van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M.
577 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical
578 arthropods through DNA metasystematics. *Proc Natl Acad Sci U S A* **111**(22): 8007-8012.
579 Goldberg CS, Pilliod DS, Arkle RS, Waits LP. 2011. Molecular detection of vertebrates in stream water: a
580 demonstration using Rocky Mountain tailed frogs and Idaho giant salamanders. *PloS one* **6**(7):
581 e22746.

582 Hajibabaei M, Spall JL, Shokralla S, van Konynenburg S. 2012. Assessing biodiversity of a freshwater
583 benthic macroinvertebrate community through non-destructive environmental barcoding of
584 DNA from preservative ethanol. *BMC Ecology* **12**: 28.
585 Hirai J, Kuriyama M, Ichikawa T, Hidaka K, Tsuda A. 2015. A metagenetic approach for revealing
586 community structure of marine planktonic copepods. *Mol Ecol Resour* **15**(1): 68-80.
587 Hocking T. 2013. Direct labels for multicolor plots in lattice or ggplot2.
588 Kelly RP, Port JA, Yamahara KM, Crowder LB. 2014. Using environmental DNA to census marine fishes in
589 a large mesocosm. *PloS one* **9**(1): e86175.
590 Mächler E, Deiner K, Steinmann P, Altermatt F. 2014. Utility of Environmental DNA for Monitoring Rare
591 and Indicator Macroinvertebrate Species. *Freshwater Sci* **33**(4): 1174-1183.
592 Mahon AR, Jerde CL, Galaska M, Bergner JL, Chadderton WL, Lodge DM, Hunter ME, Nico LG. 2013.
593 Validation of eDNA surveillance sensitivity for detection of Asian carps in controlled and field
594 experiments. *PloS one* **8**(3): e58316.
595 Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. 2012. PANDAseq: paired-end
596 assembler for illumina sequences. *BMC Bioinformatics* **13**: 31.
597 Minamoto T, Yamanaka H, Takahara T, Honjo MN, Kawabata Z. 2012. Surveillance of fish species
598 composition using environmental DNA. *Limnology* **13**(2): 193-197.
599 Ogram A, Sayler GS, Barkay T. 1987. The Extraction and Purification of Microbial DNA from Sediments. *J*
600 *Microbiol Methods* **7**(2-3): 57-66.
601 Ostell J, McEntyre S. 2003. The Taxonomy Project.
602 Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language.
603 *Bioinformatics* **20**(2): 289-290.

604 Pilliod DS, Goldberg CS, Arkle RS, Waits LP. 2014. Factors influencing detection of eDNA from a stream-
605 dwelling amphibian. *Mol Ecol Resourc* **14**(1): 109-116.
606 Pinol J, Mir G, Gomez-Polo P, Agusti N. 2015. Universal and blocking primer mismatches limit the use of
607 high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Mol Ecol*
608 *Resourc* **15**(4): 819-830.
609 Rees HC, Maddison BC, Middleditch DJ, Patmore JRM, Gough KC. 2014. REVIEW The detection of aquatic
610 animal species using environmental DNA - a review of eDNA as a survey tool in ecology. *J Appl*
611 *Ecol* **51**(5): 1450-1459.
612 Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM,
613 Remington K et al. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic
614 through eastern tropical Pacific. *PLoS biology* **5**(3): e77.
615 Sayers E. 2013. Entrez Programming Utilities Help.
616 Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks
617 DH, Robinson CJ et al. 2009. Introducing mothur: open-source, platform-independent,
618 community-supported software for describing and comparing microbial communities. *Appl*
619 *Environ Microbiol* **75**(23): 7537-7541.
620 Sherwood AR, Presting GG. 2007. Universal primers amplify a 23S rDNA plastid marker in eukaryotic
621 algae and cyanobacteria. *J Phycol* **43**(3): 605-608.
622 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et
623 al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using
624 Clustal Omega. *Mol Sys Biol* **7**: 539.
625 Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH. 2012. Environmental DNA. *Mol Ecol* **21**(8): 1789-1793.

626 Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C,
627 Willerslev E. 2007. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA
628 barcoding. *Nucleic Acids Res* **35**(3): e14.
629 Taylor PG. 1996. Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Mol Biol Evol*
630 **13**(1): 283-285.
631 Thomsen PF, Kielgast J, Iversen LL, Wiuf C, Rasmussen M, Gilbert MT, Orlando L, Willerslev E. 2012.
632 Monitoring endangered freshwater biodiversity using environmental DNA. *Mol Ecol* **21**(11):
633 2565-2573.
634 Treguier A, Paillisson JM, Dejean T, Valentini A, Schlaepfer MA, Roussel JM. 2014. Environmental DNA
635 surveillance for invertebrate species: advantages and technical limitations to detect invasive
636 crayfish Procambarus clarkii in freshwater ponds. *J Appl Ecol* **51**(4): 871-879.
637 Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ,
638 Detter JC et al. 2005. Comparative metagenomics of microbial communities. *Science* **308**(5721):
639 554-557.
640 Turner CR, Barnes MA, Xu CCY, Jones SE, Jerde CL, Lodge DM. 2014. Particle size distribution and optimal
641 capture of aqueous macrobial eDNA. *Methods Ecol Evol* **5**(7): 676-684.
642 Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. *Ggplot2: Elegant Graphics for Data*
643 *Analysis*: 1-212.
644 Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. 2012. Primer-BLAST: a tool to design
645 target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**: 134.
646 Zeale MR, Butlin RK, Barker GL, Lees DC, Jones G. 2011. Taxon-specific PCR for DNA barcoding arthropod
647 prey in bat faeces. *Mol Ecol Resourc* **11**(2): 236-244.

Zimmermann J, Jahn R, Gemeinholzer B. 2011. Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Org Divers Evol* **11**(3): 173-192.