

Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA

Kristy Deiner* | Mark A. Renshaw*  | Yiyuan Li | Brett P. Olds | David M. Lodge | Michael E. Pfrender

Department of Biological Sciences,
Environmental Change Initiative, University of
Notre Dame, Notre Dame, IN, USA

Correspondence

Kristy Deiner
Email: alpinedna@gmail.com
and
Mark A. Renshaw
Email: mrenshaw@hpu.edu

Present addresses

Kristy Deiner and David M. Lodge,
Department of Ecology and Evolutionary
Biology, Atkinson Center for a Sustainable
Future, Cornell University, Ithaca, NY, USA
Mark A. Renshaw and Brett P. Olds,
Oceanic Institute, Hawai'i Pacific University,
Waimanalo, HI, USA

Funding information

U.S. Department of Defense's Strategic
Environmental Research and Development
Program, Grant/Award Number: RC-2240

Handling Editor: Douglas Yu

Abstract

1. As environmental DNA (eDNA) from macro-organisms is often assumed to be highly degraded, current eDNA assays target small DNA fragments to estimate species richness by metabarcoding. A limitation of this approach is the inherent lack of unique species-specific single-nucleotide polymorphisms available for unequivocal species identification.
2. We designed a novel primer pair capable of amplifying whole mitochondrial genomes and evaluated it *in silico* for a wide range of ray-finned fishes (Class: Actinopterygii). We tested the primer pair using long-range PCR and Illumina sequencing *in vitro* on a mock community of fish species assembled from pooling genomic DNA extracted from tissues. *In situ* we utilized long-range PCR and Illumina sequencing to generate fragments between 16 and 17 kb from eDNA extracted from filtered water samples. Water samples were sourced from a mesocosm experiment and from a natural stream.
3. We validated our method *in silico* for 61 orders of Actinopterygii; we successfully sequenced mitogenomes *in vitro* from all six species in our mock community. *In situ* we recovered mitogenomes for all species present in our mesocosms. We additionally recovered mitogenomes from 10 of 12 species caught at the time of water sampling and two species previously only detected from eDNA metabarcoding of short DNA fragments from a natural stream.
4. Successful amplification of large fragments (>16 kb) from eDNA demonstrates that not all eDNA is highly degraded. Sequencing whole mitogenomes from filtered water samples will alleviate many problems associated with identification of species from short-fragment PCR amplicon-based methods.

KEYWORDS

Actinopterygii, eDNA, mitogenome sequencing

*Share first authorship.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2017 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society

1 | INTRODUCTION

The use of DNA found in the environment (eDNA) to catalogue biodiversity is gaining momentum (Creer et al., 2016). From surveying the three domains of life in soils (Drummond et al., 2015) to whales in the ocean (Foote et al., 2012), biodiversity information is being produced on unprecedented scales.

Perhaps because the use of eDNA to detect species was partially inspired by the field of ancient DNA (Thomsen & Willerslev, 2015), researchers assumed that eDNA was highly degraded. Because of this assumption, and coupled with current sequence length limitations of next-generation sequencers (e.g. Illumina MiSeq), researchers have focused on producing small fragments (c. 50–400 base pairs in length) using a PCR amplicon approach to characterize macro-organismal species richness (Olds et al., 2016; Valentini et al., 2016). However, reliance on short-fragment PCR amplicons from eDNA limits the current utility of the method because species-level assignments are often not possible for short reads (Deiner, Fronhofer, Mächler, Walsler, & Altermatt, 2016; Port et al., 2016).

While it may be true that some of the eDNA in environmental samples is degraded, evidence that not all eDNA is degraded has emerged. In a recent study based on water from a fish pond, most of the eDNA detected was from particles that ranged in size from 1 to 10 μm , consistent with the presence of intact tissues or cells in aquatic environments (Turner et al., 2014). This result suggests that eDNA for species currently occupying a habitat is not primarily free DNA suspended in solution, but that it could be cellular or membrane bound DNA in a tightly coiled or circular state and comparatively safe from degradative processes (Torti, Lever, & Jørgensen, 2015). We therefore hypothesized that it should be possible to long-range PCR amplify and sequence whole macro-organismal mitochondrial genomes (mitogenomes) from DNA isolated from water samples.

To test this hypothesis, we designed a novel primer set in the 16S region of the mitochondrial genome that is nearly 95% conserved at the sequence level across the class of Actinopterygii (ray-finned fishes) and could be used for long-range PCR amplification of fish mitogenomes from water samples. Long-range PCR is a viable option for the enrichment of whole mitogenomes from environmental samples because in a single amplification it can produce a fragment that encompasses the entire mitogenome (Zhang, Cui, & Wong, 2012). The amplification of the entire mitogenome in a single PCR has inherent time and cost advantages over amplification of multiple fragments, reduces the potential complications of nuclear-encoded mitochondrial pseudogenes (NUMTs) and avoids the rearrangement of targeted priming sites in mitogenomes with altered gene order (Cameron, 2014). However, the success of long-range PCR amplifications depends on the presence of relatively high-quality and high-molecular-weight DNA.

Sequencing whole mitogenomes from eDNA samples could vastly improve species assignment capabilities because full-length barcodes, such as the cytochrome c oxidase I (COI) region for animals (Hebert, Ratnasingham, & de Waard, 2003), could be recovered and used for the identification of species in communities. Other mitochondrial

genes typically used in phylogeography, systematics and conservation genetics could also be recovered in their entirety to provide a non-destructive method for sampling whole communities for studies related to community phylogenetics and conservation.

In this study, we validate a method to amplify and sequence entire mitochondrial fish genomes from water samples (Figure 1). In silico we tested newly designed mitochondrial primers and in vitro performed long-range PCR and next-generation sequencing on resulting amplifications using a mock community amassed from tissue extracted DNA. In situ we validated the method using water samples collected from a mesocosm experiment with an assembled fish community of eight species, and from a natural stream known to have 12 species present at the time of sampling. DNA extractions from the mesocosm and stream samples were the same as those used for two previous eDNA studies of fish communities (Evans et al., 2016; Olds et al., 2016), allowing us to compare the species richness estimated from a short-fragment PCR amplicon approach to that of using long-range PCR and whole mitogenome sequencing from water samples.

2 | MATERIALS AND METHODS

2.1 | Primer design and in silico evaluation

A batch download of Actinopterygii 16S sequences from GenBank was aligned using the PartTree algorithm in MAFFT version 7 (Kato & Standley, 2013). The alignment was viewed in BioEdit (Hall, 1999) and conserved regions identified by eye. These regions were then evaluated in PRIMER3 (Untergasser et al., 2012) for putative primer pairs. Primers Actinopterygii16SLRprc_F (5'-CAGGACATCCTAATGGTGCAG-3') and Actinopterygii16SLRprc_R (5'-ATCCAACATCGAGGTCGTAAC-3') were designed to be immediately adjacent to one another to PCR amplify nearly the entire mitogenome in a single reaction. The only part of the mitogenome not amplified is the 43-bp area covered by the Actinopterygii16SLRprc priming region.

To evaluate the potential taxonomic coverage of the primers, they were concatenated into a single sequence in the same reading frame (i.e. the reverse primer was reverse-complemented before it was concatenated) resulting in a 43-bp fragment. The fragment was then aligned with Actinopterygii fish mitogenomes available on MitoFish v3.02 (Figure 1a) (Iwasaki et al., 2013). One thousand five hundred and twelve fish species in the class Actinopterygii, representing 62 orders and 310 families, were considered (Appendix S1). BLASTN (blastall 2.2.26) was used to align primer fragments to mitogenomes with output in tabular format, e value = 10^{-4} , and without low-complexity filter ($-m$ 8 $-F$ $-e$ 1e-4) to ensure a single hit for each mitogenome (Altschul, Gish, Miller, Myers, & Lipman, 1990). Mismatches between each primer and the reference genomes from MitoFish are given in Appendix S1.

2.2 | In vitro evaluation using mock community

To test the performance of the primers, laboratory methods and bioinformatic pipeline, an in vitro test was performed using a mock

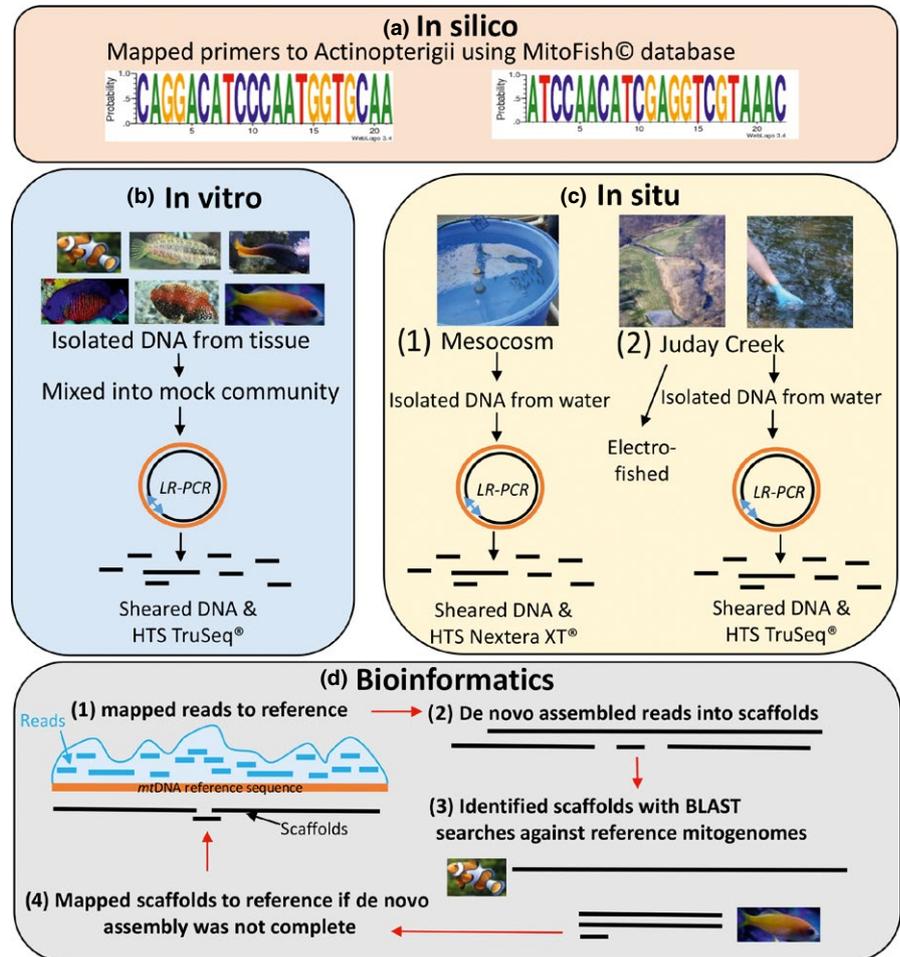


FIGURE 1 Methods overview for laboratory workflow used to design and test long-range PCR (LR-PCR) for sequencing of whole mitochondrial genomes from environmental DNA. Each box (a–d) represents the steps used in silico, in vitro and in situ to validate the method. High-throughput sequencing on the Illumina MiSeq is abbreviated as HTS

community sample at a concentration of 0.6 ng/ μ l (0.1 ng/ μ l from each species) of tissue-derived DNA from six Indo-Pacific marine fishes: *Amphiprion ocellaris*, *Salarias fasciatus*, *Ecsenius bicolor*, *Centropyge bispinosa*, *Pseudanthias dispar* and *Macropharyngodon negrosensis* (Figure 1b) (Olds et al., 2016). DNA extractions for each of the six mock community species were performed with the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany), following the protocols as outlined by the manufacturer with the exception that final elutions were made with 200 μ l of 1X TE buffer, low EDTA (USB Corporation, Cleveland, OH). The tissue-derived DNA was quantified with the Qubit® dsDNA BR Assay Kit (Life Technologies, Carlsbad, CA), and equal nanogram amounts from each of the six species were combined into a single mock community DNA extract.

PCR amplification of the mock community sample included 25 μ l of LongAmp® Taq 2X Master Mix (New England BioLabs, Ipswich, MA), 20 picomoles of each primer (forward and reverse), 5 μ l of mock community DNA extract and sterile molecular grade water to bring the total volume to 50 μ l. Cycling parameters included an initial denaturation step at 94°C for 30 s; 35 cycles of denaturation at 94°C for 30 s, annealing at 62°C for 1 min, extension at 65°C for 14 min and 10 s; and a final extension step at 65°C for 10 min. To sequence the amplified PCR product on the Illumina MiSeq, the entire 50 μ l PCR amplification was electrophoresed on a 0.75% agarose gel, a fragment of expected size (16–18 kb) for the mitogenomes was excised from the

gel with a razor blade, the PCR product was purified with the QIAquick Gel Extraction Kit (Qiagen) and eluted in 50 μ l AE buffer. The resulting DNA was quantified with the Qubit® dsDNA BR Assay Kit (Life Technologies).

Based on the Qubit reading for the cleaned PCR product, approximately 200 ng was diluted in a total volume of 52.5 μ l; the PCR product was then sheared with a S220 Focused-ultrasonicator (Covaris, Woburn, MA). Preparation of the mock community sample for sequencing on the Illumina MiSeq followed the manufacturer's suggested protocol as outlined for the TruSeq Nano LT Sample Prep Kit (low-sample protocol) for a 550-bp insert size (Illumina, San Diego, CA). Sequencing was performed with the MiSeq Reagent Kit v3 (600 cycles; Illumina), producing paired end reads each with a length of 300 bp.

2.3 | In situ evaluation from water samples

Environmental DNA used for this study was extracted from filtered water samples that were utilized in two previous studies: two samples (High Density, Skewed Abundance, Tank 3 [HS3]; High Density, Even Abundance, Tank 3 [HE3]) from a mesocosm experiment (Evans et al., 2016) and eight samples (R1-D, R1-U, R2-D, R2-U, R3-D, R3-U, R4-D, R4-U) from a stream survey (Figure 1c) (Olds et al., 2016). Details of the collection, filtration and DNA extraction have been

previously described in Evans et al. (2016) and Olds et al. (2016). DNA extracts were treated with Zymo OneStep™ PCR Inhibitor Removal (Zymo, Irvine, CA) kits prior to amplification of mitogenomes via long-range PCR. Amplifications were done exactly as described in the in vitro section with the exception that 20 µg bovine serum albumin (VWR, Radnor, PA) was added to the PCR reaction.

Based on Qubit readings for the cleaned PCR-amplified products, 1 ng (in a total volume of 5 µl) for each of the mesocosm samples was prepared for sequencing on the Illumina MiSeq following the manufacturer's suggested protocol as outlined for the Nextera XT DNA Library Preparation Kit. For each of the eight stream samples, libraries were prepared as described for the in vitro test on the mock community. For all 10 samples (including mesocosm and stream), sequencing was performed with the MiSeq Reagent Kit v3 (600 cycles; Illumina), producing paired end reads each with a length of 300 bp. The two library preparation methods were used because we were unsuccessful in producing quality data for libraries generated from the stream samples utilizing the Nextera XT DNA Library Preparation Kit and found the TruSeq Nano LT Sample Prep Kit method ultimately to be more reliable for the preparation of Illumina libraries from eDNA amplified whole mitogenomes.

2.4 | Bioinformatic filtering, mapping and de novo assembly

Raw reads were quality filtered by removing Illumina sequencing adaptor, low-quality sequences with average quality less than Q20 in any 10-bp window and short sequences with length less than 50-bp using Trimmomatic v0.32 (Bolger, Lohse, & Usadel, 2014) with "ILLUMINACLIP: MiSeq.adapter.fas:3:30:6:1:true SLIDINGWINDOW:10:20: MINLEN:50." Reads were used for alignment and assembly only when both forward and reverse reads passed quality filtering. After quality filtering, reads were merged to avoid counting the sequencing depth twice based on the same DNA fragment using USEARCH (Edgar, 2010) v8.1.1861_i86linux32 fastq_mergepairs command with minimum overlap length = 16 by default and maximum difference percentage = 1% (-fastq_maxdiffpct 0.01). If paired end reads could be merged, only the merged reads were used for mapping and de novo assembly analyses. If paired end reads could not be merged, both forward and reverse reads were mapped to mitogenomes and were used in the de novo assembly. Therefore, both merged and unmerged reads were mapped to the 6 mitogenomes represented in the mock community, the 8 mitogenomes from the two mesocosm densities (Evans et al., 2016) and the 14 mitogenomes from fish that were previously captured (or known to occur in the watershed) when water samples were taken from Juday Creek (Figure 1d) (Olds et al., 2016). BWA v0.7.15-r1140 (Li & Durbin, 2009) with a maximum difference in the seed (-k) equal to 2 and seed length equal to 32 was used for mapping. The missing probability was set under a 0.02 error rate (-n) to be equal to 0.06 which is consistent with a 97% similarity of OTU clustering used for the taxonomic assignment of amplicons from previous studies (Evans et al., 2016; Olds et al., 2016). SAI format files from both ends of the

reads were combined with "bwa sampe" command with maximum insert size equal to 1,000 bp. Merged reads were aligned as single ended reads with "bwa samse" command. Reads with mapping quality (MAPQ) <20 were removed. Only unique aligned reads (i.e. reads with "XT:A:U" in sam file) were used for reference mapping. Additionally, all reads from Juday Creek were combined before mapping to reference sequences. Samtools v1.2 (Li et al., 2009) command "stats" was used to calculate number of reads mapped for each reference. The mapping ratio was calculated as (number of mapped merged reads + number of mapped unmerged reads)/(the total number of merged and unmerged reads). Single nucleotide polymorphisms were determined from mapped reads as described in Data S1. BEDtools v2.25.0 (Quinlan & Hall, 2010) command "genomecov" was used for reference coverage and average sequencing depth calculation. Reference coverage was visualized using Geneious v9.1.5 and scaffolds from de novo assembly were mapped to the reference using default values in Geneious v9.1.5. To check for cross contamination during sample handling, reference mapping was done with all species used in this study for all libraries. Accession numbers for reference sequences are reported in Appendix S1.

For the mesocosm samples (HE3 and HS3, Evans et al., 2016), reads after Trimmomatic filtering were used for de novo assembly with the metagenomics assembler IDBA-UD v1.1.1 (Peng, Leung, Yiu, & Chin, 2012). Command "fq2fa" was used to transfer fastq format to fasta format and remove any read with an unknown base pair "N." Command "idba_ud -pre_correction -min_support 20" was used to assemble mitogenomes. For the mock community and Juday Creek samples, reads after Trimmomatic filtering were normalized based on kmer frequency (Crusoe et al., 2015) because the sequencing depth from these samples was too high. A custom Perl script was made to remove reads with sequencing depth higher than 50× based on 17-mer. Normalized reads and Perl script used for de novo assembly are provided on Dryad (<http://dx.doi.org/10.5061/dryad.q5gg0>). Normalized reads were assembled with the same parameters as the mesocosm samples with the metagenomics assembler IDBA-UD v1.1.1. Assembled scaffolds were then mapped to the reference genomes with BLAST+ (Camacho et al., 2009) to estimate coverage for each gene based on the references using a 95–97% sequence similarity cutoff.

2.5 | Long-range PCR compared with amplicon sequencing approach

Reads produced from independently sequenced gene fragments (16S, 12S and CytB) in previous studies (Evans et al., 2016; Olds et al., 2016) were mapped to the same reference mitogenomes used in the in situ evaluation with the same parameters as those used for long-range amplified mitogenomes. Only reads that passed quality control were mapped. Mitogenome coverage of the amplicons and sequence depth were evaluated and qualitatively compared to the values achieved from long-range amplified products for the entire de novo assembled gene. Differences in the estimated species richness between the two methods were also documented.

3 | RESULTS

3.1 | In silico evaluation of primers

For the in silico test of the Actinopterygii16SLRpcr primer fragment (i.e. encompassing both the forward and reverse primers), only one fish order averaged greater than two mismatches, while the other 61 orders averaged less than or equal to two mismatches (Appendix S2). The reverse primer binding site was more conserved than the forward primer binding site. Based on an average mismatch criterion of two or less mismatches, the Actinopterygii16SLRpcr primer region has the potential to long-range PCR-amplify entire mitogenomes for all orders except Osteoglossiformes. However, several of the species within many orders had mismatches of two or more base pairs in the first three base pairs of the 3' end that could lead to failed amplification of their DNA (Appendix S2). Primers were not assessed for similarity to other taxonomic groups outside of Actinopterygii.

3.2 | In vitro evaluation and de novo assembly of mitogenomes from mock community

For the mock community experiment, 77.6% of reads mapped to one of the six species included in the mixture of DNA (Appendix S3). Nearly whole mitogenomes were recovered for all species in the mock community using the reference mapping approach (Table 1). SNPs were called only in one species (Data S1).

Using the de novo assembly approach, a single scaffold was produced that covered the full-length of the reference genome for three of the six mock community species. For the other three species, two scaffolds for each species were recovered such that when combined they covered between 97.9% and 100% of their respective reference mitogenomes (Table 1). Additionally, the full length of genes commonly used in fish eDNA metabarcoding were recovered (Appendix S4). Of the reads that mapped to species not included in the mock community, four species that were only in the mesocosms and not observed in Juday Creek (*Campostoma anomalum*, *Fundulus notatus*, *Gambusia holbrooki* and *Pimephales promelas*) had no reads mapped to their reference (Appendix S5). Several species observed in Juday Creek showed moderate levels of mapping to their reference (0.04% of reads) even though they were not included in the mock community library (Appendix S5). However, the de novo assembly showed that no scaffolds for species from Juday Creek could be recovered above 1,900 bp in length (Appendix S5).

3.3 | In situ evaluation from water samples

For the high-density, even abundance (HE3) and skewed abundance (HS3) mesocosm communities, 65.2% and 75.1% of reads mapped to one of the eight species included in the mesocosm (Appendix S3). Whole mitogenomes for the eight fish were successfully recovered (99.5–99.9%, Table 1); however, one species, *P. promelas*, coverage was lower in the HE3 compared to HS3 treatment (86.2% vs. 99.5%). SNPs were called in most species (Data S1).

Using the de novo assembly approach on the mesocosm samples, long scaffolds were assembled for most species and were near complete mitogenomes (e.g. 16,524 bp for *Catostomus commersonii* in HS3, Table 1). De novo results were more consistent when abundance was even (HE3). *Pimephales promelas* de novo assembled scaffolds were short in both mesocosm communities and several species had many scaffolds that could not be merged into a single assembly without guidance of a reference sequence (Table 1). Genes commonly used in eDNA metabarcoding studies of fish were de novo assembled for all species even when the whole mitogenome could not be and ranged in coverage from 89% to 100% (Appendix S4).

For Juday Creek, 46.2% of reads could be uniquely mapped to 10 of 12 species confirmed present when water samples were collected and two additional species previously detected only from eDNA (Olds et al., 2016) (Appendix S3). Whole mitogenomes for 10 of the 12 species caught with electrofishing were successfully recovered from mapping reads to their references (94–100%, Table 1, Figure 2). The two species not recovered were *Lepomis macrochirus* and *Salmo trutta*. We additionally recovered mitogenomes from two species previously only detected from eDNA, but are known to occur in the watershed (*Cyprinus carpio* and *Micropterus salmoides*) (Table 1). SNPs were called for all species (Data S1).

The de novo assembly recovered nearly whole mitogenome scaffolds for about half the species (Table 1). The degree of coverage of these scaffolds to their reference sequences ranged from a single scaffold representing a nearly complete mitogenome to 29 smaller scaffolds covering nearly the entire mitogenome (Table 1, Figure 2). Genes commonly used in eDNA metabarcoding studies could be recovered from the de novo assemblies and the scaffolds overlapping these genes covered from 88% to 100% of the full gene length (Appendix S4). Some sequence reads from the Juday Creek sample mapped to the reference genomes of fish that were only used in the mock community and mesocosm experiment and were not present in Juday Creek (Appendix S5). In most cases the sequence depth and degree of coverage of these reads was low (0.6%–6.5%), with higher coverage of references from the mock community species (Appendix S5). None of these species had de novo assembled scaffolds longer than 2,100 bp in length.

Comparisons between long-range PCR-amplified mitogenomes and the short-fragment PCR amplicons generated from the same DNA extracts revealed some species (*S. trutta* and *L. macrochirus*) whole mitogenomes were not detected even though their smaller fragments were (Appendix S3). Additionally, two species detected previously only by eDNA, *M. salmoides* and *C. carpio* (Olds et al., 2016), were also detected; mapping revealed that their whole mitogenomes could be recovered and that long fragments could be de novo assembled (Table 1).

4 | DISCUSSION

We demonstrate that it is possible to sequence whole mitogenomes of fish from DNA extracted from water samples by coupling

TABLE 1 Long-range PCR (LRP) amplicon read mapping statistics for species in the mock community, two mesocosm communities and Juday Creek. Reference coverage for each species' mitogenome was evaluated in one of the two ways: each read mapped to a reference or de novo assembled (see methods). N is the number of individuals present at time of sampling (Juday Creek) or used in experiment (Mock or Mesocosm)

Species	Common name	Reference length (bp)	N	LRP mapped to reference			LRP de novo assembled		
				Uniquely mapped reads	Reference coverage (%)	Average sequence depth	Longest scaffold (bp)	Reference coverage (%)	No. of scaffolds
Mock community									
<i>Amphiprion ocellaris</i>	Common clownfish	16,649	1	4,559	99.4	71	16,652	99.6	1
<i>Centropyge bispinosa</i>	Twospined angelfish	16,772	1	798,978	100	11,704	13,633	97.9	3
<i>Ecsenius bicolor</i>	Flame tail blenny	16,534	1	288,587	100	4,265	13,528	98.7	3
<i>Macropharyngodon negrosensis</i>	Black leopard wrasse	16,889	1	539,213	100	8,053	16,890	100	1
<i>Pseudanthias dispar</i>	Dispar anthias	16,954	1	7,174	100	109	16,888	99.6	1
<i>Salarias fasciatus</i>	Jewelled blenny	16,496	1	318,871	100	4,728	13,867	98.2	4
Mesocosm high-density/even abundance (HE3)									
<i>Campostoma anomalum</i>	Central stoneroller	16,649	10	122,874	99.9	1,748	14,220	99.2	6
<i>Catostomus commersonii</i>	White sucker	16,622	10	8,167	99.6	118	16,524	99.4	1
<i>Cyprinus carpio</i>	Common carp	16,575	10	123,924	99.9	1,770	14,479	99.4	4
<i>Fundulus notatus</i>	Blackstripe topminnow	16,510	10	53,266	99.9	781	16,149	100	2
<i>Gambusia holbrooki</i>	Eastern mosquitofish	16,611	10	75,672	99.9	1,068	10,229	100	6
<i>Lepomis macrochirus</i>	Bluegill	16,489	10	16,677	99.9	231	16,509	100	1
<i>Pimephales promelas</i>	Fathead minnow	16,709	10	22,247	99.5	318	1,618	53.6	36
<i>Semotilus atromaculatus</i>	Creek chub	16,623	10	32,941	99.9	472	9,282	99.0	5
Mesocosm High-density/Skewed abundance (HS3)									
<i>Campostoma anomalum</i>	Central stoneroller	16,649	4	105,169	99.9	1,528	7,956	97.7	23
<i>Catostomus commersonii</i>	White sucker	16,622	4	39,822	99.9	587	11,304	99.1	10
<i>Cyprinus carpio</i>	Common carp	16,575	5	167,749	99.9	2,454	6,591	96.4	20
<i>Fundulus notatus</i>	Blackstripe topminnow	16,510	4	25,824	99.9	385	16,099	99.1	3
<i>Gambusia holbrooki</i>	Eastern mosquitofish	16,611	7	65,019	99.9	931	15,745	98.1	3
<i>Lepomis macrochirus</i>	Bluegill	16,489	18	34,285	99.9	486	15,580	97.4	7
<i>Pimephales promelas</i>	Fathead minnow	16,709	46	317	86.2	5	1,251	23.6	5
<i>Semotilus atromaculatus</i>	Creek chub	16,623	4	18,304	99.9	267	16,354	98.4	1
Juday Creek (all sites combined)									
<i>Ambloplites rupestris</i>	Rock bass	16,659	33	30,505	98.7	443	16,279	99.8	1
<i>Catostomus commersonii</i>	White sucker	16,622	27	1,367,916	100	20,988	6,644	89.5	14
<i>Cottus bairdii</i>	Mottled sculpin	16,529	295	1,296,452	100	18,926	15,873	97.2	3

(Continues)

TABLE 1 (Continued)

Species	Common name	Reference length (bp)	N	LRP mapped to reference			LRP de novo assembled		
				Uniquely mapped reads	Reference coverage (%)	Average sequence depth	Longest scaffold (bp)	Reference coverage (%)	No. of scaffolds
<i>Cyprinus carpio</i>	Common carp	16,575	0	145,017	100	2,166	16,322	99.7	2
<i>Etheostoma caeruleum</i>	Rainbow darter	16,588	1	10,940	100	165	14,429	97.9	3
<i>Etheostoma nigrum</i>	Johnny darter	16,579	89	1,266,274	100	18,643	6,651	97.7	11
<i>Lepomis cyanellus</i>	Green sunfish	16,485	48	309,297	100	4,574	15,882	98.8	3
<i>Lepomis macrochirus</i>	Bluegill	16,489	1	1	1.7	1	0		
<i>Micropterus dolomieu</i>	Smallmouth bass	16,488	19	286,733	100	4,242	5,915	95.1	12
<i>Micropterus salmoides</i>	Largemouth bass	16,484	0	232,161	100	3,434	12,879	95.7	8
<i>Oncorhynchus mykiss</i>	Rainbow trout	16,655	1	993,573	100	14,299	16,363	100	3
<i>Rhinichthys atratulus</i>	Western blacknose dace	16,651	49	8,611	93.9	143	16,036	98.6	2
<i>Salmo trutta</i>	Brown trout	16,677	1	10	5.5	2	0		
<i>Semotilus atromaculatus</i>	Creek chub	16,623	562	2,959,545	100	45,213	5,658	95.1	29

long-range PCR amplification and shotgun sequencing techniques. These results contradict the common assumption that eDNA from communities of living fishes in water is highly degraded (e.g. Bohmann et al., 2014). Our results show instead that some of the eDNA from macro-organisms currently inhabiting a water body remains intact at least at the mitochondrial genome size (c. 16 kb). Sequencing whole mitogenomes from water samples is a pioneering way to non-invasively assess communities of living macro-organisms. Additionally, it may make characterization of macro-organism eDNA relevant in studies of population and conservation genetics, systematics and phylogeography.

This methodological advance alleviates the burden of basing species identifications on short-fragment PCR amplicons. Most current eDNA studies from macro-organisms base their taxonomic assignments on DNA fragments of <150 base pairs (Port et al., 2016; Valentini et al., 2016). The amount of information in such a small fragment will always be limited, and in many cases assignments to the species level are not possible without additional information. With whole mitogenomes, entire barcode regions can be excised from a dataset and used for taxonomic assignment. For example, using the de novo assembly approach we recovered full-length mitochondrial genes (cytochrome oxidase I, cytochrome B, 12S and 16S) for the fish species in Juday Creek (e.g. *Semotilus atromaculatus*), even when their entire mitochondrial genome could not be assembled. Given the amount of missing diagnostic information in reference databases used for macro-organism taxonomic assignment (Shaw et al., 2016; Trebitz, Hoffman, Grant, Billehus, & Pilgrim, 2015), generating data with this method will allow use of any region of the mitogenome for which data exist to identify environmental sequences. This method will therefore be invaluable unless and until this void is filled with other methods such as genome skimming (Coissac, Hollingsworth, Lavergne, & Taberlet, 2016).

While the results from our in situ tests are promising, a number of questions remain about the particular environmental circumstances in which eDNA is left intact at the mitogenome level. In this study we used water sampled in September from a small third-order tributary to the St. Joseph River in Northwestern Indiana, USA. The mean annual discharge of Juday Creek is 0.44 m³/s samples (Olds et al., 2016). The average temperature on the day of sampling was 22.6°C (Shirey, Brueseke, Kenny, & Lamberti, 2016). We did not collect other variables at the time of sampling as it was not our goal to test these attributes here, but we encourage future studies to investigate additional conditions (e.g. temperature, turbidity, pH), density of individuals, flow conditions, sample type (e.g. soil, sediment), etc., that may facilitate or inhibit the amplification of whole mitogenomes.

Consideration of the species composition is also important. We observed that the de novo assembler was not able to assemble some scaffolds into entire mitogenomes (Figure 2). One possible explanation for this observed pattern is that conserved regions with high sequence similarity among species are difficult to accurately assemble from a complex mixture. Therefore, we expect the de novo assembly method, when not guided by a reference mapping approach, will be challenging in fish communities when species pairs are closely related and sequence similarity is high. Additionally, for de novo assemblies

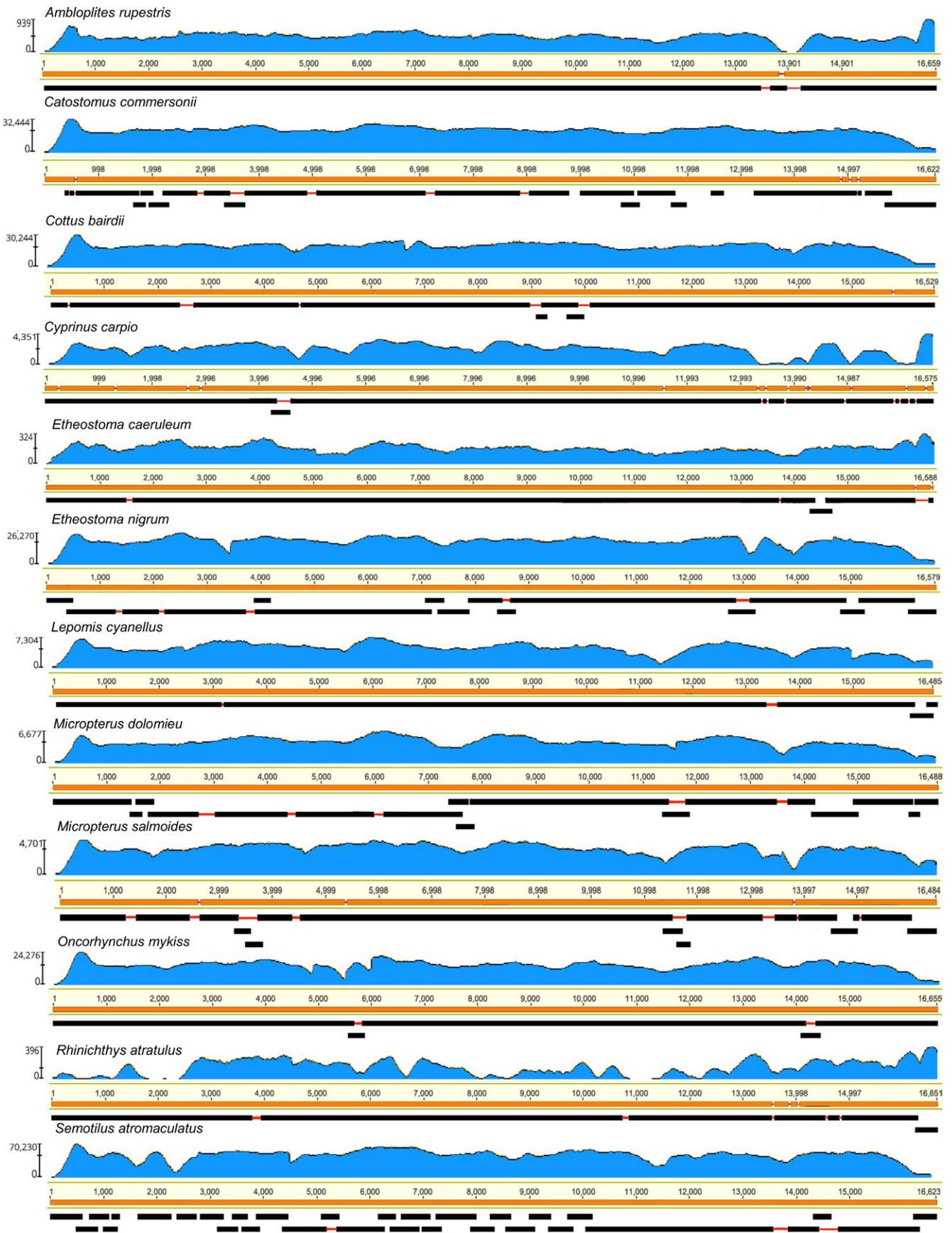


FIGURE 2 Linear visualization for long-range PCR-amplified fish mitogenomes sequenced from environmental DNA water sampled in Juday Creek, IN, USA. Species are depicted in alphabetical order from top to bottom excluding the two species (*Lepomis macrochirus* and *Salmo trutta*) that did not produce a full mitogenome (Table 1). The numbered orange bar with yellow background is the reference sequence. The blue graphic indicates the number of reads mapped at each site (i.e. sequencing depth in Table 1) and each of the black bars below the reference sequence are independent scaffolds that were de novo assembled and subsequently mapped to the reference. Scaffolds connected with a red line are a single scaffold and the line connecting them indicates a gap in the assembly. Gaps in the reference sequence (indicated by a red line) represent insertions compared to the de novo assembled scaffold

we observed that in some cases high sequencing depth inhibited the concatenation of multiple shorter scaffolds (Figure 2). While down sampling did join smaller scaffolds together more frequently, we still observed the pattern that species with the highest read depth tended to have the largest number of unassembled scaffolds (Table 1, Figure 2). Applying long-read sequencing technology such as PacBio SMRT technology to long-range amplified products could potentially solve this problem by avoiding short reads altogether (Schloss, Jenior, Koumpouras, Westcott, & Highlander, 2016).

From the laboratory perspective, there are many handling steps that could be optimized to improve detection of whole mitogenomes. For example, the filters were extracted with a Chloroform-isoamyl DNA extraction, but the use of buffered phenol (pH 8.0) in addition to chloroform-isoamyl is known to increase the quantity of high molecular weight DNA during DNA extractions (Blin & Stafford, 1976). Therefore, testing of laboratory methods from extraction to library preparation may increase the yield of eDNA suitable for long-range PCR.

We detected a small amount of contamination between samples run on the same MiSeq run (i.e. Juday Creek and the mock community). The low level of contamination between libraries did not result in high coverage of the mitogenomes nor allow de novo assembly of complete mitogenomes in these samples. Additionally, because our study design intentionally used tropical marine fish for the mock community, they could easily be excluded from occurring in Juday Creek, a temperate fresh water habitat. We cannot determine from our study's design at which point the contamination happened because the libraries showing the greatest amount of cross contamination were also processed in the laboratory at the same time (i.e. mock community and Juday Creek samples). The mock community and Juday Creek libraries were also prepared using the TruSeq Nano LT Sample Prep Kit using a single index step. Dual indexing helps to circumvent problems associated with "tag jumping," and we cannot rule out this phenomenon as a problem for our study (Schnell, Bohmann, & Gilbert, 2015). Future applications should take care to physically and temporally separate the processing of samples that may become contaminated and we recommend dual indexing samples to detect with greater accuracy false positive reads in libraries run on the same MiSeq flowcell.

The primers designed here were demonstrated *in silico* to potentially be useful for detecting a broad array of fish species of the class Actinopterygii. These primers could be made even more general by adding degenerate bases to sites showing variation in species with <95% match at the primer binding region (Appendix S2). Similarly, long-range PCR primers could be designed for other taxonomic groups, such as birds, mammals and amphibians. Our laboratory and bioinformatic approach should be generally applicable to animal mitochondrial genomes. However, given the current upper limits to the

length of long-range PCR amplifications, the method will make amplifying genomes from larger organelles, such as chloroplasts and plant mitochondria, problematic.

Extending the use of our method to research fields like population genetics will require additional studies. For example, while outside the scope of this study, there is the potential to phase haplotypes for genes or whole mitogenomes from aligned short fragments (O'Neil & Emrich, 2012). Phased haplotypes from eDNA shotgun-sequenced data would allow for population-level analyses from communities. For example, it may become possible to estimate population size from eDNA samples using a reverse inference method from population genetic equations that estimate mutation rate and haplotype diversity (Wares & Pappalardo, 2015). Using population genetic theory, at the least, the minimum number of individuals can be inferred from the haplotypes and used to estimate minimum population sizes that contributed to a sample of eDNA. However, to make use of this theory, there is a need for continued research focused on parsing out sequencing noise from real variation to determine intra-species haplotype diversity collected from environmental samples and high-throughout sequencing (Gómez-Rodríguez, Crampton-Platt, Timmermans, Baselga, & Vogler, 2015).

While it is promising that we could identify SNPs and estimate allele frequencies, there remain many questions about the validity of these estimates from eDNA. Specifically, we could not confirm our SNPs from tissues of the actual species and such tests are needed before adoption of this method is warranted. Until now, most studies generate operational taxonomic units (OTUs) and summarize the data for a species at 97%–99% similarity (Hänfling et al., 2016; Olds et al., 2016), and thus have not utilized the intraspecific level variation present in eDNA samples. A recent study by Sigsgaard et al. (2016) demonstrated that haplotype diversity is attainable from water samples at least at the single-species level using primers that targeted a small amplicon (<500 bp) in whale sharks. We expect that future studies applying our method of whole mitochondrial genome sequencing from water samples will yield similar results, but at the community scale for entire mitogenomes.

The continued advancement of single molecule and long-read technologies, such as the Oxford Nanopore MinION (Laszlo et al., 2014), will improve our approach. Here, we used Illumina sequencing which required sheering the mitogenomes, using sonication or a transposase-mediated process used in the Nextera library preparation kit, to fragment them before sequencing and subsequently remapping these reads to a reference sequence or conducting de novo assembly. Coupling long-range PCR amplifications and sequencing without fragmentation would avoid many of the problems associated with short-fragment based assembly or reference mapping. We expect that long-read sequencing technologies, once they are cost effective

and error rates are reduced, will become the method of choice for sequencing long-range PCR products and will allow population genetic analysis of eDNA samples.

ACKNOWLEDGEMENTS

We would like to thank the following individuals for tissue donations that facilitated this work: Cameron Turner, Mike Brueseke and Nathan Evans (University of Notre Dame); we additionally thank Dominic Chaloner for providing temperature data for Juday Creek. We would like to thank the Notre Dame Genomics & Bioinformatics Core Facility for their assistance with the Illumina MiSeq. We also thank Scott J. Emrich (University of Notre Dame) for his assistance in code development for down sampling our data for de novo assembly. This work was supported by the U.S. Department of Defense's Strategic Environmental Research and Development Program (RC-2240). The authors declare no competing financial interests. This is a publication of the Notre Dame Environmental Change Initiative.

AUTHORS' CONTRIBUTIONS

All authors contributed to study design; M.A.R. conducted the laboratory work; K.D., M.A.R., Y.L., B.P.O. and M.E.P. analysed the data; and K.D. and M.A.R. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

All raw data associated with this study have been deposited on the NCBI's Sequence Read Archive (SRA) (www.ncbi.nlm.nih.gov/sra) under the BioProject PRJNA317862: SRS2218772 (Mock Community), SRS2218773 (Mesocosm_HE3), SRS2218776 (Mesocosm_HS3) and SRS2218777 (Juday_Creek). All other intermediate processed data files and code used for analysis are available from Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.q5gg0> (Deiner et al., 2017).

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410.
- Blin, N., & Stafford, D. W. (1976). A general method for isolation of high molecular weight DNA from eukaryotes. *Nucleic Acids Research*, 3, 2303–2308.
- Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., ... De Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, 29, 358–367.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10, 421.
- Cameron, S. L. (2014). Insect mitochondrial genomics: Implications for evolution and phylogeny. *Annual Review of Entomology*, 59, 95–117.
- Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: Extending the concept of DNA barcoding. *Molecular Ecology*, 25, 1423–1428.
- Crampton-Platt, A., Douglas, W. Y., Zhou, X., & Vogler, A. P. (2016). Mitochondrial metagenomics: Letting the genes out of the bottle. *GigaScience*, 5, 1.
- Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., ... Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology & Evolution*, <https://doi.org/10.1111/2041-210X.12574>
- Crusoe, M. R., Alameidin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., ... Fenton, J. (2015). The khmer software package: Enabling efficient nucleotide sequence analysis. *F1000Research*, 4, 900.
- Deiner, K., Fronhofer, E. A., Mächler, E., Walsler, J. C., & Altermatt, F. (2016). Environmental DNA reveals that rivers are conveyor belts of biodiversity information. *Nature Communications*, 7, 12544.
- Deiner, K., Renshaw, M. A., Li, Y., Olds, B. P., Lodge, D. M., & Pfrender, M. E. (2017). Data from: Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Dryad Digital Repository*, <https://doi.org/10.5061/dryad.q5gg0>
- Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C., ... Nelson, N. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*, 4, 1.
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, 2460–2461.
- Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y., Jerde, C. L., ... Lodge, D. M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, 16, 29–41.
- Footo, A. D., Thomsen, P. F., Sveegaard, S., Wahlberg, M., Kielgast, J., Kynh, L. A., ... Gilbert, M. T. P. (2012). Investigating the potential use of environmental DNA (eDNA) for genetic monitoring of marine mammals. *PLoS ONE*, 7, e41781.
- Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M. J., Baselga, A., & Vogler, A. P. (2015). Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, 6, 883–894.
- Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, 95–98.
- Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., ... Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, 25, 3101–3119.
- Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, 270, S96–S99.
- Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., ... Nishida, M. (2013). MitoFish and MitoAnnotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology & Evolution*, 30, 2531–2540.
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology & Evolution*, 30, 772–780.
- Laszlo, A. H., Derrington, I. M., Ross, B. C., Brinkerhoff, H., Adey, A., Nova, I. C., ... Gundlach, J. H. (2014). Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnology*, 32, 829–833.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25, 1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
- Olds, B. P., Jerde, C. L., Renshaw, M. A., Li, Y., Evans, N. T., Turner, C. R., ... Lamberti, G. A. (2016). Estimating species richness using environmental DNA. *Ecology & Evolution*, 6, 4214–4226.
- O'Neil, S. T., & Emrich, S. J. (2012). Haplotype and minimum-chimerism consensus determination using short sequence data. *BMC Genomics*, 13, 1.

- Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*, 1420–1428.
- Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., ... Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, *25*, 527–541.
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842.
- Schloss, P. D., Jenior, M. L., Koumpouras, C. C., Westcott, S. L., & Highlander, S. K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ*, *4*, e1869.
- Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated—reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*, 1289–1303.
- Shaw, J. L., Clarke, L. J., Wedderburn, S. D., Barnes, T. C., Weyrich, L. S., & Cooper, A. (2016). Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biological Conservation*, *197*, 131–138.
- Shirey, P., Brueseke, M., Kenny, J., & Lamberti, G. (2016). Long-term fish community response to a reach-scale stream restoration. *Ecology and Society*, *21*, 11.
- Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., ... Møller, P. R. (2016). Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution*, *1*, 0004.
- Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4–18.
- Torti, A., Lever, M. A., & Jørgensen, B. B. (2015). Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Marine Genomics*, *24*, 185–196.
- Trebitz, A. S., Hoffman, J. C., Grant, G. W., Billehus, T. M., & Pilgrim, E. M. (2015). Potential for DNA-based identification of Great Lakes fauna: Match and mismatch between taxa inventories and DNA barcode libraries. *Scientific Reports*, *5*, 12162.
- Turner, C. R., Barnes, M. A., Xu, C. C., Jones, S. E., Jerde, C. L., & Lodge, D. M. (2014). Particle size distribution and optimal capture of aqueous microbial eDNA. *Methods in Ecology & Evolution*, *5*, 676–684.
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, *40*, e115.
- Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., ... Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*, 929–942.
- Wares, J. P., & Pappalardo, P. (2015). Can theory improve the scope of quantitative metazoan metabarcoding? *Diversity*, *8*, 1.
- Zhang, W., Cui, H., & Wong, L.-J. C. (2012). Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing. *Clinical Chemistry*, *58*, 1322–1331.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Deiner K, Renshaw MA, Li Y, Olds BP, Lodge DM, Pfrender ME. Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods Ecol Evol.* 2017;00:1–11. <https://doi.org/10.1111/2041-210X.12836>