Conservation Genetics Resources (2020) 12:433–446
https://doi.org/10.1007/s12686-019-01111-0

METHODS AND RESOURCES ARTICLE

Fish mitochondrial genome sequencing: expanding genetic resources to support species detection and biodiversity monitoring using environmental DNA

Julie C. Schroeter[1,2] · Aaron P. Maloy[1] · Christopher B. Rees[1] · Meredith L. Bartron[1]

Received: 10 May 2019 / Accepted: 29 August 2019 / Published online: 4 September 2019

Abstract

Conservation of aquatic resources is hampered by our limited knowledge of biological diversity and its distribution. Due to challenges with detection of rare or difficult to sample species, and the expansion of genetic technologies, fisheries professionals are supplementing traditional biodiversity field studies with emerging environmental DNA (eDNA) techniques. eDNA is generally evaluated with qPCR (primer- and probe-mediated single species detection), single-gene metabarcoding (PCR primer-mediated diversity analysis) or the emerging technique of multi-gene metagenomics (PCR independent diversity analysis). In each case, techniques are dependent on sequence databases for primer design and/or taxonomic assignment of recovered sequence data. Current reference databases contain limited mitochondrial genome information and are reliant on specific gene fragments, such as COI, which are not always suited for qPCR and metabarcoding marker design needs. To facilitate primer design and enhance taxonomic resolution for eDNA approaches, we describe a suite of order and/or family-specific long-range PCR primers sufficient for sequencing complete mitochondrial genomes. While the intent was to obtain mitochondrial genome data for freshwater fish, primers were designed on sequence alignments from all available species (including marine), and should be broadly applicable within their respective taxonomic group. We have sequenced 205 complete mitochondrial genomes representing 65 species/subspecies from 9 fish families, including novel genomes from 28 species not represented in GenBank at the time of submission. Continued expansion of species representation in mitochondrial genome databases will help move biodiversity assessment from single-gene metabarcoding approaches to multi-gene metagenomics and provide a valuable resource for eDNA applications, molecular ecology and phylogenetics.

Keywords Biodiversity assessment · Environmental DNA · Fish · Long-range PCR · Metabarcoding · Mitochondrial genome

Introduction

Environmental DNA (eDNA) techniques are quickly becoming routine in conservation research and are increasingly viewed by management professionals as a potentially cost-saving alternative to traditional field techniques (Goldberg et al. 2016). Further incorporation of eDNA techniques into management requires the continued development of genetic database resources necessary to support effective implementation. For freshwater fish species, a reasonably comprehensive reference database exists only for the barcoding region of the COI gene (Ward et al. 2009), which is publically available through the Barcode of Life Data System (BOLD; Ratnasingham and Hebert 2007). While this database remains an invaluable tool, it is limited to a small portion of the overall mitochondrial genome. Design of species-specific qPCR primers, universal primers for metabarcoding, and/or species level taxonomic resolution is not always possible using the COI gene. Other gene regions such as 12S ribosomal RNA (rRNA), 16S rRNA and CytB are also taxonomically discriminative (Miya et al. 2015; Olds et al. 2016; Evans et al. 2017) but lack comprehensive species representation

✉ Aaron P. Maloy
aaron_maloy@fws.gov

[1] Northeast Fishery Center, United States Fish and Wildlife Service, Lamar, PA 16848, USA

[2] Bozeman Fish Technology Center, United States Fish and Wildlife Service, Bozeman, MT 59715, USA

in public databases. Continued enhancement of public databases with complete mitochondrial genome sequences would provide data on all 13 protein coding genes, 2 rRNAs, 22 transfer RNAs (tRNAs) and the highly variable control region. Access to mitochondrial sequence data from additional species, and data that include increased geographic variation within a species, will provide greater flexibility in the design and application of eDNA techniques.

Currently, the design of species-specific qPCR detection assays (Farrington et al. 2015; Bronnenhuber and Wilson 2013) is constrained by the limited availability of sequence data outside of the COI barcoding region. Access to sequence data from across the mitochondrial genome provides a greater potential of locating ideal primer and probe annealing sites that convey a high level of species discrimination. Metabarcoding (Hänfling et al. 2016; Olds et al. 2016) is dependent on locating conserved priming sites that bracket a taxonomically informative area of variable sequence. Such sequence characteristics do not readily occur in protein coding genes, thus ‘universal’ primers often target the 12S or 16S rRNA region (Miya et al. 2015; Sarri et al. 2014). Additionally, metabarcoding relies on a database of high quality reference sequences to assign a taxonomic identification to the recovered sequences.

The online resource MitoFish (Iwasaki et al. 2013) provides access to a database of complete and partial fish mitochondrial genomes. Despite collating all publically available mitochondrial genomes of fish, comprehensive species representation within MitoFish is still lacking. Complete mitochondrial genome sequences are available for just 2744 (as of September 2019) of the 34,200 described fish species in FishBase (Froese and Pauly 2017), often with just a single genome representing each species. Ideally, databases should include representative genomes of all species and encompass multiple representatives of each species, including geographic variation to represent localized mutations in mitochondrial sequences that may occur across a species range. Development of such a resource is a large undertaking but is possible with next generation sequencing techniques. Genome skimming using shotgun data (Richter et al. 2015; Gan et al. 2014) and long-range PCR (Briscoe et al. 2013) are two approaches that provide a means to obtain mitochondrial genome data from a large number of individuals/species. Genome skimming is a PCR independent method in which sequenced libraries are composed of approximately 99% nontarget nuclear DNA, resulting in a process that is more expensive and requires a larger investment in computational resources to obtain a complete mitochondrial genome. Long-range PCR is used to enrich for mitochondrial DNA prior to sequencing. Highly enriched samples allow for greater levels of multiplexing resulting in lower sequencing cost and the need for fewer computation resources. However, an upfront investment is necessary to develop the long-range primer sets and obtain the sequencing data necessary for their design. Here, we describe order or family-specific primers for long-range PCR amplification for 6 orders and 9 families that we have utilized to sequence the mitochondrial genomes for 65 different species.

Methods

Tissue collection and DNA extraction

Tissues were collected for various sampling efforts over the past 2 years. When possible, whole fish voucher specimens were retained at the USFWS Northeast Fishery Center in Lamar, Pennsylvania. Fin clips were preserved in 100% ethanol and placed at − 80 °C for long-term storage. Genomic DNA was extracted from fin clips or muscle tissue using the DNeasy® Blood and Tissue Kit (Qiagen, Inc., Germantown, MD, USA). For most species, tissue from multiple individuals was obtained. With the exception of *Scaphirhynchus* species, all species were field-identified by trained fisheries biologists. *Scaphirhynchus* species were identified using a suite of microsatellite loci (McQuown et al. 2000; Schrey et al. 2007; Tranah et al. 2004). Two heuristic steps were taken to ensure quality control of sequences submitted to GenBank. The COI barcoding region was used to confirm the field identification of each specimen using BOLD (Ratnasingham and Hebert 2007) and verify sample integrity during processing. In addition, a cluster analysis of all newly sequenced full-length mitochondrial genomes and reference genomes obtained from GenBank was used to screen for potentially chimeric genomes prior to submission.

Primer design and optimization

Complete mitochondrial genome sequences for each order or family group were downloaded from GenBank (Benson et al. 2013) and aligned using the MAFFT algorithm (Katoh and Standley 2013) in Geneious R10 (Kearse et al. 2012). In most cases, this included a non-redundant list of every species with an available NCBI RefSeq (O’Leary et al. 2016) sequence. The total number of available sequences used in long-range primer design alignments varied between taxonomic groups as follows: Acipenseridae/Polyodontidae (15), Clupeidae (13), Catostomidae (13), Cyprinidae (538), Centrarchidae/Percidae (21), Salmonidae (50), Ictaluridae (7). Highly conserved regions were identified visually and Primer3 (Untergasser et al. 2012) was run within Geneious R10 (Biomatters Ltd., Newark, NJ, USA) to locate suitable primer annealing sites. Primer annealing sites were generally located in the rRNA genes and tRNAs due to their higher levels of conservation within order or family groups. Primer sets were chosen to amplify the entire mitochondrial genome in four overlapping sections each with a length of 3000 to 7000 base pairs (Table 1). All primer pairs, except for those designed for Cyprinidae, were designed based on a 100% consensus of the aligned sequences to ensure primer specificity across all species within the targeted taxonomic group. Due to the sequence diversity within Cyprinidae, a 90% consensus threshold was used to identify primer locations that minimized the need for redundant bases. Redundant bases were used sparingly and avoided in the 3′ end of any primer. Primer design for Clupeidae was restricted to genera with representation in freshwater habitats due to primer design difficulty with broader family representation. Primer sets for each mitochondrial genome region were optimized by running temperature gradients and template concentration dilutions to identify optimal amplification conditions, assessed by product evaluation on 1.5% agarose gels (Table 1).

PCR and gel electrophoresis

Long-range PCR was used to amplify the complete mitochondrial genome in four overlapping sections (Fig. 1). Each 25 µl PCR was amplified with either Q5 Hot Start High-Fidelity Master Mix (New England Biolabs, Ipswich, MA, USA) or Kapa HiFi HotStart ReadyMix (Kapa Biosystems, Wilmington, MA, USA) following the manufacturer's recommended concentrations. Reactions were run under the following conditions: enzyme activation for 2 min at 98 °C, followed by 35 cycles of 20 s denaturing at 98 °C, 20 s at primer annealing temperature (Table 1) and 3 min at 72 °C, followed by a final 7 min elongation at 72 °C. All PCR products were visualized on a 1.5% agarose gel to verify amplification success.

Illumina sequencing

Successful amplification products were quantified using a Qubit™2.0 (Life Technologies, Carlsbad, CA, USA) and corresponding fragments from the same specimen were pooled in equimolar ratios. Pooled PCR amplicons were bead purified, fluorometrically quantified and diluted to a standard concentration of 0.2 ng/µl. DNA libraries were created using the Nextera XT Library Prep Kit following the manufacturer's instructions (Illumina, Inc., San Diego, CA, USA). Bead normalized libraries were pooled and sequenced using the MiSeq Reagent V2 Kit with 2 × 250 paired end reads. Sequences were sorted into FASTQ files and trimmed to remove remaining adaptor and index sequences using the onboard Illumina FASTQ workflow.

Mitochondrial genome assembly

After trimming Illumina adaptor and index sequences, FASTQ files were uploaded into Geneious R10 (Kearse et al. 2012) for quality control and assembly. Low quality ($Q < 20$) bases were trimmed from each end and short reads (< 25 bp) and reads with an average read quality of Q 20 or less were discarded. Reads were merged (merge rate = normal) before error correction and normalization (BBNorm version 37.25; error correction, default settings; normalization, target coverage 60 and minimum depth = 6). The normalized merged reads were de novo assembled using the Geneious assembler under medium sensitivity with the circularize contigs function turned on. A maximum mismatch of 5% was allowed. Occasionally the full mitochondrial genome was not obtained as a single complete circular contig using normalized data. In these cases, a de novo assembly was done using all available merged reads. Consensus sequences were based on the majority base call for each nucleotide position. Gene annotations were mapped to new genomes in Geneious R10 (Kearse et al. 2012) using existing NCBI reference sequences of the same or closely related species as the source genome. All complete genomes were submitted to GenBank (Table 2). Mitochondrial genomes were aligned using MAFFT (Katoh and Standley 2013) and the maximum percent of base pair differences were calculated within each species.

Results

The novel order and family-level primer sets presented here allowed for the successful amplification and sequencing of 205 mitochondrial genomes from 9 families of fish representing 65 species/subspecies, 28 of which were not available in GenBank at the time of submission. It was not uncommon to observe a failed PCR reaction in one of the four regions being amplified under the initial PCR conditions. However, the majority of these instances could be corrected by adjusting the annealing temperature, template concentration or a change in *Taq* polymerase. Overall, amplification success across all primers sets was greater than 90%, with only a few species failing to amplify one or more of the four regions. All mitochondrial genomes assembled had a length (16,486–16,832 bp) and gene composition typical of most fish species (Satoh et al. 2016) including: two rRNA genes, 13 protein coding genes, 22 tRNAs and the highly variable displacement loop (control region). With the exception of ND6, all protein coding and rRNA genes were coded on the heavy strand. Eight tRNA genes (tRNA-Ala, tRNA-Asn, tRNA-Cys, tRNA-Gln, tRNA-Glu, tRNA-Pro, tRNA-Ser, tRNA-Tyr) were coded on the light strand with the remaining 14 coded on the heavy strand.

Sequencing multiple mitochondrial genomes from the same species revealed varying levels of intraspecies genetic variation. The total intraspecies base pair composition differed by a maximum of 2.65% (black crappie) with an

Table 1 Long-range primers used to amplify complete mitochondrial genomes in four overlapping regions

Primer	Sequence (5′–3′)	Length (bp)	Annealing temperature (°C)	Annealing location	Approximate size (bp)
Order Acipenseriformes					
ACI_R1-F	GTTGTTAATTCAACTATAAAAACC	24	57	tRNA-Glu	4285
ACI_R1-R	TTCATTTAAAAGACAAGTGATTAC	24		16S rRNA	
ACI_R2-F	AACCTAACGAGCCTAGTAATAG	22	58	16S rRNA	5804
ACI_R2-R	GTCTTGGAATCCTAATTGTG	20		COII	
ACI_R3-F	ATCCTACAAAATCTTAGTTAAC	22	56	tRNA-Asn	6614
ACI_R3-R	AGAATTAGCAGTTCTTAGTG	20		tRNA-Ser	
ACI_R4-F	ATTTCGGCTCAACTAATTAT	20	57	tRNA-Arg	5633
ACI_R4-R	AGTTTAATGTAGAATCTTAGCTTT	24		tRNA-Pro	
Family Catostomidae					
CAT_R1F	CCCGTCACTCTCCCCTGTTA	20	65	12S rRNA	6196
CAT_R1R	AAGGAAGTGGCAGAGTGGTT	20		tRNA-Ser	
CAT_R2F	CTCTGTCTTCGGGGCTACAA	20	65	tRNA-Tyr	6517
CAT_R2R	TTGCACCAAGAGTTTYTGGTTCC	23		tRNA-Leu	
CAT_R3F	AAGACCTCTGATTTCGRCTCAGA	23	65	tRNA-Arg	5611
CAT_R3R	CAGGGGTGGGAGTTAAAATCT	21		tRNA-Pro	
CAT_R4F	TGAAGAACCACCGTTGTYATTCA	23	65	tRNA-Glu	3792
CAT_R4R	TAGGCAACCAGCTATCACCA	20		16S rRNA	
Family Centrarchidae and Percidae					
CP_R1-F	GCAATCACTTGTCTTTTAA	19	56	16S rRNA	5794
CP_R1-R	CTTAAAAGGCTAACGCTA	18		tRNA-Lys	
CP_R2-F	TTACCGCTCTGTCACT	16	56	tRNA-Ser	4822
CP_R2-R	AGTTTTTGGTTCCTAAGAC	19		tRNA-Leu	

Table 1 (continued)

Primer	Sequence (5′–3′)	Length (bp)	Annealing temperature (°C)	Annealing location	Approximate size (bp)
CP_R3-F	CCAAGGAAAGATAATG	16	53	tRNA-Gly/ND3	6354
CP_R3-R	AATAGTTGTCCCTCAC	16		D-Loop	
CP_R4-F	CCCAAAGCTAGGATTCTA	18	55	tRNA-Pro	3528
CP_R4-R	TAGATAGAAACTGACCTGGA	20		16S rRNA	
Family Clupeidae					
CLU_R1-F	ACCAAAAGTTTAACGGCCGC	20	65	16S rRNA	4980
CLU_R1-R	GGGGTTCGATTCCTCCCTTT	20		tRNA-Ser	
CLU_R2-F	CCTTCAAAGCTCCAAGCAGG	20	65	tRNA-Trp	4917
CLU_R2-R	CTGAGCCGAAATCAGAAGTCT	21		tRNA-Arg	
CLU_R3-F	TACGTCTCYATCTACTGATGAGGATC	26	65	tRNA-Gly	6080
CLU_R3-R	GCTTTGGGAGTTAGAGGTGGA	22		tRNA-Pro	
CLU_R4-F	YGAAAAACCACCGTTGTTATTCAA	24	65	tRNA-Glu	4888
CLU_R4-R	GAACCCTTAATAGCGGCTGC	20		16S rRNA	
Family Cyprinidae					
CYP_R1-F	TAAAACTCGTGCCAGCCACC	20	61	12S rRNA	4959
CYP_R1-R	TTGTAGGATCGAGGCCTTCC	20		tRNA-Asn	
CYP_R2-F	AAGCTTTCGGGCCCATACC	19	65	tRNA-Met	6048
CYP_R2-R	TCTGAGCCGAAATCAGAGGTC	21		tRNA-Arg	
CYP_R3-F	AGCCCATGACCMCTAACCGGA	21	65	tRNA-Gly/ND3	4752
CYP_R3-R	DGTTTTTCGTAGGCTTGCCAT	21		CytB	
CYP_R4-F	TTGGTCTTAGGAACCAAAAACTCT	24	65	tRNA-Leu	5272
CYP_R4-R	CCGTCAGGTCCTTTGGGTTT	20		12S rRNA	

Table 1 (continued)

Primer	Sequence (5′–3′)	Length (bp)	Annealing temperature (°C)	Annealing location	Approximate size (bp)
Family Ictaluridae					
ICT_R1-F	TCAGACCCACCTAGAGGAGC	20	65	12S rRNA	6595
ICT_R1-R	GCCGCGTCTTGGAATCCTAG	20		COII	
ICT_R2-F	AGATGAGAAGGCCTCGATCCT	21	65	tRNA-Asn	6690
ICT_R2-R	TTGGTTCCTAAGACCAAYGGATGA	24		tRNA-Leu	
ICT_R3-F	AAGACCTCTGATTTCGRCTCAGA	23	67	tRNA-Arg	5524
ICT_R3-R	TCTCCGGATTACAAGACCGG	20		tRNA-Thr	
ICT_R4-F	TAACCAGGACYAATGACT	18	65	tRNA-Glu	3213
ICT_R4-R	CTTACCATGTTACGACTTG	19		12S rRNA	
Family Salmonidae					
SAL_R1-F	CTATATACCACCGTCGTC	18	55	12S rRNA	4845
SAL_R1-R	AATGTCTTTGTGGTTGG	17		COI	
SAL_R2-F	AAGTCCCCTCAATTCTAG	18	55	tRNA-lle/tRNA-Gln	4244
SAL_R2-R	GCTTARTGTCATGGTCAG	18		ATP6/ATP8	
SAL_R3-F	GTTAGCCTTTTAAGCTAAAG	20	55	tRNA-Lys	6451
SAL_R3-R	GTGGTTTTTCAAGTCATTA	19		tRNA-Glu	
SAL_R4-F	GCRCAATTTGGACTTC	16	55	ND5	4829
SAL_R4-R	TAGAGAATGTAGCCCATT	18		12S rRNA	

Primers were designed to amplify species within an order or family group

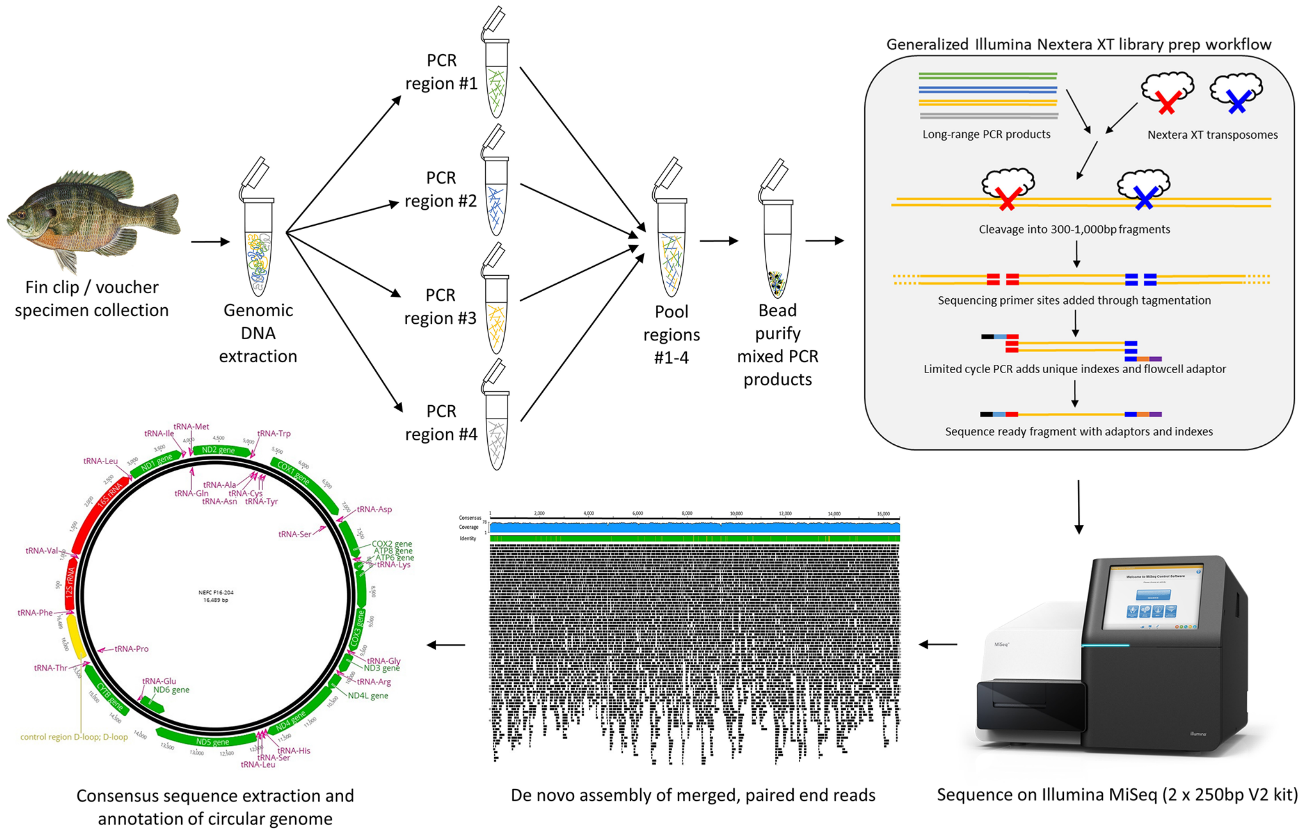

Fig. 1 Mitochondrial genome sequencing workflow. Voucher specimens and fin clips samples are used to obtain genomic DNA from fish. Independent long-range PCR reactions amplify the mitochondrial genome in four overlapping regions. Regions are pooled, purified and prepared for sequencing using the Illumina Nextera XT Prep Kit workflow. After Illumina sequencing, reads are demultiplexed and de novo assembled into a circular contig. The consensus sequence is extracted and annotated prior to final quality assurance and submission to GenBank

average of 0.38%. Seven species showed levels of base pair variation over 1.0%. Four of these (channel catfish, redbreast sunfish, golden shiner, black crappie) come from a geographically disperse area ranging from South Carolina to New York to Michigan, while emerald shiner all originated from New York water. The remaining two species (round whitefish, Alaska; lake sturgeon and emerald shiner, New York) originated from a geographically similar area and the elevated level of variation was due to a variable number of tandem repeats in the displacement loop (Table 2).

Discussion

The primer sets presented here offer a way to obtain mitochondrial genomes from 9 families of freshwater and marine fish species/subspecies and cover approximately 60% of the estimated 1050 species native to North America (Lundberg et al. 2000). While truly universal primers may not be possible, amplification of novel genomes from species not available during initial primer design suggests a broad application of the primers within their respective taxonomic target group. Overall, primers performed as expected and amplified a range of families/genera within their targeted taxonomic group. Primers designed at the order level were successful when the order contained a limited number of families. In the case of Acipenseriformes, there are only two extant family groups with relatively limited species diversity. Primer sets occasionally failed to amplify one of the target regions. However, adjustments in annealing temperature, template concentration or a change in the type of *Taq* polymerase used generally resulted in successful amplification. The families of Centrarchidae and Percidae were consolidated and a suite of primers were developed to target both families simultaneously. Cyprinidae is a very large family and mitochondrial genomes were available from 538 species for sequence alignment and primer design. In this instance, a 90% consensus of the sequence alignment was used to design family-specific primers. Under the 90% criteria, primer mismatching is possible and primers may show reduced performance with certain species. Design precautions ensured that potential mismatches were reduced in the 3′ end of the

Table 2 Sequenced mitochondrial genomes

Order	Family	Scientific name	Common name	Number of individuals sequenced	Accession numbers	Origin	Percent sequence divergence
Acipenseriformes	Acipenseridae	*Acipenser brevirostrum**	Shortnose sturgeon	10	KX817302–KX817311	GA (h)	0.04
		*Acipenser fulvescens**	Lake sturgeon	5	KU985070, KU985081–KU985084	NY	1.13
		Acipenser oxyrinchus oxyrinchus	Atlantic sturgeon	2	KU985073, KU985074	PA (h)	0.00
		*Scaphirhynchus albus**	Pallid sturgeon	10	KU985069, KU985079, KU985080, MF101777, MF101778, MF101780, MF101781, MF101785, MF101786, MF101790	MO, NE, ND	0.66
		Scaphirhynchus albus × *S. platorynchus**	Hybrid pallid × shovelnose sturgeon	3	KU985072, KU985077, KU985078	MO, ND	0.63
		*Scaphirhynchus platorynchus**	Shovelnose sturgeon	10	KU985071, KU985075, KU985076, MF101779, MF101782–MF101784, MF101787–MF101789	MO, NE, ND	0.68
		*Scaphirhynchus suttkusi**	Alabama sturgeon	2	MF101791, MF101792	AL	0.00
	Polyodontidae	*Polyodon spathula*	American paddlefish	2	KU985085, KU985086	PA (h)	0.00
Clupeiformes	Clupeidae	*Alosa aestivalis**	Blueback herring	1	MG570463	SC	–
		*Alosa mediocris**	Hickory shad	1	MG570462	SC	–
		Alosa pseudoharengus	alewife	3	MG570421, MG570439, MG570440	Lake Ontario, Lake Michigan	0.04
		Alosa sapidissima	American shad	2	MG570422, MG570434	MD, SC	0.44
		Dorosoma cepedianum	American gizzard shad	3	MG570459, MG570415, MG570429	NY, SC	0.60

Table 2 (continued)

Order	Family	Scientific name	Common name	Number of individuals sequenced	Accession numbers	Origin	Percent sequence divergence
Cypriniformes	Catostomidae	*Catostomus catostomus**	Longnose sucker	2	MG570441, MG570442	Lake Michigan	0.01
		Catostomus commersonii	White sucker	4	MG570423, MG570424, MH324426, MH324427	PA, Lake Erie	0.54
		Erimyzon oblongus	Creek chubsucker	2	MG570410, MG570411	PA	0.08
		Minytrema melanops	Spotted sucker	2	MG570436, MG570430	SC	0.04
	Cyprinidae	*Ctenopharyngodon idella*	Grass carp	1	MG570437	SC	–
		*Cyprinella leedsi**	Bannerfisn shiner	4	MG570431, MG570432, MH324417, MH324418	SC	0.12
		Cyprinus carpio	Common carp	3	MG570426, MG570435, MG570427	Lake Erie, PA, SC	0.03
		*Exoglossum maxillingua**	Cutlips minnow	1	MG570453	NY	–
		Luxilus cornutus	Common shiner	2	MG570449, MG570447	NY	0.04
		*Nocomis micropogon**	River chub	1	MH324421	NY	–
		Notemigonus crysoleucas	Golden shiner	4	MG570428, MG570438, MG570425, MG570412	Lake Erie, NY, PA	1.42
		Notropis atherinoides	Emerald shiner	2	MG570455, MG570456	NY	1.14
		Notropis bifrenatus	Bridle shiner	3	MG570408, MG507409, MG570451	NY, PA	0.25
		*Notropis chalybaeus**	Ironcolor shiner	1	MG570407	PA	–
		*Notropis heterolepis**	Blacknose shiner	2	MG570413, MG570414	PA	0.00
		*Notropis hudsonius**	Spottail shiner	1	MG570443	Lake Michigan	–
		Pimephales notatus	Bluntnose minnow	4	MG570450, MG570420, MG570457, MG570458	NY, PA	0.17
		Pimephales promelas	Fathead minnow	2	MG570452, MG570454	NY	0.20
		Rhinichthys atratulus	Eastern blacknose dace	2	MG570444, MG570445	NY	0.32
		Rhinichthys cataractae	Longnose dace	2	MG570446, MG570448	NY	0.07
		*Rhinichthys obtusus**	Western blacknose dace	2	MG570416, MG570417	WI	0.02
		Semotilus atromaculatus	Creek chub	2	MG570418, MG570419	WI	0.08

Table 2 (continued)

Order	Family	Scientific name	Common name	Number of individuals sequenced	Accession numbers	Origin	Percent sequence divergence
Perciformes	Centrarchidae	*Lepomis auritus**	Redbreast sunfish	3	MF621723, MH301065, MH301066	PA, SC	1.17
		Lepomis gibbosus	Pumpkinseed	3	MF621724–MF621726	Lake Erie, NY, WI	0.09
		*Lepomis gulosus**	Warmouth	2	MH301067, MH301068	SC	0.01
		Lepomis macrochirus	Bluegill	3	MF621712–MF621714	Lake Erie, PA, SC	0.50
		*Lepomis punctatus**	Spotted sunfish	2	MF621732, MH301069	SC	0.05
		Micropterus dolomieu	Smallmouth bass	2	MF621710, MF621711	Lake Erie, NY	0.07
		Micropterus salmoides floridanus	Largemouth bass: northern strain	3	MH301070–MH301072	Lake Erie, NY, PA	0.10
		Micropterus salmoides salmoides	Largemouth bass: southern strain	4	MH301073–MH301076	AL, TN, TX	0.60
		Pomoxis nigromaculatus	Black crappie	5	MF621715, MF621719, MH301081, MH324419, MH324420	Lake Erie, NY, PA, SC	2.65
	Percidae	*Etheostoma flabellare**	Fantail darter	2	MH301059, MH301060	NY	0.24
		*Etheostoma olmstedi**	Tessellated darter	4	MH301061–MH301064	NY, PA	0.51
		Perca flavescens	Yellow perch	5	MF621736, MH301077–MH301080	Lake Erie, PA, IL	0.77
		Sander canadensis	Sauger	1	MH301082	QC	–
		Sander vitreus	Walleye	5	MH301083–MH301086, MH324422	Lake Erie, Lake Ontario, NJ, NY	0.28

Table 2 (continued)

Order	Family	Scientific name	Common name	Number of individuals sequenced	Accession numbers	Origin	Percent sequence divergence
Salmoniformes	Salmonidae	_Coregonus artedi_*	Cisco	4	MF621765, MF621766, MH301055, MH301056	Lake Ontario, Lake Huron (h), PA (h)	0.81
		_Coregonus clupeaformis_*	Lake whitefish	2	MH301057, MH301058	Lake Ontario	0.04
		Prosopium cylindraceum	Round whitefish	4	MF621759, MF621764, MF621767, MF621768	AK	1.12
		Oncorhynchus kisutch	Coho salmon	2	MF621749	WI	0.15
		Oncorhynchus mykiss	Rainbow trout	1	MF621750	WI	–
		Salvelinus alpinus	Arctic char	3	MF621740, MF621741, MF621743	AK	0.03
		Salvelinus fontinalis	Brook trout	3	MF621737–MF621739	NY, PA, WI	0.44
		_Salvelinus namaycush_*	Lake trout	6	MF621742–MF621748	AK, Lake Erie, Lake Ontario, PA	0.33
		Salmo trutta	Brown trout	4	MF621760–MF621763	Lake Erie, NY, WI	0.17
		Thymallus arcticus	Arctic grayling	7	MF621752–MF621758	AK	0.32
Siluriformes	Ictaluridae	_Ameiurus catus_*	White catfish	6	MG570433, MG570464, MG570465, MH324423–MH324425	SC	0.31
		_Ameiurus natalis_*	Yellow bullhead	2	MF621735, MG570406	Lake Erie, PA	0.07
		_Ameiurus nebulosus_*	Brown bullhead	3	MF621733, MF621734	Lake Erie	0.00
		Ictalurus furcatus	Blue catfish	2	MG570460, MG570461	SC	0.33
		Ictalurus punctatus	Channel catfish	6	MF621716–MF621722	Lake Michigan, PA, SC	1.10
		_Pylodictis olivaris_*	Flathead catfish	4	MF621727–MF621730	SC	0.23
Six orders	Nine families	34 Genera	65 Species/subspecies (28 novel)	205 Mitochondrial genomes			

Species marked with * represent novel mitochondrial genomes without species representation in GenBank at the time of submission. Percent sequence divergence represents the maximum nucleotide variation observed within each species across the mitochondrial genome. Origin locations denoted with (h) indicate hatchery fish primer. All cyprinid species evaluated to date have successfully produced PCR amplicons for all four mitochondrial genome fragments. Sequence alignment of available Clupeidae mitochondrial genomes lacked sufficient conservation to design robust primers at the family level. Primer design was thus restricted to those species found in North American freshwater habitats. The broader family-level applicability of the Clupeidae primer sets remains uncertain.

Amplification of complete mitochondrial genomes in two overlapping regions is possible (Zhu et al. 2013). However, amplification of long PCR templates is sensitive to DNA quality (Deagle et al. 2006). In our laboratory, recently-obtained fin clips stored in 95% ethanol at room temperature showed poor amplification success after more than 3 months of storage. In contrast, fin clips stored in 95% ethanol at − 80 °C provided consistent amplification after storage in excess of 2 years. Overall, we experienced better consistency in amplification success when targeting shorter fragments to obtain the complete mitochondrial genome in four overlapping regions, though amplification in two fragments is possible. Other tissue preservation methods may offer advantages over ethanol (Kilpatrick 2002) and allow better preservation of high molecular weight DNA. The need for large intact fragments of mitochondrial DNA is a methodological weakness of long-range PCR and generally precludes the analysis of archived museum specimens. In this instance, the PCR independent approach of genome skimmer offers a viable strategy to mine the wealth of voucher specimens available through museums.

Emerging techniques in the field of molecular ecology and eDNA are dependent on the continued development of representative reference databases. In particular, multi-gene metagenomics avoids potential PCR bias associated with metabarcoding biodiversity studies by directly sequencing eDNA without a proceeding PCR amplification step (Bista et al. 2018; Tang et al. 2014). Environmental metagenomic strategies have proven effective for detecting insect species from eDNA water samples (Crampton-Platt et al. 2016), but are hindered by low level recovery of mitochondrial DNA relative to non-target genomic material. Both non PCR-mediated (Liu et al. 2016) and PCR-based methods (Deiner et al. 2017) are being used to enrich the mitochondrial DNA fraction prior to sequencing. In each instance, reference genomes are used to enhance the recovery and identification of sequencing reads obtained from mixed species assemblages and are central to the multi-gene metagenomics approach.

Continued expansion of mitochondrial genome databases to include both a greater number of species and increased representation of species from throughout their range will provide an improved basis for analysis. For example, we observed intraspecies variation across the black crappie genome of 2.65%, a comparatively high value relative to the average of 0.38% (Table 2). Further examination clearly shows a geographic component: within species variation from northern locations (New York, Pennsylvania, Lake Erie) was only 0.06% while those originating from a southern location (South Carolina) variated by 0.07%. Two additional species, round whitefish and lake sturgeon, were obtained from geographically similar areas, Alaska and New York respectively, but still had variation in excess of 1%. This variation was attributed to a variable number of tandem repeats found in the displacement loop region. Excluding the displacement loop, both species (round whitefish, 0.17%; lake sturgeon, 0.33%) had a level of variation less than the average across all species examined. Tandem repeats within the displacement loop have been previously described and attributed to adaptation to harsh environments (Hirayama et al. 2010).

Based on the limited data presented here it is not possible to discern the full extent of intraspecies variation, but it does suggest a comprehensive evaluation of the issue is warranted. Intraspecies variation is of concern when designing species-specific eDNA markers and assigning taxonomic designations to sequencing reads in metabarcoding applications. Reference datasets that lack sufficient sequence diversity can result in qPCR markers that perform poorly across a species geographic range. The lack of sufficient sequence diversity will also negatively impact metabarcoding read classification with sequence variants remaining unclassified due to the lack of matching sequences in the reference dataset. Continued expansion of reference data sets to include additional species and sequence diversity is an essential foundational aspect of current and future eDNA applications. Additional sequencing with greater geographic representation will also allow future studies to explore intraspecific variation in a broader context and identify mitochondrial regions most suitable for marker development.

Use of order or family specific primers to easily obtain mitochondrial genome data from a large number of fish species is a valuable asset for applications such as eDNA, molecular ecology, conservation genetics, and phylogenetics. Improved species representation and geographic diversity will increase the efficiency of species-specific primer design for qPCR assays, provide more robust reference sequences for species identification in metabarcoding applications, and provide a basis for increased use of multigene metagenomics applications. Utilization of large mitochondrial genome databases will allow the most taxonomically discriminative marker or marker combinations to be identified, which may require targeting different regions within the mitochondrial genome. It is anticipated that continued expansion and public availability of mitochondrial genome data for fish (and all species in general) will greatly expand future applications of genomic research.

Acknowledgements The authors would like to thank our collection partners at Mansfield University, Samford University, Auburn University, Montana State University, University of California, Davis, Missouri Department of Conservation, Alabama Department of Conservation and Natural Resources, Nebraska Game and Parks Commission, Montana Fish, Wildlife and Parks, New York State Department of Environmental Conservation, Pennsylvania Fish and Boat Commission, United States Fish and Wildlife Service, United States Geological Survey, University of Connecticut. The authors would also like to thank J. Kalie for extensive laboratory assistance. Funding for this work was provided by the U.S. Fish and Wildlife Service and the Great Lakes Restoration Initiative. The findings and conclusions in this article are those of the author(s) and do not necessarily represent the views of the U.S. Fish and Wildlife Service.

Author contributions A.M., J.S., and C.R. designed the research. A.M. and J.S. performed the research and analyzed the data. A.M., J.S., C.R., and M.B. contributed to writing the manuscript.

Data availability All whole mitochondrial genome sequences are available on GenBank with the following accession numbers: KU985069–KU985086, KX817302–KX817311, MF101777–MF101792, MF621710–MF621768, MG570406–MG570465, MH301055–MH301086, and MH324417–MH324427.

References

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Res 41:D36–D42. https://doi.org/10.1093/nar/gks1195

Bista I, Carvalho GR, Tang M, Walsh K, Zhou X, Hajibabaei M, Shokralla S, Seymour M, Bradley D, Liu S, Christmas M, Creer S (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Mol Ecol Resour. https://doi.org/10.1111/1755-0998.12888

Briscoe AG, Goodacre S, Masta SE, Taylor MI, Arnedo MA, Penney D, Kenny J, Creer S (2013) Can long-range PCR be used to amplify genetically divergent mitochondrial genomes for comparative phylogenetics? A case study within spiders (Arthropoda: Araneae). PLoS ONE 8:e62404. https://doi.org/10.1371/journal.pone.0062404

Bronnenhuber JE, Wilson CC (2013) Combining species-specific COI primers with environmental DNA analysis for targeted detection of rare freshwater species. Conserv Genet Resour 5:971–975. https://doi.org/10.1007/s12686-013-9946-0

Crampton-Platt A, Yu DW, Zhou X, Vogler AP (2016) Mitochondrial metagenomics: letting the genes out of the bottle. GigaScience 5:15. https://doi.org/10.1186/s13742-016-0120-y

Deagle BE, Eveson JP, Jarman SN (2006) Quantification of damage in DNA recovered from highly degraded samples—a case study on DNA in faeces. Front Zool 3:11. https://doi.org/10.1186/1742-9994-3-11

Deiner K, Renshaw MA, Li Y, Olds BP, Lodge DM, Pfrender ME (2017) Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. Methods Ecol Evol 8:1888–1898. https://doi.org/10.1111/2041-210x.12836

Evans NT, Li Y, Renshaw MA, Olds BP, Deiner K, Turner CR, Jerde CL, Lodge DM, Lamberti GA, Pfrender ME (2017) Fish community assessment with eDNA metabarcoding: effects of sampling design and bioinformatic filtering. Can J Fish Aquat Sci 74:1362–1374. https://doi.org/10.1139/cjfas-2016-0306

Farrington HL, Edwards CE, Guan X, Carr MR, Baerwaldt K, Lance RF (2015) Mitochondrial genome sequencing and development of genetic markers for the detection of DNA of invasive bighead and silver carp (*Hypophthalmichthys nobilis* and *H. molitrix*) in environmental water samples from the United States. PLoS ONE 10:17. https://doi.org/10.1371/journal.pone.0117803

Froese R, Pauly D (2019) FishBase. World Wide Web electronic publication. https://www.fishbase.org. Accessed 04 2019

Gan HM, Schultz MB, Austin CM (2014) Integrated shotgun sequencing and bioinformatics pipeline allows ultra-fast mitogenome recovery and confirms substantial gene rearrangements in Australian freshwater crayfishes. BMC Evol Biol 14:19. https://doi.org/10.1186/1471-2148-14-19

Goldberg CS, Turner CR, Deiner K, Klymus KE, Thomsen PF, Murphy MA, Spear SF, Mckee A, Oyler-Mccance SJ, Cornman RS, Laramie MB, Mahon AR, Lance RF, Pilliod DS, Strickler KM, Waits LP, Fremier AK, Takahara T, Herder JE, Taberlet P (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods Ecol Evol 7:1299–1307. https://doi.org/10.1111/2041-210x.12595

Hänfling B, Lawson Handley L, Read DS, Hahn C, Li J, Nichols P, Blackman RC, Oliver A, Winfield IJ (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. Mol Ecol 25:3101–3119. https://doi.org/10.1111/mec.13660

Hirayama M, Mukai T, Miya M, Murata Y, Sekiya Y, Yamashita T, Nishida M, Watabe S, Oda S, Mitani H (2010) Intraspecific variation in the mitochondrial genome among local populations of Medaka *Oryzias latipes*. Gene 457:13–24. https://doi.org/10.1016/j.gene.2010.02.012

Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K, Takeshima H, Miya M, Nishida M (2013) MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol Biol Evol 30:2531–2540. https://doi.org/10.1093/molbev/mst141

Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kilpatrick CW (2002) Noncryogenic preservation of mammalian tissues for DNA extraction: an assessment of storage methods. Biochem Genet 40:53–62

Liu S, Wang X, Xie L, Tan M, Li Z, Su X, Zhang H, Misof B, Kjer KM, Tang M, Niehuis O, Jiang H, Zhou X (2016) Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. Mol Ecol Resour 16:470–479. https://doi.org/10.1111/1755-0998.12472

Lundberg JG, Kottelat M, Smith GR, Melanie LJS, Gill AC (2000) So many fishes, so little time: an overview of recent ichthyological discovery in continental waters. Ann Mo Bot Gard 87:26–62. https://doi.org/10.2307/2666207

Mcquown EC, Sloss BL, Sheehan RJ, Rodzen J, Tranah GJ, May B (2000) Microsatellite analysis of genetic variation in sturgeon: new primer sequences for *Scaphirhynchus* and *Acipenser*. Trans Am Fish Soc 129:1380–1388. https://doi.org/10.1577/1548-8659(2000)129%3c1380:maogvi%3e2.0.co;2

Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, Minamoto T, Yamamoto S, Yamanaka H, Araki H, Kondoh M, Iwasaki W (2015) MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than

230 subtropical marine species. R Soc Open Sci. https://doi.org/10.1098/rsos.150088

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, Mcveigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, Mcgarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, Dicuccio M, Kitts P, Murphy TD, Pruitt KD (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189

Olds BP, Jerde CL, Renshaw MA, Li Y, Evans NT, Turner CR, Deiner K, Mahon AR, Brueseke MA, Shirey PD, Pfrender ME, Lodge DM, Lamberti GA (2016) Estimating species richness using environmental DNA. Ecol Evol 6:4214–4226. https://doi.org/10.1002/ece3.2186

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes 7:355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Richter S, Schwarz F, Hering L, Böggemann M, Bleidorn C (2015) The utility of genome skimming for phylogenomic analyses as demonstrated for glycerid relationships (Annelida, Glyceridae). Genome Biol Evol 7:3443–3462. https://doi.org/10.1093/gbe/evv224

Sarri C, Stamatis C, Sarafidou T, Galara I, Godosopoulos V, Kolovos M, Liakou C, Tastsoglou S, Mamuris Z (2014) A new set of 16S rRNA universal primers for identification of animal species. Food Control 43:35–41. https://doi.org/10.1016/j.foodcont.2014.02.036

Satoh TP, Miya M, Mabuchi K, Nishida M (2016) Structure and variation of the mitochondrial genome of fishes. BMC Genomics 17:719. https://doi.org/10.1186/s12864-016-3054-y

Schrey AW, Sloss BL, Sheehan RJ, Heidinger RC, Heist EJ (2007) Genetic discrimination of middle Mississippi River *Scaphirhynchus* sturgeon into pallid, shovelnose, and putative hybrids with multiple microsatellite loci. Conserv Genet 8:683–693. https://doi.org/10.1007/s10592-006-9215-9

Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A, Zhou X (2014) Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. Nucleic Acids Res 42:e166–e166. https://doi.org/10.1093/nar/gku917

Tranah G, Campton DE, May B (2004) Genetic evidence for hybridization of pallid and shovelnose sturgeon. J Hered 95:474–480. https://doi.org/10.1093/jhered/esh077

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40:e115–e115. https://doi.org/10.1093/nar/gks596

Ward RD, Hanner R, Hebert PDN (2009) The campaign to DNA barcode all fishes, FISH-BOL. J Fish Biol 74:329–356. https://doi.org/10.1111/j.1095-8649.2008.02080.x

Zhu S-R, Ma K-Y, Xing Z-J, Xie N, Wang Y-X, Wang Q, Li J-L (2013) The complete mitochondrial genome of *Channa argus*, *Channa maculata* and hybrid snakehead fish [*Channa maculata* (♀) × *Channa argus* (♂)]. Mitochondrial DNA 24:217–218. https://doi.org/10.3109/19401736.2012.752469

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.