

Creating, curating and evaluating a mitogenomic reference database to improve regional species identification using environmental DNA

Emily Dziedzic¹  | Brian Sidlauskas¹  | Richard Cronn² | James Anthony³ |
Trevan Cornwell³ | Thomas A. Friesen³ | Peter Konstantinidis¹ | Brooke E. Penaluna² |
Staci Stein³ | Taal Levi¹ 

¹Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Corvallis, Oregon, USA

²Pacific Northwest Research Station, US Department of Agriculture Forest Service, Corvallis, Oregon, USA

³Oregon Department of Fish and Wildlife, Corvallis Research Laboratory, Corvallis, Oregon, USA

Correspondence

Emily Dziedzic, Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Corvallis, OR 97331, USA.

Email: emily.dziedzic@oregonstate.edu

Handling Editor: Aurélie Bonin

Abstract

Species detection using eDNA is revolutionizing global capacity to monitor biodiversity. However, the lack of regional, vouchered, genomic sequence information—especially sequence information that includes intraspecific variation—creates a bottleneck for management agencies wanting to harness the complete power of eDNA to monitor taxa and implement eDNA analyses. eDNA studies depend upon regional databases of mitogenomic sequence information to evaluate the effectiveness of such data to detect and identify taxa. We created the Oregon Biodiversity Genome Project to create a database of complete, nearly error-free mitogenomic sequences for all of Oregon's fishes. We have successfully assembled the complete mitogenomes of 313 specimens of freshwater, anadromous and estuarine fishes representing 24 families, 55 genera and 129 species and lineages. Comparative analyses of these sequences illustrate that many regions of the mitogenome are taxonomically informative, that the short (~150 bp) mitochondrial 'barcode' regions typically used for eDNA assays do not consistently diagnose for species and that complete single or multiple genes of the mitogenome are preferable for identifying Oregon's fishes. This project provides a blueprint for other researchers to follow as they build regional databases, illustrates the taxonomic value and limits of complete mitogenomic sequences and offers clues as to how current eDNA assays and environmental genomics methods of the future can best leverage this information.

KEYWORDS

eDNA, environmental DNA, environmental genomics, mtDNA, reference sequence database

1 | INTRODUCTION

The use of ambient genetic material—environmental DNA (eDNA)—to detect and identify metazoans in soil, air, marine environments and freshwater habitats is transforming biodiversity monitoring (Andersen et al., 2012; Clare et al., 2022; Deiner et al., 2016; Drummond et al., 2015; Hauck et al., 2019; Lim et al., 2016; Lynggaard et al., 2022; Port et al., 2016; Valentini et al., 2016; Yamamoto

et al., 2017). Monitoring methods using eDNA require comprehensive reference databases of sequence information for target and nontarget species in the clade of interest. The oft-cited lack of comprehensive, reliably vouchered sequence information for many species (Bohmann et al., 2014; Collins et al., 2013; Cordier et al., 2021; Porter & Hajibabaei, 2018; Schnell et al., 2010) exposes the need to build reference databases using standardized sample collection, data and specimen curation and data-sharing protocols (Goldberg

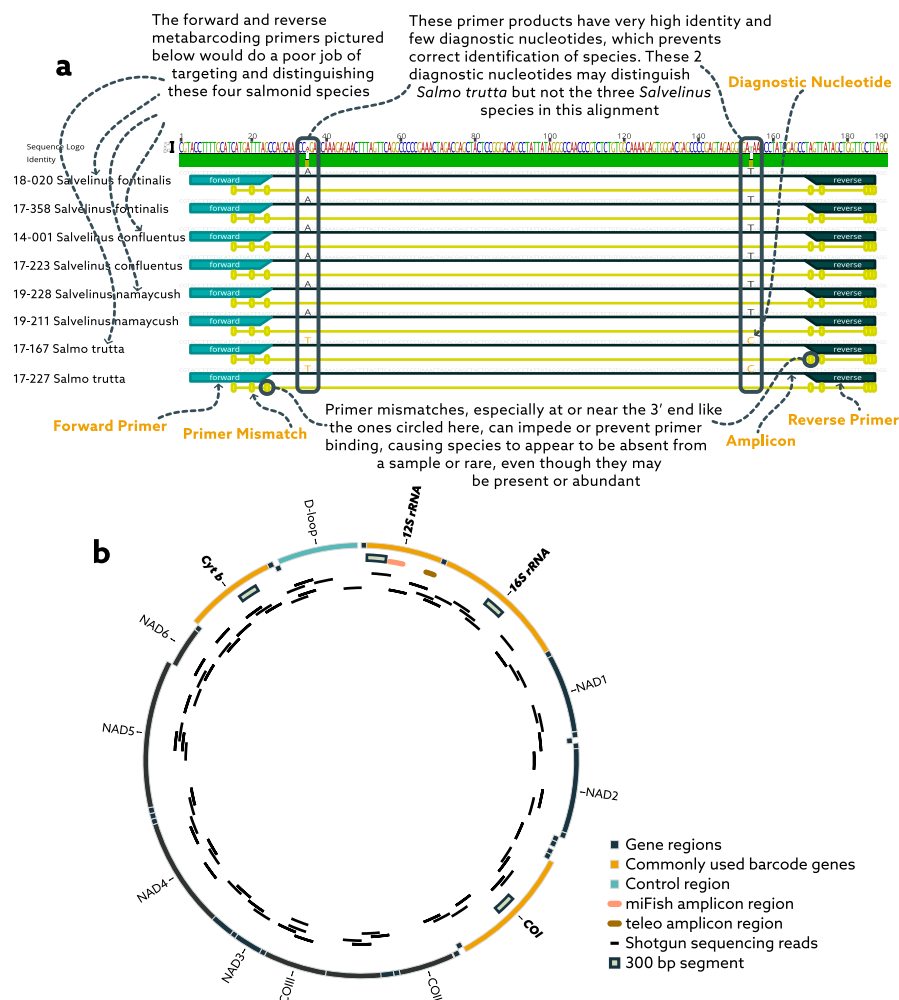
et al., 2016). This contribution documents and details the process of creating such a database for Oregon's freshwater fishes and provides a roadmap for others to follow in similar endeavours.

Molecular taxonomists have recommended microgenomic methods (metabarcoding, barcoding and single-species detection) for decades to work around limitations with morphology-based identification (Hebert, Cywinska, et al., 2003; Hebert, Ratnasingham, & de Waard, 2003), but molecular species detection methods do have some drawbacks. They rely on diagnostic sequence information from prototypical specimens to ensure correct identification of genetic material found in environmental samples, and that sequence information is not always available. Additionally, the gene- and taxon-specific primers needed to amplify barcode regions introduce a key source of error and bias in PCR amplification (Yang et al., 2021) because they select certain DNA sequences over others (Figure 1a; Deiner, Bik, et al., 2017). This primer bias not only allows researchers to sort metazoan targets from microorganismal backgrounds but also leads to unwanted loss of information when targeted populations and species amplify variably. Even minor binding biases among target sequences can affect PCR amplification substantially (Nichols et al., 2018; Piñol et al., 2015; Stadhouders et al., 2010), preventing reliable measurements of species presence

and/or relative abundance (Yang et al., 2021). Additionally, if species have not diverged at the locus targeted by a primer set, the assay will neither diagnose those taxa nor properly assess their presence or abundance (Figure 1a). Incomplete mitogenomic sequence information prevents in silico verification that primers will bind to species' DNA or that the amplicon will correctly diagnose species. In addition, missing sequence data and improper taxonomic assignments hinder accurate species identification when querying eDNA metabarcoding results.

Comprehensive databases of error-free, taxonomically verified, full mitogenomic data can solve issues related to unreliable genetic data and greatly improve the accuracy of novel environmental genomics methods that involve sequencing all the DNA in an environmental sample. Such approaches are known as 'shotgun sequencing', 'ecogenomics' or 'community genomics' (Béjà, 2004; Bragg & Tyson, 2014; Taberlet et al., 2012). Researchers focusing on animals typically target areas within the mitochondrial genome (mitogenome) for eDNA applications because mitochondrial DNA frequently diagnoses taxa accurately (Hebert, Cywinska, et al., 2003), resists environmental degradation (Foran, 2006) and is more easily recovered from degraded samples than lower copy nuclear DNA (Hartmann et al., 2011). Once isolated and

FIGURE 1 Primer issues and mapping shotgun reads. (a) The problems with primers. A sequence alignment of PCR products produced in silico illustrates why amplicons produced by some primer pairs, like the ones depicted here, might fail to detect or identify species. (b) Mapping shotgun reads. Samples that are shotgun-sequenced will produce reads from the entire mitogenome. This graphic depicts how, with whole mitogenomic information, it is possible to make use of all the data produced from shotgun sequencing. The length of barcodes is limited due to technological constraints and can only harness a small (maximum 300–500bp using Illumina next generation sequencing) fragment of genetic information from the mitogenome while far more genetic information can be accessed by genome skimming from shotgun sequencing, where all the DNA that is present in a sample is sequenced.



sequenced, whole mitogenomes can be used to assign taxonomic identifications in mitochondrial metagenomics (Crampton-Platt et al., 2016), multilocus metabarcoding (Arulandhu et al., 2017; Curd et al., 2019)—where multiple barcode markers are used to identify taxa in a sample—and ‘ultra-barcoding’ (Kane et al., 2012) also known as ‘super-barcoding’ (Li et al., 2015) where much longer barcodes or entire organelles are targeted. Mitogenomic approaches like these can help overcome key challenges with metabarcoding such as primer mismatches, which lead to taxonomic dropout (Cristescu & Hebert, 2018; Piñol et al., 2015; Sipos et al., 2010), reduced quantitative information (Bru et al., 2008; Wu et al., 2009) and incomplete taxonomic resolution (Piñol et al., 2015). For example, Tang et al. (2015) demonstrated that mapping shotgun-sequenced data to complete mitogenomes improved identification and quantitation of species in bee mock communities. Mitochondrial metagenomics can also harness more complete mitogenome sequences to infer phylogenetic relationships from bulk samples (Andújar et al., 2015; Crampton-Platt et al., 2015).

While advancements in sequencing technology have made it feasible to generate the voluminous data on which environmental mitogenomics depend, the lack of well-curated mitogenomic sequence databases presents a bottleneck (Arulandhu et al., 2017). Environmental mitogenomics depend on these databases because they allow matching of any mtDNA fragment to complete, taxonomically verified mitogenomes (Figure 1b; Crampton-Platt et al., 2016; Deiner, Renshaw, et al., 2017). When such databases exist, any recovered fragment can improve inference about abundance (Braukmann et al., 2019) and yield valuable information on species presence, which together with high spatial resolution can provide distributional information.

Existing genetic information in public reference databases can facilitate assay design, but issues with data collection make them potentially unreliable. GenBank® (Benson, 1996; Clark et al., 2016) cannot fill the need for curated reference databases because GenBank's sequence data are not uniformly linked to taxonomically verified vouchers. The lack of vouchers weakens the link between DNA sequence and taxonomic identity and prevents independent verification (Meiklejohn et al., 2019). In addition, GenBank does not always screen for contamination or check for errors in protein-coding sequences. Quality-checking at GenBank has improved (Leray et al., 2019), but the sequence data it holds can still be draft quality, contain errors and be incorrectly assigned taxonomically (Meiklejohn et al., 2019), particularly at the species or subspecies level (Locatelli et al., 2020). Identification errors may particularly plague speciose invertebrate groups (Leray et al., 2020). RefSeq (O'Leary et al., 2016), GenBank's curated and well-annotated sequence dataset, solves some of these problems by incorporating additional rounds of error-checking and provides information for the entire mitogenome. However, it is not comprehensive—when this study began, RefSeq contained sequence data for just 44% of Oregon's freshwater fish species and

these data were largely derived from specimens collected outside the state.

The Barcode of Life Data System (BOLD) provides an alternative to GenBank with more rigorous voucher requirements, consistent use of validation via translation to detect pseudogenes, contaminant checking and other tools to identify data anomalies and low-quality records (Ratnasingham & Hebert, 2007). However, BOLD skews heavily to information from cytochrome c oxidase I (COI) due to BOLD's initial development around a single >500bp COI barcode region. As of this writing, COI sequences represent 80.8% of the chordate data available and 82.4% for ray-finned fishes. Although remarkably diagnostic for many species, COI markers often fail to discern recently diverged sister species pairs and may fail to amplify certain taxa due to poorly conserved primer binding regions (Deagle et al., 2014). In some cases, other mitochondrial regions contain better conserved sites across taxa for primer placement and do a comparable job of distinguishing species. For example, Miya et al. (2015) found that two conserved regions flanking a hypervariable segment of the 12S mitochondrial gene provided more suitable fish metabarcoding primers.

Reference databases also require denser intraspecific sampling to identify loci that best diagnose species or geographically structured variation within species. Without multiple sequences for each species, researchers cannot test primer binding specificity and species diagnosability in silico. Here again, available reference sequence databases fall short. GenBank's curated RefSeq database, for example, is nonredundant by design (NCBI, 2021), with each species associated with only one complete mitogenome. Overall, the data gaps associated with GenBank and BOLD introduce uncertainty and potential error into the eDNA assay design process, making sole reliance on these resources for sequence data problematic.

Regional databases have the benefit of being able to control error-checking, curate full mitogenomic data, provide sequence information for genes and intergenic regions and identify and resolve taxonomic/genetic inconsistencies through resampling and revalidation (Astrin et al., 2013; Buckner et al., 2021; de Santana et al., 2021). Ideally, management-quality eDNA biodiversity surveys would extend the ‘BOLD model’ to create curated reference databases of complete mitogenome sequences tied to vouchered specimens collected throughout discrete regions. The need to expand the range of species with full mitogenomic sequence information has been elucidated by Langlois et al. (2021). The authors call for full mitochondrial genome sequences for multiple examples per species so that robust, comprehensive sequence alignments can support the development of assays that avoid cross-binding of primers to nontarget taxa or nonbinding of primers to target DNA (Langlois et al., 2021). In addition to improving qPCR primer design, databases of full mitogenomic sequence information provide the data needed to correctly assign metabarcoding sequence queries to taxa regardless of the primer binding site in the mitogenome. Finally, whole mitochondrial genomes also permit the exploration

of numerous markers for both quantitative and conventional PCR (Farrington et al., 2015). Fortunately, it is now practical and affordable for a small consortium to sequence and assemble hundreds of mitogenomes using a single Illumina Novaseq sequencing lane. This means that little impedes the development of the curated mitogenomic reference sequence databases needed to prepare for PCR-free mitogenomics and to develop, test and query single-species and metabarcoding eDNA assays.

Here, we provide a roadmap for constructing such a curated mitogenomic reference library and additionally evaluate the resulting data to determine the effectiveness of subsections of the mitogenome and the organelle in its entirety to identify species. This effort was motivated by the Oregon Biodiversity Genome Project (OBGP; www.obgp.org), a multi-institution collaboration between scientists and managers at Oregon State University, the Oregon Department of Fish and Wildlife and the United States Forest Service. The OBGP aims to develop a regional genetic reference database to facilitate statewide eDNA monitoring programmes for Oregon's resident freshwater fishes. The specific goals of the OBGP, as outlined in our roadmap (Figure 2), are to: (1) use sterile laboratory methods to collect 10 georeferenced full-bodied vouchers of each freshwater fish species from dispersed watersheds in Oregon; (2) archive and link voucher specimens, tissues and metadata for taxonomic verification and revision; (3) sequence full mitogenomes from multiple specimens per species; and (4) make all curated data publicly available. While biodiversity and geographic complexity differ from region to region, this study provides a realistic sense of the effort needed to construct a database covering ~150 species spread across ~250,000 km². By curating this reference database of full mitogenomes, we created the taxonomic reference information needed to identify freshwater, anadromous and estuarine fish species found in Oregon and bordering states by any mitogenome-based single-species eDNA or metabarcoding assay and set the stage for future PCR-free environmental mitogenomics methods. Our approach also analyses the efficacy of various regions of the mitogenome to identify species and provides pipelines that can guide other organizations as they develop reference sequence databases for their taxa and regions of interest.

2 | MATERIALS AND METHODS

2.1 | Voucher specimen and tissue collection

The study area initially encompassed the state of Oregon—the region of interest for our eDNA monitoring programme—and expanded to a few sites in northern California and Washington State (Figure 3). To strategize sample collection, we examined historical location records in fish collections, such as the Oregon State Ichthyology Collection, and conferred with local biologists to identify current distributions. For cases where we knew or suspected

that deeply divergent evolutionary lineages existed in the present concept of a species, we aimed to include representatives of all lineages. We ultimately identified 146 native and non-native freshwater fish species and lineages that are currently found in Oregon and strategized collections to span watersheds throughout the state (Appendix S1).

To facilitate consistent sampling, we provided sampling kits (Box S1 in Appendix S2) to collectors that contained a 500-mL Nalgene bottle filled with 10% formalin, a 2.0-mL cryotube filled with 95% EtOH, a sterile scalpel, scissors and tweezers, a bleach wipe, latex gloves, a detailed sampling protocol to ensure consistent tissue sampling and data collection (Box S2 in Appendix S2) and a field notes sheet (Box S3 in Appendix S2) for metadata collection. Collectors anaesthetised and euthanized all fish specimens before tissue collection by immersion in an aqueous solution of Tricaine mesylate (MS-222; 400 mg MS-222, 400 mg sodium bicarbonate, 1 L water). We instructed all partners to collect a minimum of ~0.5 cm³ of tissue from each specimen, which was then placed in 95% EtOH. Euthanized fish were placed in 10% formalin as vouchers, thereby ensuring preservation of diagnostic features. The cost outlay for kit components to collect 1500 individuals totalled US\$16,185.20 (Appendix S6).

2.2 | Taxonomic verification, accession and cataloguing

Fish biologists provisionally identified specimens in the field before Oregon State Ichthyology Collection taxonomists verified or refined those identifications by morphological examination and reference to published keys (Markle & Tomelleri, 2016; Wydoski & Whitney, 2003). The Oregon State Ichthyology Collection is in the process of accessioning and cataloguing all vouchers and tissues. During that process, the curators input the metadata associated with each specimen and collection event into a relational database, and full-bodied vouchers are transferred from formalin to isopropyl alcohol for permanent storage in a dedicated collection facility that complies with modern fire and earthquake safety codes. Tissues are stored in 2.0-mL cryotubes in -80°C freezers. Total expenditures on storage supplies such as jars, lids and preservation fluid came to US\$8624.21 (Appendix S6).

2.3 | Mitogenome assembly

To capture geographic genetic variation of each resident species across its distribution in Oregon, we sequenced the first collected representative of each species and subsequently sequenced specimens collected from separate watersheds, when possible. We stored gzipped FASTq sequencing files on 2 × 1 TB hard drives and performed mitogenome assemblies on 4 × 2.30 GHz 16-core processors using 512 GB RAM (total hardware cost \$12,368.33; Appendix S6). Mitochondrial genomes were assembled de novo

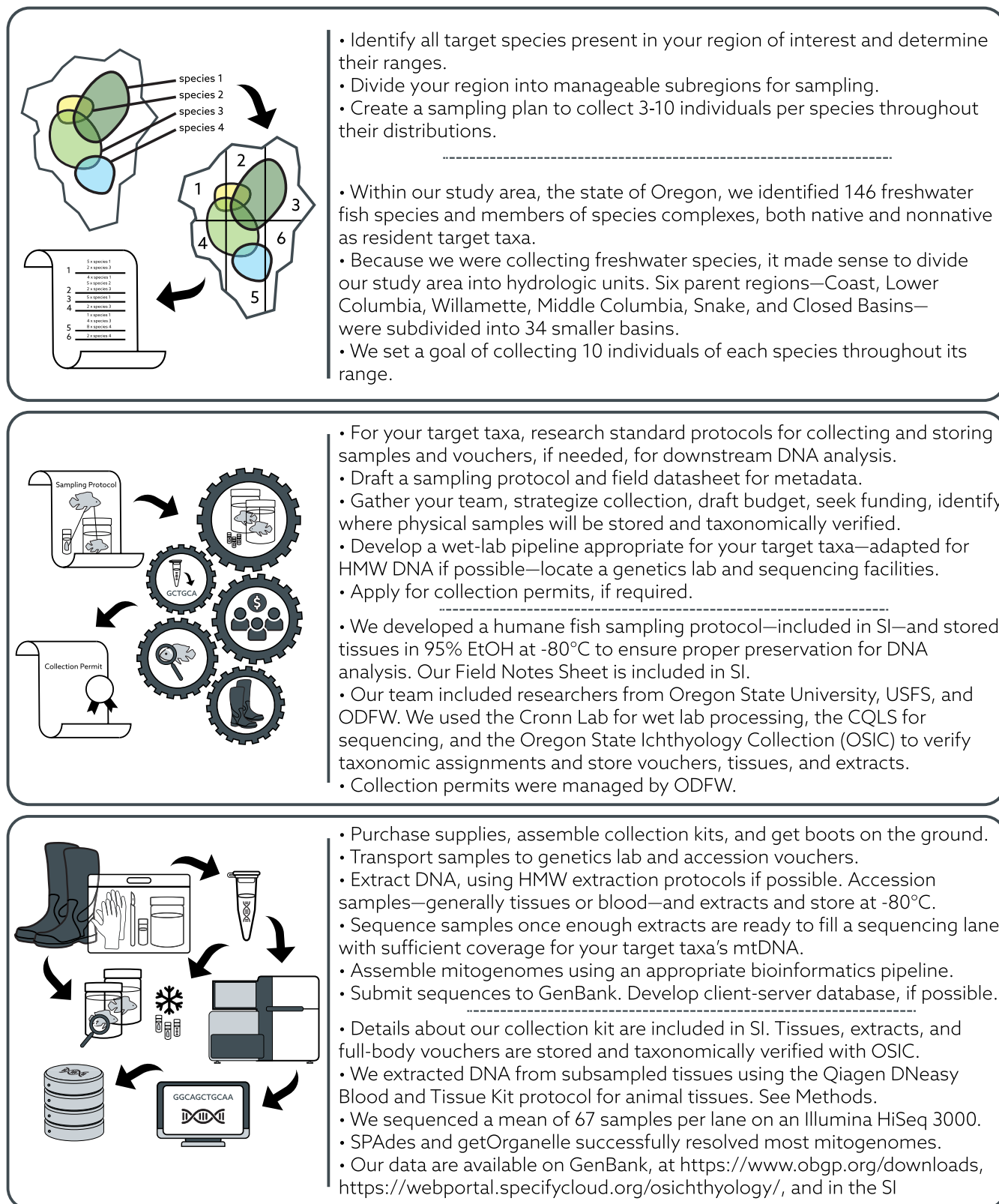


FIGURE 2 Creating a reference sequence database for monitoring with eDNA. A one-page blueprint designed to provide researchers with the basic steps to follow to create a regional database of full mitogenome sequences for target taxa.

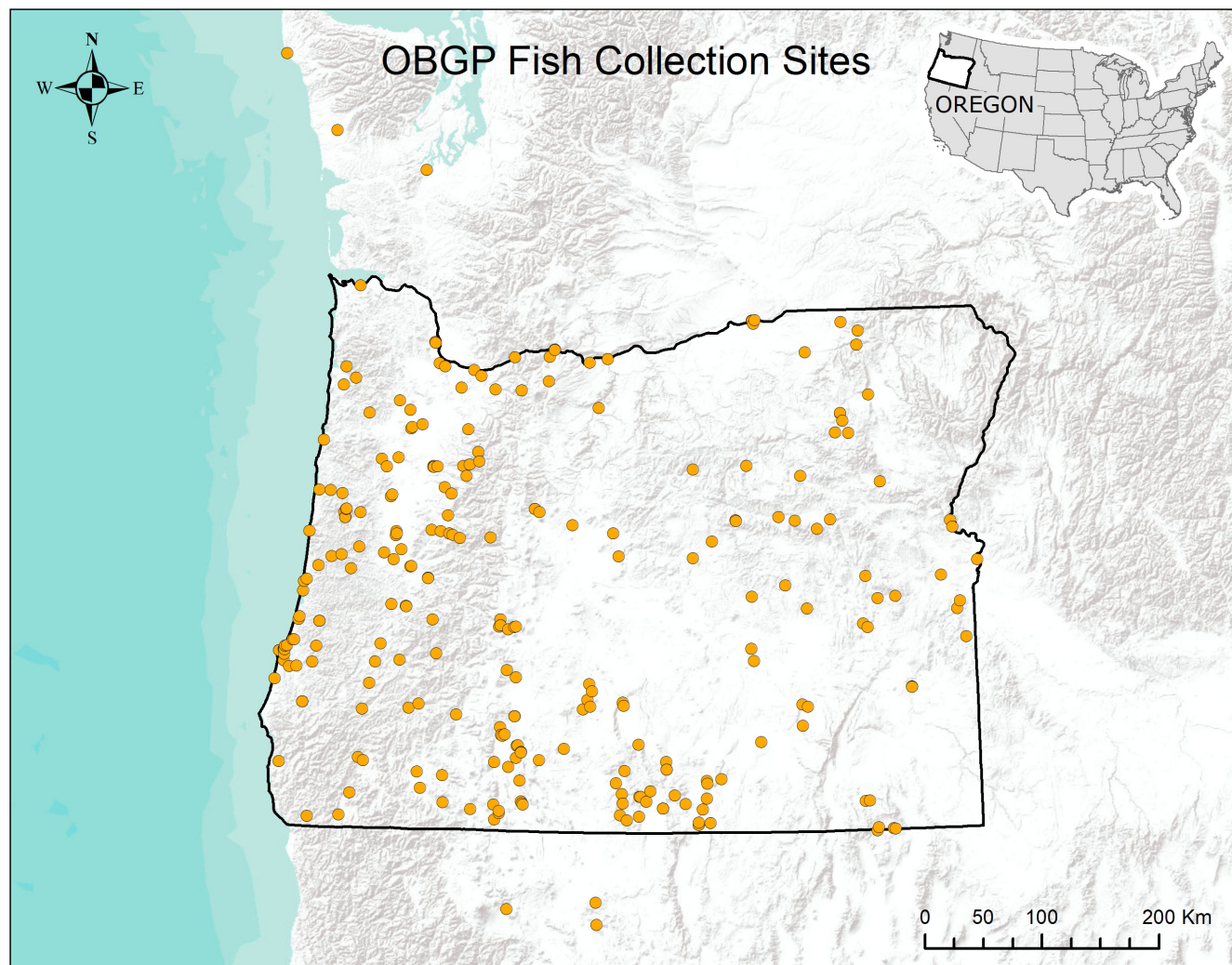


FIGURE 3 Map of study area and sampling sites. Each orange circle represents a single sampling location. An interactive map can be viewed [here](#).

from raw paired reads using SPAdes assembler (versions 3.12.0–3.15.3; Bankevich et al., 2012) and getOrganelle 1.6.2 or 1.7.5 (Jin et al., 2020) upon release. We annotated all mitochondrial sequences using a combination of MITOS² WebServer (Donath et al., 2019) and Geneious 10.2.6 using annotations from identical or closely related species.

2.4 | Mitogenome variability

To analyse intra- and interspecies mitogenome variability, assembled mitogenomes from each species were aligned with MUSCLE (Edgar, 2004) in Geneious 10.2.6 using default parameters. After reciprocal rounds of morphological examination and molecular clustering, we aligned sequences of species from within the same family and then aligned these family clusters to create a master alignment of all sequences. To identify taxonomically diagnostic regions for efficient eDNA assay development, we used the R package SPIDER (Brown et al., 2012) to perform a sliding window analysis on the master

alignment and locate areas with the highest density of taxonomically diagnostic nucleotides (TDN)—defined as locations where a nucleotide is fixed within species and different or unaligned in all other species. We determined mean taxonomically diagnostic nucleotides per 150-base window shifted at 20-base intervals ($TDN/w_{150i_{20}}$) by calculating the mean of all $TDN/w_{150i_{20}}$ counts across a particular region. We also identified regions containing TDN 'spikes', where $\max(TDN/w_{150i_{20}}) > 2 * \text{mean}(TDN/w_{150i_{20}})$. To identify species and genes with high variability, we used identity matrices generated in Geneious and plotted variability with heat maps, parallel coordinate plots and radar charts using the R packages Superheat (Barter & Yu, 2018), GGally (Schloerke et al., 2021) and fmsb (Nakazawa, 2021), respectively. Gene regions <690 base pairs in length—ATP6, ATP8, NAD3, NAD4l, NAD6 and all tRNA genes—were not included in our analyses of individual genes. We grouped described subspecies with their full parent species for the purpose of calculating mean percent identities. Calculation of intraspecies, intrafamily/interspecies and interfamily/interspecies mitogenome identities and the proportional relationships among these identities required mitogenomes of multiple specimens for each species within

a family. Seven families satisfied these requirements: Catostomidae, Centrarchidae, Cottidae, Cyprinidae, Ictaluridae, Petromyzontidae and Salmonidae.

2.5 | Utility of different mitogenomic regions for species identification

We evaluated the relative success of subregions of the mitogenome to identify fish species by first extracting subsets from the alignment of 313 mitogenomes. In Geneious 10.2.6, we created miFish (Miya et al., 2015) and Teleo (Valentini et al., 2016) amplicons using Primer3 (Untergasser et al., 2012) and extracted 12S, 16S, NAD2, CO1, NAD4, NAD5 and D-loop regions from the alignment using annotations as guidelines. We examined the entire mitogenome along with regions spanning the 12S, trnV and 16S genes and the NAD4, trnH, trnS1, trnL1 and NAD5 genes. We performed BLASTn queries of these regions against a local database created from our 313 mitogenomes. We parsed the results from this BLASTn query to determine the effectiveness of the different regions to successfully identify species.

2.6 | Data sharing

Mitogenome data generated for this project have been deposited in GenBank under the Oregon Biodiversity Genome Project BioProject. GenBank accession numbers and sequence data are included in Supplemental Information (Appendices S3 and S7 contain mitogenome FASTAs). Sequence data are also available at www.obgp.org/downloads. As of the time of this writing, linked voucher, tissue and DNA extract accessioning into the Oregon State Ichthyology Collection was ongoing. The voucher specimens accessioned and catalogued to date are available and searchable via webportal at <https://webportal.specifycloud.org/osichthyology/>, which the collection's curators updated periodically.

Further details of voucher specimen and tissue collection, taxonomic verification, accession and cataloguing, mitogenome assembly and utility of different mitogenomic regions for species identification are available in Appendix S8. Information regarding DNA extraction and sequencing is solely available in Appendix S8.

3 | RESULTS

3.1 | Voucher specimen and tissue collection

We collected 625 specimens representing 129 fish species or distinct lineages within species complexes. These specimens originated from more than 240 localities in Oregon. Twelve additional tissue samples of four species were acquired from natural history collections, bringing the total number of species represented in the database to 132. Of these 132 species, 119 represent the original 146

fish species identified by Oregon Department of Fish and Wildlife as native or naturalized in Oregon. The remaining 13 species belong to 11 coastal estuarine species not included in our initial freshwater collection plan, plus one species endemic to western Washington (Olympic Mudminnow, *Novumbra hubbsi*) and one newly identified lineage of Paiute Sculpin (*Cottus beldingii* ssp.) from the John Day River Basin in central Oregon.

3.2 | Taxonomic verification

After clustering sequences based on maximum-likelihood (ML) inference and examining voucher specimens in the laboratory, we refined or corrected 31 field identifications (9.9%; Appendix S3; Figures S1–S8 in Appendix S2). Specimens from Cottidae represented the highest proportion of conflicting identifications ($n=16$; 52%) likely because the group is the subject of active taxonomic research to identify the geographic, genetic and anatomical limits of their component evolutionary lineages (Young et al., 2022). The most current morphological identification keys to *Cottus* do not reliably separate all the species and species complexes in this genus.

Of the 31 total refinements, 15 (48%; Cottidae, $n=12$; Catostomidae, $n=1$; Petromyzontidae, $n=1$; Ictaluridae, $n=1$) had ambiguous or conflicting taxonomic assessments and were assigned to a species or species complex based on ML inference. Of the cottids, nine resided in the *Cottus gulosus/perplexus* complex and were assigned to either *C. perplexus* ($n=8$) or *C. gulosus* ($n=1$), while OBG-2017-269 resided in the *Cottus asper/perplexus* complex and was assigned to *C. asper*. OBG-2017-148 matched some but not all the features normally diagnosing *C. beldingii* and clustered with the newly identified *C. beldingii* lineage, and OBG-2018-012 was assigned to *C. confusus* as this was suggested by ML inference and supported by the morphological identification key for the species. Of the three remaining specimens with ambiguous taxonomic assignments, the sucker specimen OBG-2018-023 clustered with *Catostomus bondi* yet lacked all the diagnostic features of this species and was assumed to be a hybrid with *Catostomus columbianus*. OBG-2016-007 was a transforming lamprey microphthalmia with caudal fin pigmentation suggesting *Entosphenus lethophagus*, and a catfish specimen OBG-2018-046 that had some of the diagnostic features of both *Ameiurus natalis* and *A. nebulosus* but clustered with *A. nebulosus* was assumed to be a hybrid.

Morphological assignment superseded ML clustering for four individuals (13%) consisting of three cottids (OBG-2017-273, OBG-2017-346 and OBG-2019-178) and one catostomid (OBG-2017-101). ML clustering disagreed with in-field assignment and was concordant with morphological assessment for six specimens (19%). These consisted of challenging-to-identify species from Centrarchidae (OBG-2017-308, OBG-2017-381 and OBG-2018-159) and Petromyzontidae (OBG-2017-248 and OBG-2017-250), along with a single cyprinid that was a xanthic morph of *Pimephales promelas* (OBG-2018-033). Three vouchers (10%), all centrarchids, were of poor quality and made morphological

identification challenging (OBGP-2017-275, OBGP-2017-360 and OBGP-2019-057). An additional two catostomids (7%; OBGP-2018-006 and OBGP-2018-092) appeared to have been switched during processing, and one cottid (3%) was erroneously noted to be a subspecies of *Cottus klamathensis* (OBGP-2017-218) during wet lab processing.

3.3 | Mitogenome sequencing and assembly

In total, 313 assembled mitogenomes representing 129 collected species and lineages were used for downstream analysis (Table 1; Table S3 in Appendix S4). Most de novo assemblies (96.8%; $n=303$) resolved as a single mitochondrial contig with an overlapping splice point. The remaining assemblies were derived from either: (a) multiple contigs with overlapping splice points ($n=3$); (b) a single contig with a nonoverlapping splice point in an intergenic area with mononucleotide C repeats ($n=6$); or (c) multiple contigs from different SPAdes runs with overlapping splice points ($n=1$). All mitogenomes had GC content between 38.90% and 49.50% (mean 45.14%) except for within Petromyzontidae, where GC content ranged from 37.90% to 38.70% (mean 38.05%; Appendix S3). Mitogenome sizes ranged from 16,098 to 17,185 bp in length (mean 16,590). All but 17 assembled mitogenomes had error-free contigs when measured with $k=31$ using Merqury. Assembled mitogenomes with errors had QVs between 40.7507 and 57.0952 (Appendix S3) indicating errors in the range of 1 in ~10,000 bp to 1 in ~1,000,000 bp, respectively. Read mapping showed anomalous coverage in intergenic regions of 36 assemblies. These anomalies were generally located in areas with nucleotide repeats and manifested as dips or spikes in coverage. Dips were likely due to sequencing and/or assembly errors, and we assumed spikes were the result of nuclear mitochondrial DNA. These anomalies were not sufficient to exclude mitogenomes from downstream analyses (Appendix S3).

3.4 | Mitogenome variability

The sliding window analysis of our alignment of 313 complete mitogenomes revealed that mean taxonomically diagnostic nucleotides per 150-base window shifted at 20-base intervals ($TDN/w_{150}i_{20}$) in analysed gene regions were as follows: COI (7.257), CytB (9.451), NAD1 (13.381), NAD5 (17.092), NAD4 (18.226), NAD2 (20.065), 12S (20.814) and 16S (25.726) (Figure 4b). The highest concentrations of TDNs occurred in the D-loop and the intergenic region between the NAD2 and COI genes (Figure 4a). Sliding window analyses of aligned taxonomic subsets of the entire mitogenome suggested that the density of TDN varies by taxonomic group (Figure 5) with mean $TDN/w_{150}i_{20}$ values as follows: catostomids (10.656), centrarchids (26.082), cottids (18.871), cyprinids (19.316), ictalurids (21.024), petromyzontids (12.104) and salmonids (20.209). TDN spikes occurred in the

D-loop of all families analysed and the 16S region of all families apart from Ictaluridae. These spikes also arose in the 12S region of Salmonidae and Cottidae, the COI region of Cyprinidae and the NAD5 region of Centrarchidae (Table 2). A sliding window analysis of 274 mitogenomes that excluded singletons exhibited the same pattern with TDN concentrations dispersed across the entire mitogenome with the highest concentration of TDNs occurring in the 16S region (Figure S11 in Appendix S4). Mean TDNs for analysed gene regions in this alignment were as follows: COI (8.069), CytB (10.863), NAD1 (12.952), NAD5 (15.291), NAD4 (16.714), 12S (17.953), NAD2 (18.739) and 16S (29.155).

We used heat maps (Figure 6a; Figure S10 in Appendix S4; raw data in Appendix S5) to illustrate the degrees of similarity between families in different gene regions, although all genes have diverged sufficiently to diagnose familial lineages. These results showed that sequence identity among families in the COI gene exceeded that of other coding and noncoding gene regions (Figure 6a) suggesting that COI is the mitogenomic region in which families differ the least, a conclusion that concurs with the results from the sliding window analysis. NAD2, NAD5 and 16S contrasted the most among families and species in overall percent identity. Despite different degrees of intrafamily identity in the mitochondrial regions we examined, our analysis suggested that divergence in all mitochondrial genes and the D-loop is sufficient to identify taxa at the family level (Figure 6a).

Zooming in to species differences within families, after in-field identification errors were corrected, all species resembled members of their own species in mtDNA sequences more than they resembled members of other species within the same family, as expected (Table S1 in Appendix S4). Species within the same family differed most in the percent identity of the D-loop (0.855), followed by the NAD2 (mean 0.885), NAD5 (0.896) and NAD4 (0.896) genes. The 12S (0.965) and 16S (0.957) genes were the least differentiable regions (Figure 7). Intraspecies mean percent identities for the 12S and 16S rRNA genes and all coding genes >690 bp ranged from 98.259% to 99.975% and from 92.428% to 98.249% for the D-loop (Table S2 in Appendix S4). This comparison illustrates that the D-loop is less conserved within species than are any of the other gene regions (Figure 8). The most conserved genes were 12S, 16S and COX2, with the lowest mean values found in the NAD2 gene nevertheless still exceeding 99% identity (Table S2 in Appendix S4). Radar charts of mean percent intraspecies, intrafamily interspecies and interfamilial interspecies identity (Figure 8) illustrated that different genes are more conserved among species within certain families than others. For example, species in Catostomidae varied little in sequences from rRNA and all three COX genes, while the 12S and 16S genes were fairly conserved among salmonid and cottid species. Non-rRNA regions in Salmonidae and Cottidae, and all gene regions and the D-loop in Cyprinidae, Centrarchidae and Ictaluridae contained diverged interspecies sequences. Full mitogenomes were highly conserved within species (mean 99.493% identity) and had sufficient divergence among species in the same family to suggest

TABLE 1 Assembled Mitogenome Taxa Counts: OBGp specimens with assembled mitogenomes are grouped according to taxonomic designation with counts for each taxonomic level (see [Table S3](#) in Appendix [S4](#) for table with individual specimen IDs).

Family		Genus		Species		Subspecies/Lineage	
Acipenseridae	3	<i>Acipenser</i>	3	<i>medirostris</i>	1		
				<i>transmontanus</i>	2		
Atherinopsidae	2	<i>Atherinops</i>	2	<i>affinis</i>	2		
Catostomidae	61	<i>Catostomus</i>	48	<i>bondi</i>	3		
				<i>columbianus</i>	1		
				<i>macrocheilus</i>	2		
				<i>microps</i>	9		
				<i>occidentalis</i>	8	<i>lacusanserinus</i>	5
				<i>rimiculus</i>	6	Jenny Creek	3
						Klamath	2
						Rogue	1
				<i>snyderi</i>	10		
				<i>tahoensis</i>	1		
				<i>tsiltcoosensis</i>	6	Coos	2
						Coquille	1
						Siuslaw	1
						Umpqua	2
				<i>warnerensis</i>	2		
		<i>Chasmistes</i>	2	<i>brevirostris</i>	2		
		<i>Deltistes</i>	11	<i>luxatus</i>	11		
Centrarchidae	24	<i>Archoplites</i>	2	<i>interruptus</i>	2		
		<i>Lepomis</i>	10	<i>cyanellus</i>	1		
				<i>gibbosus</i>	5		
				<i>gulosus</i>	1		
				<i>macrochirus</i>	3		
				<i>microlophus</i>	1		
		<i>Micropterus</i>	7	<i>dolomieu</i>	2		
				<i>salmoides</i>	5		
		<i>Pomoxis</i>	5	<i>nigromaculatus</i>	5		
Clupeidae	2	<i>Alosa</i>	2	<i>sapidissima</i>	2		
Cobitidae	2	<i>Misgurnus</i>	2	<i>anguillicaudatus</i>	2		
Cottidae	46	<i>Cottus</i>	45	<i>aleuticus</i>	2		
				<i>asper</i>	4		
				<i>bairdii</i>	1		
				<i>beldingii</i>	6	ssp.	3
				<i>bendirei</i>	1		
				<i>confusus</i>	2		
				<i>gulosus</i>	5		
				<i>klamathensis</i>	3		
				<i>marginatus</i>	2		
				<i>perplexus</i>	10		
				<i>pitensis</i>	2		
				<i>polyporus</i>	1		
				<i>princeps</i>	1		
				<i>rhotheus</i>	3		
				<i>tenuis</i>	2		
		<i>Enophrys</i>	1	<i>bison</i>	1		

TABLE 1 (Continued)

Family		Genus		Species		Subspecies/Lineage	
Cyprinidae	79	<i>Acrocheilus</i>	2	<i>alutaceus</i>	2		
		<i>Carassius</i>	2	<i>auratus</i>	2		
		<i>Cyprinus</i>	2	<i>carpio</i>	2		
		<i>Gila</i>	3	<i>coerulea</i>	3		
		<i>Hesperoleucus</i>	2	<i>symmetricus</i>	2		
		<i>Mylocheilus</i>	1	<i>caurinus</i>	1		
		<i>Notemigonus</i>	1	<i>crysoleucas</i>	1		
		<i>Oregonichthys</i>	5	<i>crameri</i>	2		
				<i>kalawatseti</i>	3		
		<i>Pimephales</i>	3	<i>promelas</i>	3		
		<i>Ptychocheilus</i>	5	<i>oregonensis</i>	2		
				<i>umpquae</i>	3	Siuslaw	1
		<i>Rhinichthys</i>	26	<i>catractae</i>	4		
						'Millicoma Dace'	1
				<i>evermanni</i>	2		
				<i>falcatus</i>	2		
				<i>osculus</i>	16	Black Lined	5
						Closed Basin	2
						Foskett Spring	2
						Klamath	2
						Malheur Stream	2
						Stinking Lake Spring	1
						Western	2
				<i>umatilla</i>	2		
		<i>Richardsonius</i>	14	<i>balteatus</i>	12	<i>balteatus</i>	5
						<i>hydrophlox</i>	2
						Siuslaw	5
		<i>Siphateles</i>	12	<i>egregius</i>	2		
				<i>alvordensis</i>	2		
				<i>bicolor</i>	8	<i>eurysoma</i>	1
						Klamath	2
						<i>obesus</i>	2
Embiotocidae	3	<i>Cymatogaster</i>	1	<i>aggregata</i>	1		
		<i>Embiotoca</i>	1	<i>lateralis</i>	1		
		<i>Phanerodon</i>	1	<i>furcatus</i>	1		
Esocidae	1	<i>Esox</i>	1	<i>lucius</i>	1		
Fundulidae	2	<i>Fundulus</i>	2	<i>diaphanus</i>	2		
Gasterosteidae	1	<i>Gasterosteus</i>	1	<i>aculeatus</i>	1		
Ictaluridae	10	<i>Ameiurus</i>	7	<i>natalis</i>	3		
				<i>nebulosus</i>	4		
		<i>Ictalurus</i>	2	<i>punctatus</i>	2		
		<i>Noturus</i>	1	<i>gyrinus</i>	1		

(Continues)

TABLE 1 (Continued)

Family		Genus		Species		Subspecies/Lineage	
Osmeridae	2	<i>Hypomesus</i>	1	<i>pretiosus</i>	1		
		<i>Thaleichthys</i>	1	<i>pacificus</i>	1		
Oxudercidae	1	<i>Rhinogobius</i>	1	<i>brunneus</i>	1		
Percidae	4	<i>Perca</i>	3	<i>flavescens</i>	3		
		<i>Sander</i>	1	<i>vitreus</i>	1		
Percopsidae	1	<i>Percopsis</i>	1	<i>transmontana</i>	1		
Petromyzontidae	13	<i>Entosphenus</i>	12	<i>lethophagus</i>	3		
				<i>minimus</i>	1		
				<i>similis</i>	3		
				<i>tridentatus</i>	5		
		<i>Lampetra</i>	1	<i>richardsoni</i>	1		
Pholidae	2	<i>Apodichthys</i>	1	<i>flavidus</i>	1		
		<i>Pholis</i>	1	<i>ornata</i>	1		
Pleuronectidae	3	<i>Parophrys</i>	1	<i>vetulus</i>	1		
		<i>Platichthys</i>	2	<i>stellatus</i>	2		
Poeciliidae	2	<i>Gambusia</i>	2	<i>affinis</i>	2		
Salmonidae	47	<i>Oncorhynchus</i>	27	<i>clarkii</i>	8		
						<i>henshawi</i>	3
						<i>lewisi</i>	2
				<i>keta</i>	2		
				<i>kisutch</i>	3		
				<i>mykiss</i>	9	<i>gairdneri</i>	3
						<i>irideus</i>	3
						<i>newberrii</i>	3
				<i>nerka</i>	2		
				<i>tshawytscha</i>	3		
		<i>Prosopium</i>	2	<i>williamsoni</i>	2		
		<i>Salmo</i>	3	<i>trutta</i>	3		
		<i>Salvelinus</i>	15	<i>confluentus</i>	6		
				<i>fontinalis</i>	6		
				<i>namaycush</i>	3		
Sebastidae	1	<i>Sebastes</i>	1	<i>caurinus</i>	1		
Umbridae	1	<i>Novumbra</i>	1	<i>hubbsi</i>	1		

they would be diagnostic at the species level for Oregon fishes (Figure 9).

3.5 | Utility of mitogenomic regions for species identification

Plots of parsed results from BLASTn queries suggest that the full mitogenome, concatenated gene regions from the mitogenome, and the NAD5 gene are superior to miFish and Teleo metabarcoding primers and most individual mitochondrial genes for producing first hits that match specimens to species and to described subspecies (Figure 10). Nevertheless, queries using every gene region and

Teleo and miFish primer amplicons produce significant E-value hits matching specimens to the species or subspecies level 100% of the time. Shorter amplicons are not as successful at producing top E-value matches as longer gene regions. For full mitogenomes, the top E-value hit always matches the target species or subspecies in 99.225% of queried specimens. All other regions do an inferior job of always producing top E-value hits that match the target species or subspecies. Percent of specimens queried that had top E-value hits matching the species or subspecies are as follows, in order of increasing success: teleo (68.992), miFish (75.969), 12S (86.047), NAD4 (92.248), COXI (93.023), D-loop (93.798), 16S (93.798), NAD2 (96.124), NAD5 (96.899), concatenated NAD4 and NAD5 (96.899) and concatenated 12S and 16S (98.450).

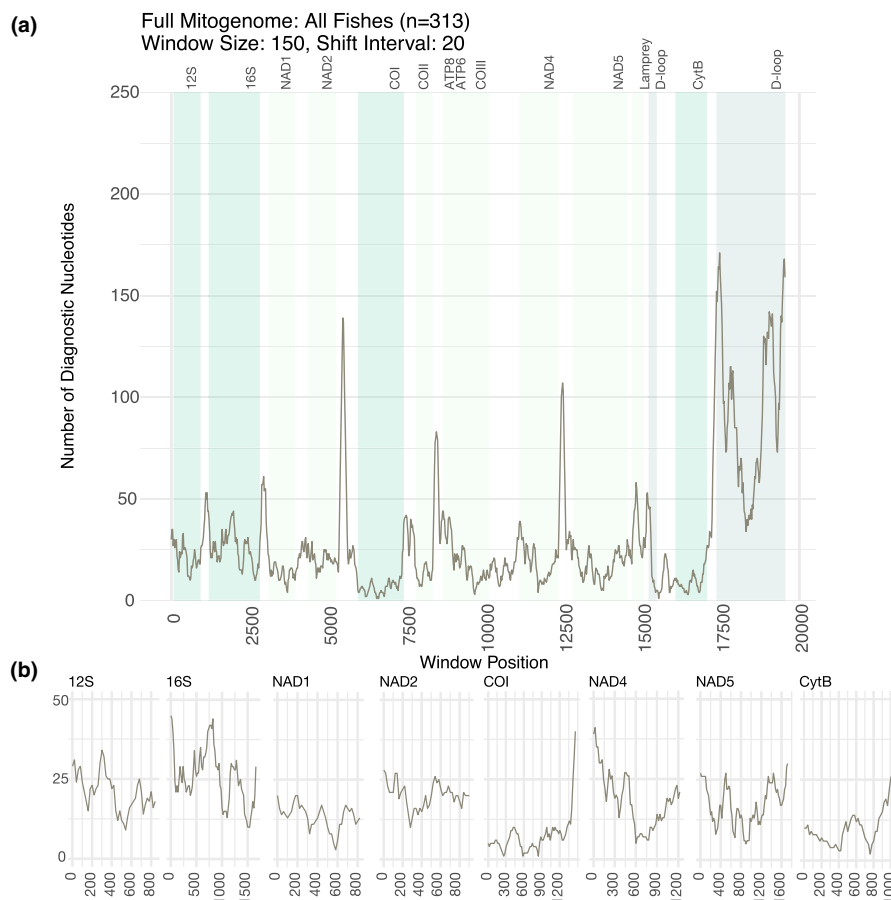


FIGURE 4 Sliding window analysis: A 150-base-wide window is placed at the beginning of an alignment of 313 mitogenome sequences and shifted right at 20 base intervals. At each window position, the number of taxonomically diagnostic nucleotides (TDN)—where a nucleotide is shared within a species but is either different or unaligned with other species—is counted. Areas with high concentrations of TDNs are likely to be superior for species identification in eDNA assays. (a) Full mitogenome: Gene regions and the D-loop are shaded. Areas with clusters of TDNs are distributed across the entire mitogenome and the highest TDN concentrations are in noncoding regions suggesting the whole genome is of potential use. (b) Individual genes: These plots zoom in on a subset of individual genes to focus on the number of TDNs within 8 barcode regions. The 16S region has the highest taxonomically diagnostic nucleotides per 150-base window shifted at 20-base intervals. Mean TDN/ $w_{150}i_{20}$ in each region: COI, 7.257; CytB, 9.451; NAD1, 13.381; NAD5, 17.092; NAD4, 18.226; NAD2, 20.065; 12S, 20.814; 16S, 25.726.

4 | DISCUSSION

4.1 | A blueprint for constructing mitogenomic databases

We demonstrate a robust, affordable, feasible blueprint for constructing mitogenomic databases. The workflow begins with the collection of reference specimens and progresses through taxonomic verification, accessioning of specimens, tissues and DNA, mitogenome assembly and open-source provisioning of complete mitogenomes. Such databases can help to refine the taxonomy of understudied or difficult groups, guide the discovery and delineation of cryptic species or distinct population segments and facilitate the transition to eDNA-based monitoring of aquatic biodiversity.

Though beginning such a project appears daunting, the steps for carrying out a similar endeavour are straightforward: (1) using historical collection data and local knowledge, determine all focal

species and their distributions, (2) break up the region of interest into manageable subregions for sampling, (3) create a sampling plan to collect 3–10 individuals per species/lineages of interest and begin the sampling effort using accepted standards for metadata collection (Rimet et al., 2021), acquiring tissues from vouchered specimens in natural history collections whenever possible and (4) sequence and assemble specimens as they accumulate, measuring intraspecific sequence variability to inform continued collection.

We have catalogued the collection strategy and wet and dry laboratory pipelines we used for our bottom-up development of an eDNA biodiversity reference collection and sequence database and provide an easy-to-follow roadmap (Figure 2). This bottom-up approach harnesses the expertise, knowledge and resources of researchers and managers within their region and taxa of interest, an essential strategy as these individuals possess the intimate knowledge of species, taxonomy and geography needed to plan expeditions, carry out collections and identify specimens.

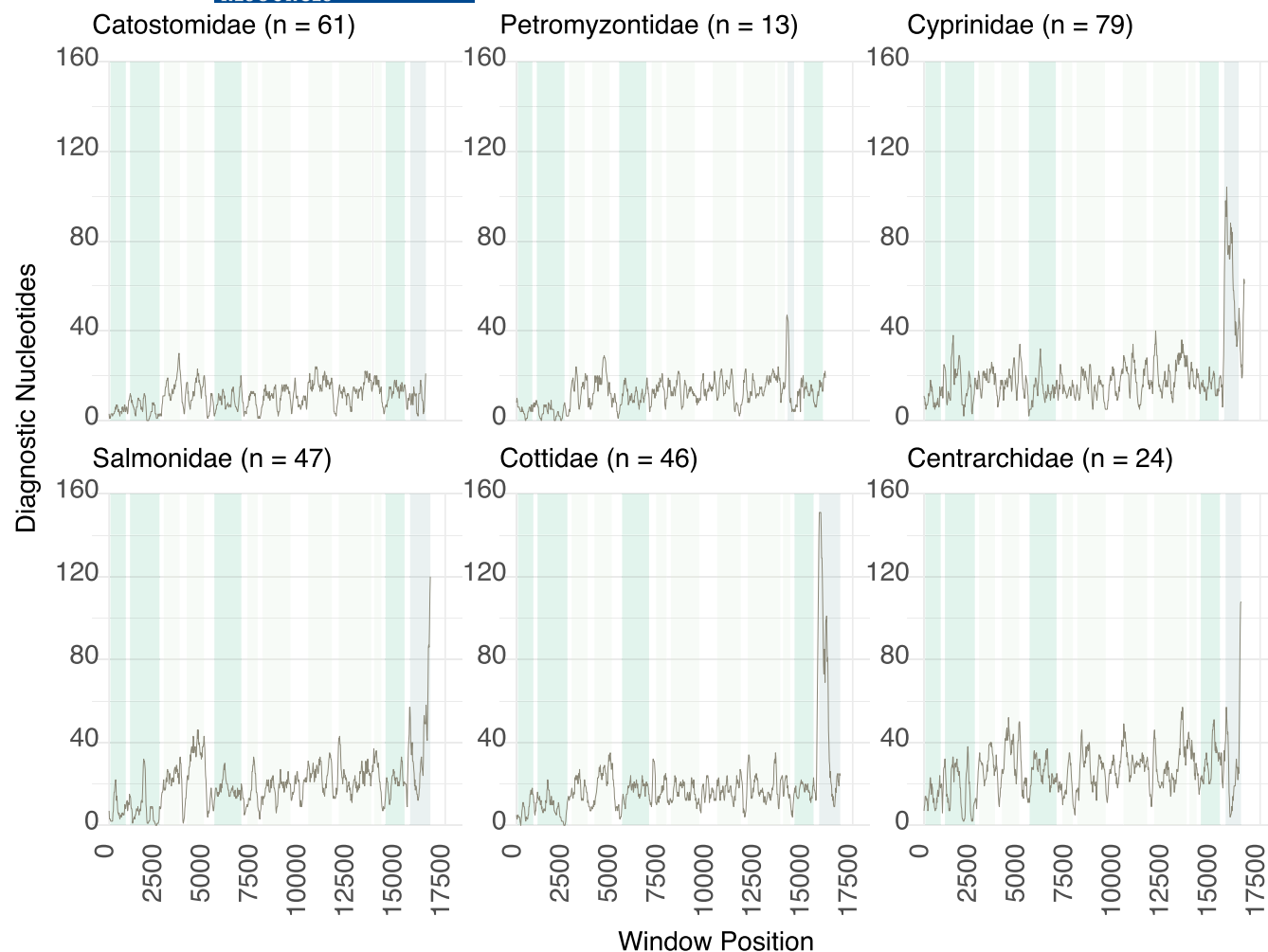


FIGURE 5 Sliding window analysis, by family, full mitogenome. The full alignment of 313 specimens is separated into individual families and a window 150bp in length is placed along the length of an alignment of each family alignment and moved at 20-bp intervals. At each position, the number of diagnostic nucleotides—where a base is shared within a species but is either different or unaligned with other species—is counted. This illustrates that different families have different levels of variability across the mitogenome. Commonly used barcode genes are highlighted in deep bluegreen, from left to right, 12S, 16S, COI, CytB. Petromyzontidae mitogenome is structured with its control region upstream of the CytB gene. Means (TDN/ $w_{150i_{20}}$): Catostomidae, 10.656; Petromyzontidae, 12.104; Cottidae, 18.871; Cyprinidae, 19.316; Salmonidae, 20.209; Centrarchidae, 26.082.

TABLE 2 Taxonomically diagnostic nucleotides (TDN) within families: For each of 7 families, maximum and mean TDNs in a 150-base window shifted at 20-base intervals along an intrafamily alignment of mitochondrial gene regions are listed here.

	Salmonidae			Cyprinidae			Catostomidae		
	Max	Mean	id_prop	Max	Mean	id_prop	Max	Mean	id_prop
rrnS	<u>21</u>	8.050	0.971	17	9.220	0.961	8	4.275	0.991
rrnL	<u>31</u>	7.727	0.970	<u>38</u>	15.962	0.926	<u>13</u>	5.481	0.985
nad1	30	23.452	0.882	28	18.952	0.849	29	16.167	0.950
nad2	47	35.022	0.871	29	18.244	0.826	23	13.844	0.940
cox1	29	17.634	0.908	<u>31</u>	14.380	0.892	18	9.014	0.973
cox2	<u>32</u>	18.821	0.932	19	13.429	0.906	13	9.536	0.977
cox3	27	20.281	0.912	21	13.469	0.890	17	12.344	0.970
nad4	35	25.710	0.885	33	19.645	0.846	24	17.435	0.954
nad5	43	23.318	0.889	39	21.988	0.832	22	13.588	0.958
cytb	<u>33</u>	21.300	0.894	24	15.660	0.864	18	13.200	0.952
d-loop	<u>118</u>	38.755	0.792	<u>104</u>	57.040	0.791	<u>21</u>	9.927	0.948

Note: TDN 'spikes', where $\max(\text{TDN}) > 2 * \text{mean}(\text{TDN})$ are underlined and in bold. Proportional relationship between mean within-family intraspecies and interspecies identity (id_prop) is also listed.

Because projects of this scale require moderate financial support and substantial human effort, we strongly recommend assessing available resources before launching a new endeavour. We were able to complete this project on a relatively low budget (estimated cost without labour ~\$250 per mitogenome; Appendix S6) because individuals donated considerable amounts of their time collecting throughout the state of Oregon and because collaborating institutions provided genetic laboratory facilities and sequencing at reduced costs.

There are many factors to consider when constructing a genomic database. Contamination must be prevented during tissue sampling, and voucher archiving is key for taxonomic verification (Rimet et al., 2021). Adequate tissue volumes must be harvested and properly preserved (Nagy, 2010) to derive sufficient volumes of high-quality DNA for sequencing requirements. Additionally, if sample processing cannot be outsourced, access to equipment, facilities and the expertise to perform DNA extractions, library preparation and genetic sequencing is necessary. The choice of extraction method requires consideration as DNA quality can differ depending on taxa (Panova et al., 2016) and high-molecular-weight DNA is required for long-read sequencing methods (Trigodet et al., 2022). We have detailed the extraction and sequencing protocols used for this project (Appendix S8), but trials should be performed to optimize collection, extraction and sequencing strategies for deriving a desired dataset for a particular taxonomic group. The workflow also depends on taxonomic expertise to identify specimens within difficult families, namely those featuring many morphologically similar species or undescribed cryptic species. In addition, access to the infrastructure and archival capacity of a natural history collection is vital because voucher specimens must be catalogued properly and preserved in perpetuity for the science to be verifiable and repeatable (Astrin et al., 2013; Buckner et al., 2021; Prendini et al., 2002). Finally, the use of an appropriate repository—in our case the Oregon State University Ichthyology Collection—that links DNA extracts, tissues, specimens and genomic data permits the re-evaluation of specimen assignment when taxonomies are revised, ensuring that

the collection remains valid and current. For example, a 2023 assessment of the *Rhinichthys osculus* species complex (Moyle et al., 2023) revised the taxonomic designations of our dace specimens during final review of our manuscript. This highlights the point that a collection of this nature is not static and needs maintenance. It requires collaborators with a robust understanding of local species along with the expertise to perform morphological evaluations and update classifications as taxonomies are revised. In addition, any taxonomic or other changes to specimen metadata need to be reflected in public repositories.

Researchers seeking to construct reference libraries for non-piscine taxa should first understand the structure and makeup of the mitochondrial and nuclear genomes of those groups prior to curating mitogenomic sequences. They should ensure that the proposed wet and dry laboratory pipelines can successfully resolve mitogenomes in those groups. Fish mitogenomes contain fewer repeats, insertions and deletions than those of other vertebrates (Formenti et al., 2021), which can cause problems in the sequencing and assembly pipeline (Tørresen et al., 2019). The structural simplicity of many fish mitochondrial genomes enables mitogenome recovery while sequencing at shallow depths. Given 2×150 bp sequencing (L), 65 specimens (S) and ~300 million reads (N) per lane and a hypothetical 2200Mb of DNA (G) per specimen, coverage will be <1× for the nuclear genome (LN/GS; Sims et al., 2014) but many multiples of this for the mitogenome as there are many mitochondria per cell. Between 30× and 50× coverage is recommended to reconstruct mitogenomes de novo when using short-read sequencers, so increased sequencing depth may be needed for taxa with bulky, complex genomic material, such as certain salamander species. Alternatively, a combination of long-read and short-read sequencing may readily resolve complete mitogenomes in these circumstances (Formenti et al., 2021).

An essential concern for a sequence database of this type is to properly represent each taxon's genetic variability. Polymorphisms exist within species, so having multiple examples is necessary to assess intraspecific variability. After collecting multiple examples of a species, we compared intra- and interspecific identity to determine

Centrarchidae			Cottidae			Ictaluridae			Petromyzontidae		
Max	Mean	id_prop	Max	Mean	id_prop	Max	Mean	id_prop	Max	Mean	id_prop
24	14.366	0.909	18	8.073	0.983	24	15.268	0.948	7	4.026	0.995
38	17.718	0.904	23	8.821	0.976	30	15.184	0.944	8	3.708	0.997
42	30.786	0.827	26	18.619	0.938	48	32.976	0.868	24	13.902	0.988
51	36.111	0.795	35	24.267	0.927	44	33.756	0.861	30	18.711	0.984
38	26.113	0.853	24	17.479	0.957	41	29.479	0.883	19	10.803	0.991
36	21.750	0.869	25	16.679	0.964	27	17.429	0.911	20	14.370	0.987
35	25.844	0.855	24	14.938	0.956	39	27.938	0.896	18	12.656	0.989
50	31.258	0.809	28	17.548	0.938	44	34.355	0.860	23	15.113	0.987
59	28.482	0.812	35	18.435	0.94	51	34.024	0.862	23	15.301	0.985
51	31.680	0.826	22	16.900	0.952	41	31.900	0.873	18	12.520	0.989
107	30.950	0.834	103	34.655	0.919	46	22.040	0.885	46	15.588	0.982

whether further collections were necessary. High interspecies identity within Catostomidae and Cottidae suggested the need for more examples per species, but the ratio of intra- to interspecific identity was otherwise low (Figure 7), suggesting that, for most of Oregon's fishes, only a few examples per species were needed. Unfortunately, we were unable to collect multiple individuals for all taxa and were therefore unable to assess intraspecific variation for all species. To properly capture intraspecific variation, one could collect multiple examples of each taxon across its distribution, repeatedly measure the diminution of new polymorphisms and use rarefaction curves to strategize collection. Nevertheless, databases containing full mitogenomic data that do not completely capture intraspecific variation or include all target taxa are still useful because queries with multiple nucleotide mismatches spread across the mitogenome suggest a species has been detected that is not represented in the database.

4.2 | Taxonomically informative genes for fishes

Our reference sequence database provides a valuable genetic resource for analysing mitochondrial genetic variability among Oregon's freshwater fishes and gauging capacity to identify species in eDNA assays. Our analyses illustrate that mitochondrial sequences at every level, from individual genes to the entire mitogenome, are sufficiently conserved within species to provide reliable identifications. However, not all mtDNA regions are equally good at distinguishing between closely related species and certain complete genes and concatenated gene regions (such as the NAD regions) are generally superior (Figure 10). Sequences must have diverged sufficiently among taxa to discern them, so longer regions and faster evolving sections of the mitogenome are preferable for diagnosing recently separated lineages.

Sections of the COI gene have been recommended for barcoding animals (Hebert, Cywinska, et al., 2003) and metabarcoding eukaryotes (Meusnier et al., 2008) or metazoans (Andújar et al., 2018), but they may not be optimal for identifying certain taxa, especially when only short stretches of DNA are retrievable. For certain taxonomic groups, COI sequences do not consistently cluster or associate sequences with their assigned species when barcoding (Waugh, 2007) or metabarcoding (Collins et al., 2019). Previous analyses by Hebert, Cywinska, et al. (2003) and Hebert, Ratnasingham, and de Waard (2003) examined the use of COI for species identification in Lepidoptera by quantifying sequence divergence, using NJ analyses and multidimensional scaling to assign sequences to species. They extrapolated the general suitability of their COI barcode to diagnose animal species from their success in these taxonomically narrow trials. Though their arguments in favour of the COI as the core of a global bioidentification system for animals were logical, they were also speculative (Hebert, Ratnasingham, & de Waard, 2003). They did not assess the comparative merits of the COI over other mitochondrial genes and explicitly stated the need to validate the diagnosability of the COI gene

for different taxonomic groups (Hebert, Cywinska, et al., 2003). This has been done for the COI barcode for a variety of taxonomic groups over the intervening decades (invertebrates: (Cywinska et al., 2006; Sheffield et al., 2009; M. R. Young et al., 2019); fish: (Zemlak et al., 2009); birds: (Hebert et al., 2004; Kerr et al., 2009); amphibians: (Smith et al., 2008); mammals: (Francis et al., 2010)) with results based on sequence divergence and NJ clustering analyses suggesting that, for arthropods and vertebrates, this barcode is useful for parsing these groups taxonomically. It is unclear, however, to what degree the COI is taxonomically diagnostic to the species level for all metazoans. Relatively low percent identity between multiple species does not necessarily equate to species-level diagnosticity (Figure S9 in Appendix S4). A more complete evaluation of the comparative diagnosability of different parts of the mitogenome is therefore needed for a broad range of taxa. Here, we demonstrate that for Oregon's freshwater fishes, regions other than the COI have the potential to better identify species in certain taxa.

A variety of factors contribute to how successfully a region parses the independently evolving lineages we typically call species. Areas with relatively high interspecies genetic distance and concentrations of taxonomically diagnostic nucleotides within families are likely candidates for diagnosing to the species level, whether barcoding, metabarcoding or performing single-species detection. To 'capture' these regions for single-species qPCR assays, the goal is to detect eDNA from the target species and no other species, so areas with high intrafamily distance, high intra-species identity and high mean concentrations of TDNs are likely the best candidates. Unlike single-species qPCR assays, metabarcoding primers need to capture eDNA from a broad range of taxa—different families, orders, classes or even phyla—so there need to be shared regions (typically between 18 and 27 bases long) that can permit primer binding and avoid species dropout. Essentially, a 'Goldilocks' zone is needed for metabarcoding: a region with sufficient genetic divergence to distinguish between species, but not so divergent that conserved flanking regions are unavailable for primer binding. For this reason, the hairpin-loop structure of rRNA regions makes them appropriate for metabarcoding and explains why the most referenced fish metabarcoding primers are found in the 12S region (Miya et al., 2015) despite this region's high within-family interspecies percent identity relative to other regions. To guide the discovery of these areas in future studies, our analysis suggests that these regions contain TDN 'spikes' in otherwise conserved stretches of the mitogenome.

For Oregon's freshwater fishes, we found multiple gene regions and the D-loop had high interspecies genetic distance and concentrations of TDNs within families, suggesting there are numerous alternatives to the COI gene for species identification using eDNA. For single-species analyses, the results were mixed depending on the family examined. Salmonidae, Cyprinidae, Cottidae and Petromyzontidae had the highest mean TDN/ w_{150} in the D-loop (Table 2; Figure 5), although this region also had

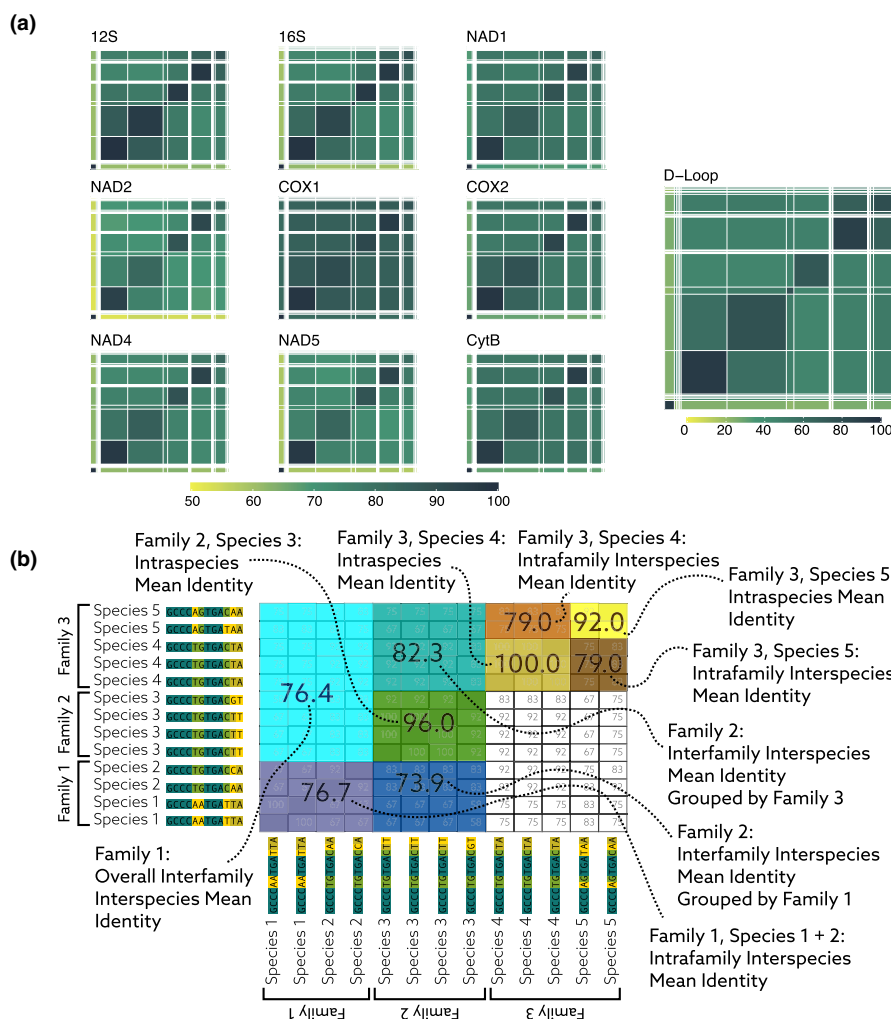


FIGURE 6 (a) Percent identity heatmaps: Family level. This is a graphical representation of interfamily mitogenome percent identity. Yellow colours indicate greater dissimilarity, while 100% identity is represented in dark purple; higher contrast therefore indicates greater distance in identity between families. Y- and X-axes are identical with each block on an axis representing one family in the following order from top to bottom and right to left: Poeciliidae, Fundulidae, Pleuronectidae, Oxudercidae, Embiotocidae, Centrarchidae, Percidae, Pholidae, Gasterosteidae, Sebastidae, Cottidae, Percopsidae, Osmeridae, Umbridae, Esocidae, Salmonidae, Ictaluridae, Cyprinidae, Catostomidae, Cobitidae, Clupeidae, Atherinopsidae, Acipenseridae and Petromyzontidae. Raw numbers can be referenced in Appendix S5. A species-level heatmap is available in Figure S9 in Appendix S4. (b) Schematic View of an Identity Matrix: An alignment of sequences is compared in a pairwise fashion to determine the distance between one sequence and all other sequences in the alignment. The resulting matrix is symmetric along the diagonal with the central diagonal—where a sequence is compared with itself—remaining blank. Data can be grouped, and mean percent identities can be calculated as depicted here.

low relative intraspecies identity compared with other mitochondrial regions (Figure 8). Regions with a combination of all three important factors—high intraspecies identity, high intrafamily distance and high mean TDN/w₁₅₀i₂₀—were the NAD2 gene for Salmonidae, Centrarchidae, Cottidae and Petromyzontidae, the NAD5 gene for Cyprinidae, and the NAD4 gene for Catostomidae and Ictaluridae (Table 2; Figure 7; Table S2 in Appendix S4). The 12S and 16S genes contain TDN ‘spikes’ (Table 2) within loop structures, where mutations are introduced among otherwise conserved complementary hairpin regions. These clusters of TDNs within conserved hairpin areas illustrate why these rRNA regions are better candidates for metabarcoding primers than the COI. Although the 16S rRNA region had these ‘spikes’ across more

families than other genes (Table 2), our analysis of diagnostic nucleotides suggests that for metabarcoding, no single gene region is clearly best for all families of resident freshwater fish in Oregon, and primer binding requirements further restrict which areas can be used.

To ultimately be of use for metabarcoding, a region needs to diagnose species effectively and consistently and be sequenceable using contemporary technologies. Our comparative BLAST analysis provides insight into how consistently various regions of the mitogenome diagnose species or subspecies (Figure 10). The results suggest that, when performing BLASTn sequence queries—the standard query used for eDNA metabarcoding results—multiple genes or the full mitogenome are generally more successful at

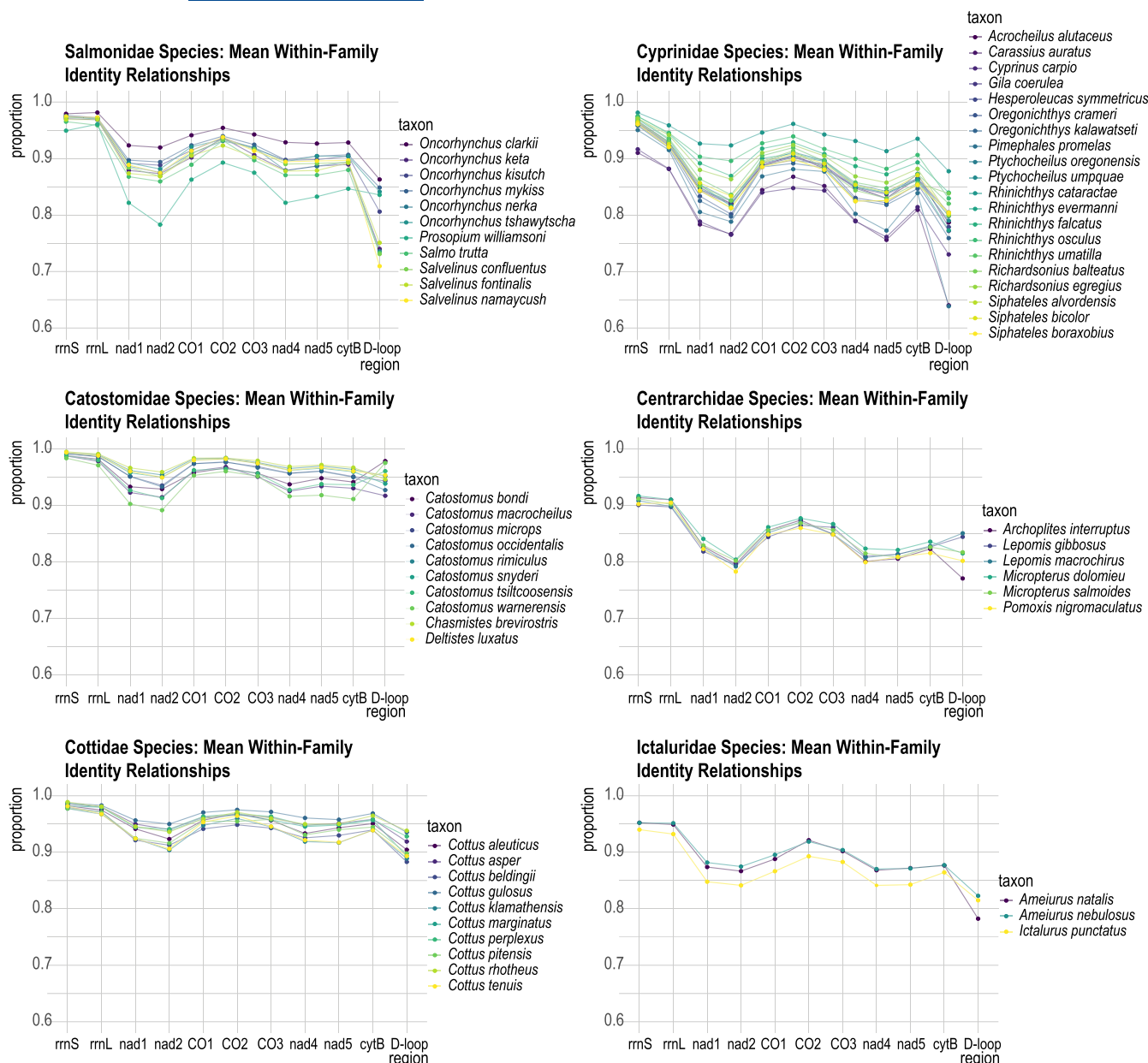


FIGURE 7 Within-family relationships in mean percent identity: Parallel coordinate plots. These plots illustrate the relationship between intraspecies and interspecies genetic distance within the six plotted families at 11 regions (10 gene regions and the D-loop). At each of these 11 mitogenomic regions, mean intraspecies percent identity is calculated and then mean percent identity is calculated between that species and all other species within a given family. For each species at each region, the proportional relationship between these two means is plotted. A higher proportional value indicates higher genetic similarity between species within a family, so identifying species within these families may be more difficult using eDNA assays. Species within families exhibiting greater genetic distance here—Cyprinidae and Centrarchidae—should therefore be easier to identify using shorter regions of the mitogenome than species within families with higher genetic similarity (e.g. Cottidae and Catostomidae). See Table S1 in Appendix S4 for proportional identity figures.

identification than most individual genes and mini-barcodes. The exception is the NAD5 gene, which produced the most target-specific results along with queries using the full mitogenome and concatenated NAD4/NAD5 or 12S/16S regions. This suggests that the NAD5 gene and certain concatenated genes within the mitogenome may be useful for single-species assays and with future PCR-free approaches. However, the low interspecies percent identity of the NAD genes (Figure 8) may prevent location

of nondegenerate multispecies metabarcoding primers that capture regions short enough to be sequenced using Illumina technologies. In addition, entire single and concatenated genes are too long to be sequenced using next-generation methods. For example, the ~650bp COI region published by Hebert, Ratnasingham, and de Waard (2003) exceeds the length feasible for Illumina high-throughput sequencing (Meusnier et al., 2008), so subsections need to be used for metabarcoding. However, capturing

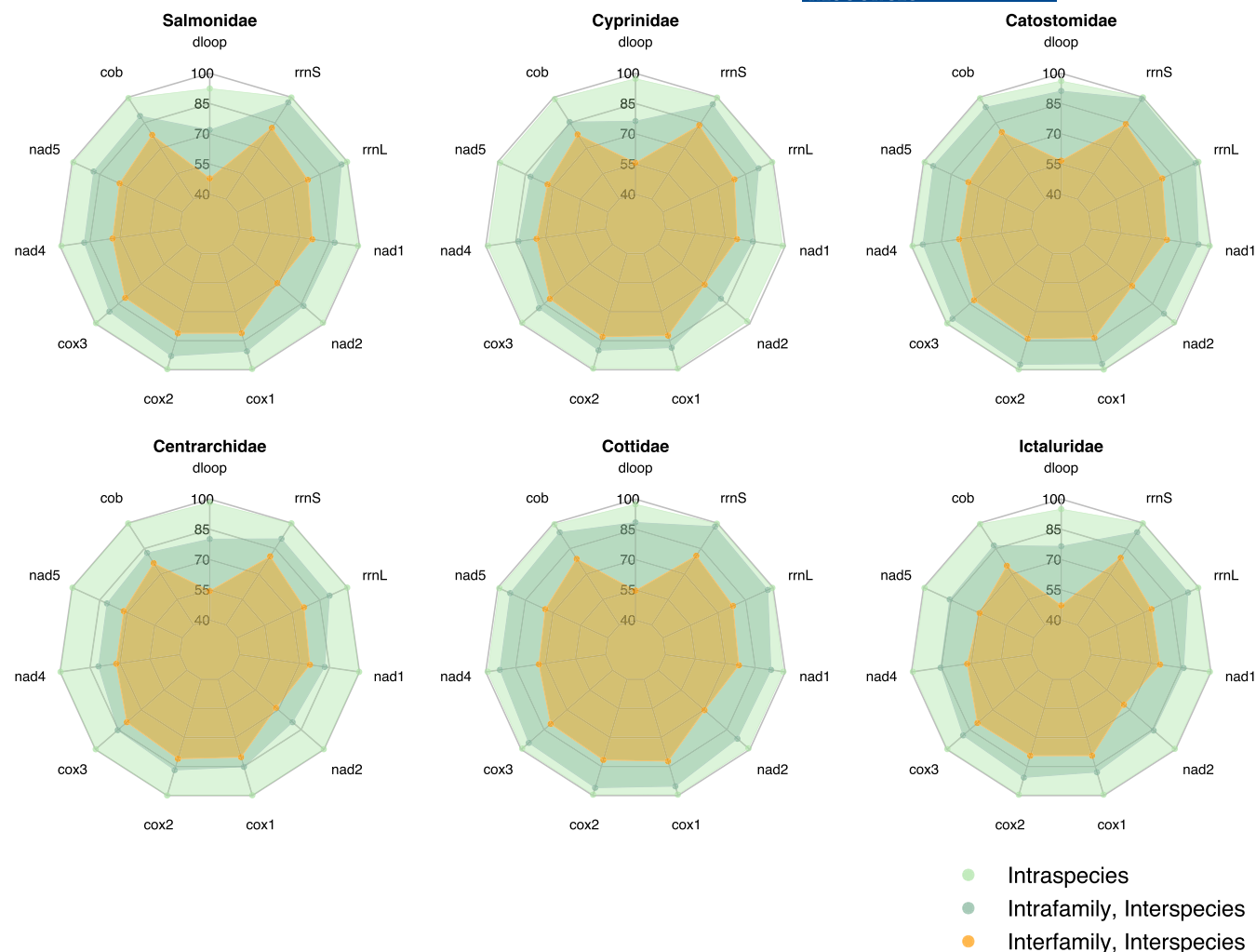


FIGURE 8 Percent identity in a subset of families at regions within the mitogenome. Within each family, mean intraspecies percent identities are calculated and plotted for each gene and the D-loop (pale green), along with interspecies intrafamily percent identities (dark green), and interspecies interfamily percent identities (orange). Interfamily calculations are computed between species from all families, not just the families depicted here. Values on radar chart axes span from 40% identity at the innermost ring to 100% identity at the outermost ring. Genes are arranged in the order in which they occur in the circular mitogenome. Petromyzontidae is not included in plots due to highly skewed interspecies/interfamily identity. To identify species in eDNA assays, intraspecies identity should ideally be high and interspecies identity should be low. All families exhibit high intraspecies identity across all regions of the entire mitogenome with the lowest intraspecies identity found in the D-loop. Between families, there is sufficient interspecies variation to distinguish species from different families and identify sequences to the genus level. Within Cottidae and Catostomidae, there is relatively high interspecies identity, suggesting that species within these families may be difficult to identify. Within Salmonidae, there is high interspecies identity at the 12S (rrnS) and 16S (rrnL) gene regions suggesting these regions may not be ideal for identifying salmonid species.

sufficiently short regions of the COI for Illumina sequencing often requires highly degenerate primers (Collins et al., 2019; Deagle et al., 2014). Because of their wide range of optimal primer melting temperatures, degenerate primers may differentially amplify target species, of particular concern when targeting extraorganismal eDNA (Hajibabaei et al., 2019). The 16S gene may present a potential alternative to 12S mini-barcode regions for eDNA metabarcoding of fishes because it identifies targets to species or subspecies relatively well (Figure 10) and has high interspecies percent identity (Figure 8) and TDN 'spikes' (Table 2). Additional studies evaluating subsections of the 16S gene for this purpose would be informative.

In cases where no single barcode can separate all targeted species, several viable options exist. For single-species and metabarcoding assays that rely on short barcode regions, it may be necessary to use multiple regions for metabarcoding or perhaps a diagnostic region in the nuclear genome, such as the ITS1 gene, to discern closely related congeners (Dysthe et al., 2018). Having more extensive mitogenomic data is necessary for multilocus metabarcoding, a strategy that can increase specificity and sensitivity in single-species eDNA assays (Brys et al., 2023). In addition, full mitogenomes contain significant within-family interspecies variability that can be harnessed with PCR-free approaches to improve species identification over single genes, even among

Full Mitogenome Identity

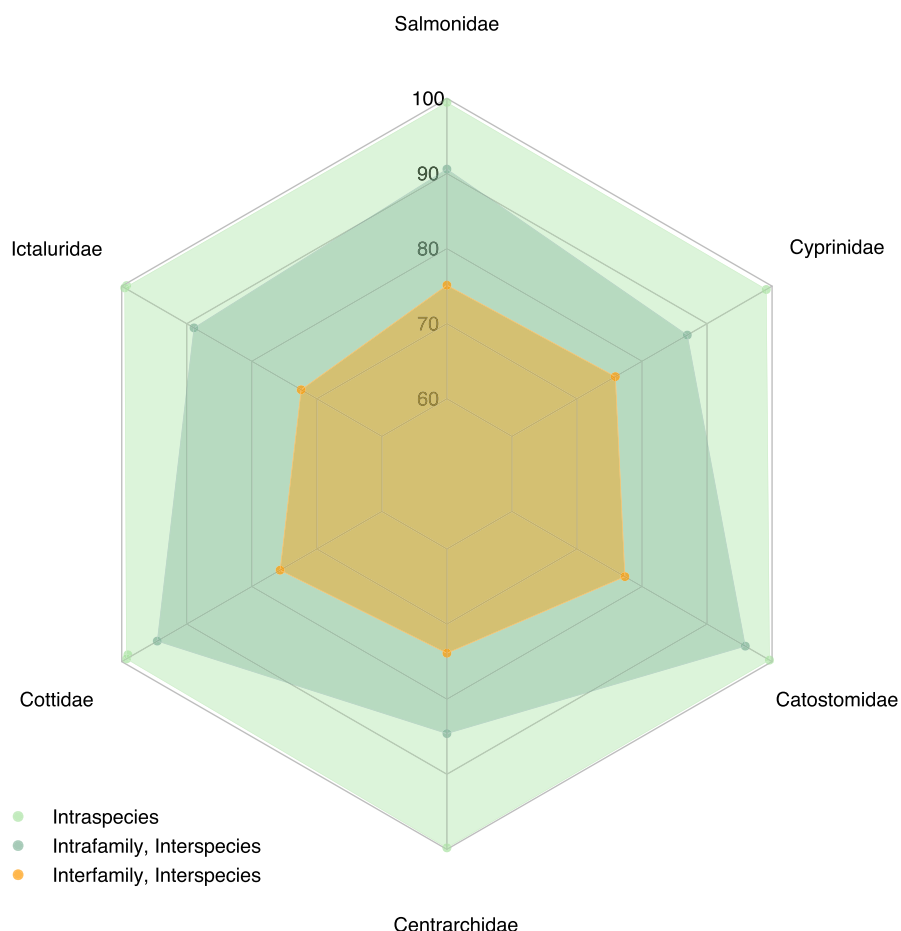


FIGURE 9 Whole mitogenome percent identity in a subset of families. Within each family, mean intraspecies percent identities are calculated and plotted for the whole mitogenome (pale green), along with interspecies intrafamily percent identities (dark green), and interspecies interfamily percent identities (orange). Interfamily calculations are computed between species from all families, not just the families depicted here. Values on radar chart axes span from 60% identity at the innermost ring to 100% identity at the outermost ring. Petromyzontidae is not included in this plot due to highly skewed interspecies/interfamily identity. To identify species in mitochondrial metagenomic assays, intraspecies identity should ideally be high and interspecies identity should be low. All families exhibit high intraspecies identity across the entire mitogenome. All families exhibit very low intrafamily interspecies identity and relatively low within-family interspecies identity suggesting the entire mitogenome should successfully identify species from all families.

relatively conserved taxa. The benefits of using full mitogenomes or a combination of strategically valuable mitochondrial genes derived from full mitogenomes to discern Oregon's resident fish species are illustrated by our analysis of *in silico* BLASTn queries (Figure 10). An advantage of genome skimming from shotgun sequencing (also 'whole-genome shotgun sequencing' or WGS)—where all genetic material from a specimen is sequenced at shallow depth (Trevisan et al., 2019), as undertaken in this project (Appendix S8)—is that the resulting raw sequence data includes high-copy-number nuclear markers such as 18S rRNA at sufficient coverage for potential assembly and use to discern species should mitochondrial markers fail. Indeed, genome skimming produces a plethora of useful data that can be utilized for specimen identification in a variety of applications (Bohmann et al., 2020).

4.3 | Biological challenges in distinguishing species

It is important to note that the failure to correctly identify the presence of a species using eDNA may be the biological reality rather than a fault of the method. Species that often confound taxonomists

and are difficult to distinguish morphologically also appear to be difficult to resolve genetically; not even full, reliable, mitogenomic sequences may be able to resolve such cases. For example, hybridization and organelle introgression from secondary contact can obscure the relationships of different species in an environment (Dowling et al., 2016; Forsythe et al., 2020). Because it is inherited matrilineally, mitogenomic information on its own cannot distinguish hybrid species, and nuclear genetic information may be needed to untangle the genetic complexities of introgression resulting from secondary contact and hybridization—the likely culprit behind difficulties identifying certain catostomid species (Dowling et al., 2016). Difficulties with cottid identification may result from insufficiently diverged lineages or widely variable morphology (Rowsey & Egge, 2017), and the tendency of lamprey taxa to rapidly derive nonparasitic 'satellite species' from parasitic species in sympatry (Salewski, 2003; Vladikov & Kott, 1979) often leaves insufficient time for genetic divergence to accrue (April et al., 2011; Brownstein & Near, 2023). Understanding these shortcomings is essential when undertaking biodiversity assessments using eDNA, and even though the method may not be perfect, the ability to assess a broad variety of taxa using noninvasively collected environmental samples is transformative.

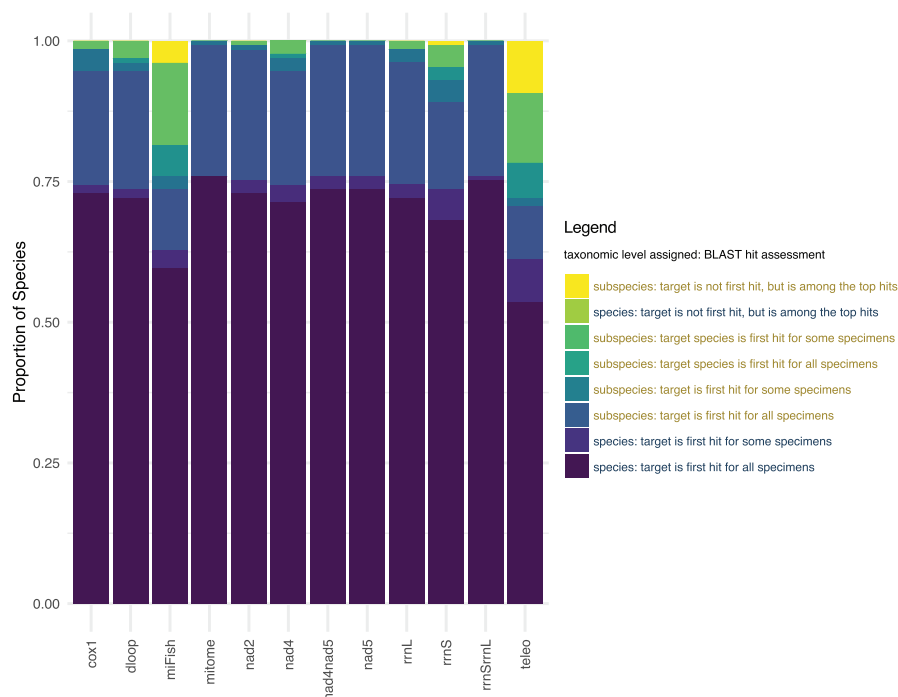


FIGURE 10 Mitogenome regions and PCR amplicons queried against local OBP database: Analysis of All Highest E-Value BLAST Hits: Regions of the mitogenome were extracted from the alignment of 313 OBP sequences, and all sequence regions were blasted against a local database created from the same alignment. Single genes queried were COI, NAD2, NAD4, NAD5, 16S (rrnL) and 12S (rrnS) regions. The entire region spanning from the beginning of the NAD4 to the end of the NAD5 (nad4nad5) genes and from the beginning of the 12S to the end of the 16S (rrnSrrnL) genes was also queried along with the entire mitogenome (mitome) as well as the amplicons produced by both miFish and Teleo primer pairs. All hits with the highest E-Value measured in BLAST were analysed to determine the proportion of species having the target species or subspecies as the first hit. Species or subspecies with multiple specimens represented in the alignment were evaluated to see whether the target species or subspecies was the first hit for some or all representatives of that species. For the full mitogenome, the target species was the first hit for all specimens assigned to the species level and the target subspecies was the first hit for most subspecies that had specimens assigned to the subspecies level. For concatenated regions and the NAD5 gene, the results were similar to the full mitogenome, although the correct species was not always the first hit for all representatives of a species. The remaining regions were less likely to have the target as the first hit with miFish and Teleo amplicons having fewer first target hits than other, longer regions. This illustrates that a reliance on single short amplicons produced with PCR limits the capacity to identify species in an eDNA sample.

4.4 | Mitogenomes in a nuclear future

Despite the challenges involved with a project of this scale and the limitations of mitogenomes to genetically resolve hybrids and certain closely related species, full mitogenomic data provide both a useful genetic reference for species identification and the genetic information needed to develop primers for single-species and metabarcoding assays. It also furnishes researchers with the data needed to move away from microgenomics such as barcoding or metabarcoding and into capture enrichment (Wilcox et al., 2018) or PCR-free environmental genomics, which solve the problems associated with PCR amplification biases (Piñol et al., 2015). Such methods make accurate quantification of relative species abundance in a sample a real possibility (Yang et al., 2021). Full mitogenomic data also frequently diagnose individuals to species more reliably than do shorter sequences (Figure 10). Compiling regional mitogenome databases expands the global repository of available genetic data and builds local capacity to query metabarcoding reads and develop or select the qPCR and metabarcoding primers that best support eDNA-based monitoring of biodiversity.

In anticipation of future applications, organizations are assembling whole nuclear genomes. Examples include the 'Bat1K' (Teeling et al., 2018) and '1000 Fungal Genomes' (Grigoriev et al., 2014) projects, sequencing 1000 species of bats and fungi, respectively, the 'i5K' consortium sequencing 5000 arthropod genomes (i5K Consortium, 2013), the '10KP' and 'P10K' projects sequencing 10,000 plant (Cheng et al., 2018) and protist (Miao et al., 2020) species and the 'GIGA' project (GIGA, 2014) dedicated to sequencing invertebrate genomes. Even more ambitious projects are efforts to sequence genomes from representatives of every vertebrate species (Vertebrate Genomes Project (VGP); Rhie et al., 2021), and all of Earth's eukaryotic biodiversity by 2027 (Earth Biogenome Project; Lewin et al., 2018). These large-scale, expensive and top-down efforts will create nuclear genome reference databases for many species in the coming decades, but their global focus is unlikely to provide comprehensive genetic information for any specific geographic region or taxonomic group of interest. For example, as of the time of this writing the VGP has produced 110 nuclear assemblies and published data for 125 mitogenomes (Formenti et al., 2021). This is an invaluable contribution but lacks redundant sequence information. An exception to these

large-scale projects is the California Conservation Genomics Project (Shaffer et al., 2022), a regional effort to create nuclear reference genomes and multiple resequenced nuclear genomes for numerous species within the state of California. This presents an exemplar for researchers to pursue similar efforts, and if the affordability and accuracy promised by PacBio for its new generation of long-read sequencers becomes a reality, projects of this scale will effectively become democratized. While we should strive to create reference databases of full nuclear genomes for all organisms on the planet, nuclear DNA may not turn up reliably in extraorganismal eDNA from aggregating, nonspawning metazoans (Jensen et al., 2021; Olson et al., 2012), making nuclear reference sequence databases less useful for eDNA applications. There will therefore always be a need for mitogenomic reference data, and much can still be gleaned from complete mitogenomes on their own (Blair, 2023).

5 | CONCLUSION

We hope these protocols and insights into mitogenomic variability will encourage researchers around the globe to follow suit and develop their own regional databases and archives of voucher specimens. Widely ranging mitogenomic databases would expand eDNA monitoring potential to more regions. Repositories of vouchered samples and full mitogenomic information as described herein not only provide the genetic information needed to use eDNA effectively for biodiversity studies (de Santana et al., 2021) but also can support investigations of taxonomy, population structure, landscape genetics and multilocus metabarcoding. Coupled with high-molecular-weight DNA extraction for nuclear genome sequencing, a project of this scope grows the global repository of sequence information in anticipation of an environmental genomics future and lays the groundwork to compile all the available genetic information for freshwater fishes or other taxa in a region of interest.

AUTHOR CONTRIBUTIONS

Emily Dziedzic, Richard Cronn, Brooke E. Penaluna, Taal Levi and Peter Konstantinidis conceived the project. James Anthony, Trevan Cornwell, Thomas A. Friesen, Brian Sidlauskas, Brooke E. Penaluna and Staci Stein strategized and carried out specimen collection. Emily Dziedzic analysed the data. Emily Dziedzic, Taal Levi, Brian Sidlauskas, Richard Cronn, Brooke E. Penaluna, James Anthony, Peter Konstantinidis and Thomas A. Friesen wrote the paper.

ACKNOWLEDGEMENTS

We thank the Oregon Department of Fish and Wildlife, United States Forest Service Pacific Northwest Research Station and Oregon State University for financial and logistical support. We are also grateful to the Klamath Tribes, the United States Fish and Wildlife Service and the California Department of Fish and Wildlife for providing valuable contributions towards our collection goals. In addition, we greatly appreciate specimen donations from Mark Buettner, Nolan Banish, Bruce Hansen, Paul Divine and Dave Hering. Laura Hauck and Lucas

Longway provided invaluable assistance with laboratory training and Shawn Clements helped to get this project off the ground. Many thanks to the University of Washington, the University of Alaska Museum of the North and the Oregon State University Ichthyology Collection for providing us with vouchered tissue samples. We also thank Doug Markle and Álvaro Cortés for providing taxonomic identification assistance. In addition, Tom Stahl, Marc Johnson and members of the Levi Lab reviewed the manuscript and provided valuable feedback for which we are also grateful.

DATA AVAILABILITY STATEMENT

Appendix S3 contains all mitogenomic sequences and details regarding the methodological processes used to derive the data for downstream analyses in this paper; Appendix S7 is a zipped file containing all mitogenome sequences assembled for this paper in FASTA format. These files along with all additional data and supplemental materials are available at <https://doi.org/10.5061/dryad.2jm63xsv4> via the Dryad data repository.

ORCID

Emily Dziedzic  <https://orcid.org/0000-0002-5889-4110>

Brian Sidlauskas  <https://orcid.org/0000-0003-0597-4085>

Taal Levi  <https://orcid.org/0000-0003-1853-8311>

REFERENCES

- Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjær, K. H., Orlando, L., Gilbert, M. T. P., & Willerslev, E. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, 21(8), 1966–1979. <https://doi.org/10.1111/j.1365-294X.2011.05261.x>
- Andújar, C., Arribas, P., Ruzicka, F., Crampton-Platt, A., Timmermans, M. J. T. N., & Vogler, A. P. (2015). Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Molecular Ecology*, 24(14), 3603–3617. <https://doi.org/10.1111/mec.13195>
- Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, 27(20), 3968–3975. <https://doi.org/10.1111/mec.14844>
- April, J., Mayden, R. L., Hanner, R. H., & Bernatchez, L. (2011). Genetic calibration of species diversity among North America's freshwater fishes. *Proceedings of the National Academy of Sciences*, 108(26), 10602–10607. <https://doi.org/10.1073/pnas.1016437108>
- Arulandhu, A. J., Staats, M., Hagelaar, R., Voorhuijzen, M. M., Prins, T. W., Scholtens, I., Costessi, A., Duijsings, D., Rechenmann, F., Gaspar, F. B., Barreto Crespo, M. T., Holst-Jensen, A., Birck, M., Burns, M., Haynes, E., Hohegger, R., Klingl, A., Lundberg, L., Natale, C., ... Kok, E. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience*, 6(10), 1–18. <https://doi.org/10.1093/gigascience/gix080>
- Astrin, J., Zhou, X., & Misof, B. (2013). The importance of biobanking in molecular taxonomy, with proposed definitions for vouchers in a molecular context. *ZooKeys*, 365, 67–70. <https://doi.org/10.3897/zookeys.365.5875>
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Pribelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of*

- Computational Biology*, 19(5), 455–477. <https://doi.org/10.1089/cmb.2012.0021>
- Barter, R. L., & Yu, B. (2018). Superheat: An R package for creating beautiful and extendable heatmaps for visualizing complex data. *Journal of Computational and Graphical Statistics*, 27(4), 910–922. <https://doi.org/10.1080/10618600.2018.1473780>
- Béjà, O. (2004). To BAC or not to BAC: Marine ecogenomics. *Current Opinion in Biotechnology*, 15(3), 187–190. <https://doi.org/10.1016/j.copbio.2004.03.005>
- Benson, D. (1996). GenBank. *Nucleic Acids Research*, 24(1), 1–5. <https://doi.org/10.1093/nar/24.1.1>
- Blair, C. (2023). Organellar DNA continues to provide a rich source of information in the genomics era. *Molecular Ecology*, 32(9), 2144–2150. <https://doi.org/10.1111/mec.16872>
- Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, 29(6), 358–367. <https://doi.org/10.1016/j.tree.2014.04.003>
- Bohmann, K., Mirarab, S., Bafna, V., & Gilbert, M. T. P. (2020). Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology*, 29(14), 2521–2534. <https://doi.org/10.1111/mec.15507>
- Bragg, L., & Tyson, G. W. (2014). Metagenomics using next-generation sequencing. In I. T. Paulsen & A. J. Holmes (Eds.), *Environmental microbiology* (Vol. 1096, pp. 183–201). Humana Press. https://doi.org/10.1007/978-1-62703-712-9_15
- Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., de Waard, J. R., Sones, J. E., Zakharov, E. V., & Hebert, P. D. N. (2019). Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources*, 19(3), 711–727. <https://doi.org/10.1111/1755-0998.13008>
- Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M., Malumbres-Olarte, J., Vink, C. J., & Cruickshank, R. H. (2012). Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, 12(3), 562–565. <https://doi.org/10.1111/j.1755-0998.2011.03108.x>
- Brownstein, C. D., & Near, T. J. (2023). Phylogenetics and the Cenozoic radiation of lampreys. *Current Biology*, 33(2), 397–404. <https://doi.org/10.1016/j.cub.2022.12.018>
- Bru, D., Martin-Laurent, F., & Philippot, L. (2008). Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Applied and Environmental Microbiology*, 74(5), 1660–1663. <https://doi.org/10.1128/AEM.02403-07>
- Brys, R., Halfmaerten, D., Everts, T., Van Driessche, C., & Neyrinck, S. (2023). Combining multiple markers significantly increases the sensitivity and precision of eDNA-based single-species analyses. *Environmental DNA*. <https://doi.org/10.1002/edn3.420>
- Buckner, J. C., Sanders, R. C., Faircloth, B. C., & Chakrabarty, P. (2021). The critical importance of vouchers in genomics. *eLife*, 10, e68264. <https://doi.org/10.7554/eLife.68264>
- Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P. M., Li, F. W., Melkonian, B., Mavrodiev, E. V., Sun, W., Fu, Y., Yang, H., Soltis, D. E., Graham, S. W., Soltis, P. S., Liu, X., Xu, X., & Wong, G. K. S. (2018). 10KP: A phylodiverse genome sequencing plan. *GigaScience*, 7(3), 1–9. <https://doi.org/10.1093/gigascience/giy013>
- Clare, E. L., Economou, C. K., Bennett, F. J., Dyer, C. E., Adams, K., McRobie, B., Drinkwater, R., & Littlefair, J. E. (2022). Measuring biodiversity from DNA in the air. *Current Biology*, 32(3), 693–700. <https://doi.org/10.1016/j.cub.2021.11.064>
- Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, 44(D1), D67–D72. <https://doi.org/10.1093/nar/gkv1276>
- Collins, R. A., Armstrong, K. F., Holyoake, A. J., & Keeling, S. (2013). Something in the water: Biosecurity monitoring of ornamental fish imports using environmental DNA. *Biological Invasions*, 15(6), 1209–1215. <https://doi.org/10.1007/s10530-012-0376-9>
- Collins, R. A., Bakker, J., Wangenstein, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, 10(11), 1985–2001. <https://doi.org/10.1111/2041-210X.13276>
- Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, 30(13), 2937–2958. <https://doi.org/10.1111/mec.15472>
- Crampton-Platt, A., Timmermans, M. J. T. N., Gimmel, M. L., Kutty, S. N., Cockerill, T. D., Vun Khen, C., & Vogler, A. P. (2015). Soup to tree: The phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution*, 32(9), 2302–2316. <https://doi.org/10.1093/molbev/msv111>
- Crampton-Platt, A., Yu, D. W., Zhou, X., & Vogler, A. P. (2016). Mitochondrial metagenomics: Letting the genes out of the bottle. *GigaScience*, 5(1), 15. <https://doi.org/10.1186/s13742-016-0120-y>
- Cristescu, M. E., & Hebert, P. D. N. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. *Annual Review of Ecology, Evolution, and Systematics*, 49(1), 209–230. <https://doi.org/10.1146/annurev-ecolsys-110617-062306>
- Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., Pipes, L., Schweizer, T. M., Rabichow, L., Lin, M., Shi, B., Barber, P. H., Kraft, N., Wayne, R., & Meyer, R. S. (2019). *Anacapa toolkit*: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, 10(9), 1469–1475. <https://doi.org/10.1111/2041-210X.13214>
- Cywinska, A., Hunter, F. F., & Hebert, P. D. N. (2006). Identifying Canadian mosquito species through DNA barcodes. *Medical and Veterinary Entomology*, 20(4), 413–424. <https://doi.org/10.1111/j.1365-2915.2006.00653.x>
- de Santana, C. D., Parenti, L. R., Dillman, C. B., Coddington, J. A., Bastos, D. A., Baldwin, C. C., Zuanon, J., Torrente-Vilara, G., Covain, R., Menezes, N. A., Datovo, A., Sado, T., & Miya, M. (2021). The critical role of natural history museums in advancing eDNA for biodiversity studies: A case study with Amazonian fishes. *Scientific Reports*, 11(1), 18159. <https://doi.org/10.1038/s41598-021-97128-3>
- Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, 10(9), 20140562. <https://doi.org/10.1098/rsbl.2014.0562>
- Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21), 5872–5895. <https://doi.org/10.1111/mec.14350>
- Deiner, K., Fronhofer, E. A., Mächler, E., Walser, J. C., & Altermatt, F. (2016). Environmental DNA reveals that rivers are conveyor belts of biodiversity information. *Nature Communications*, 7, 12544. <https://doi.org/10.1038/ncomms12544>
- Deiner, K., Renshaw, M. A., Li, Y., Olds, B. P., Lodge, D. M., & Pfrender, M. E. (2017). Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods in Ecology and Evolution*, 8(12), 1888–1898. <https://doi.org/10.1111/2041-210X.12836>
- Donath, A., Jühling, F., Al-Arab, M., Bernhart, S. H., Reinhardt, F., Stadler, P. F., Middendorf, M., & Bernt, M. (2019). Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Research*, 47(20), 10543–10552. <https://doi.org/10.1093/nar/gkz833>

- Dowling, T. E., Markle, D. F., Tranah, G. J., Carson, E. W., Wagman, D. W., & May, B. P. (2016). Introgressive hybridization and the evolution of lake-adapted catostomid fishes. *PLoS One*, 11(3), e0149884. <https://doi.org/10.1371/journal.pone.0149884>
- Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C., Heled, J., Ross, H. A., Tooman, L., Grosser, S., Park, D., Demetras, N. J., Stevens, M. I., Russell, J. C., Anderson, S. H., Carter, A., & Nelson, N. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*, 4(1), 46. <https://doi.org/10.1186/s13742-015-0086-1>
- Dysthe, J. C., Franklin, T. W., McKelvey, K. S., Young, M. K., & Schwartz, M. K. (2018). An improved environmental DNA assay for bull trout (*Salvelinus confluentus*) based on the ribosomal internal transcribed spacer 1. *PLoS One*, 13(11), e0206851. <https://doi.org/10.1371/journal.pone.0206851>
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Farrington, H. L., Edwards, C. E., Guan, X., Carr, M. R., Baerwaldt, K., & Lance, R. F. (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(2), e0117803. <https://doi.org/10.1371/journal.pone.0117803>
- Foran, D. R. (2006). Relative degradation of nuclear and mitochondrial DNA: An experimental approach*. *Journal of Forensic Sciences*, 51(4), 766–770. <https://doi.org/10.1111/j.1556-4029.2006.00176.x>
- Formenti, G., Rhie, A., Balacco, J., Haase, B., Mountcastle, J., Fedrigo, O., Brown, S., Capodiferro, M. R., Al-Ajli, F. O., Ambrosini, R., Houde, P., Koren, S., Oliver, K., Smith, M., Skelton, J., Betteridge, E., Dolucan, J., Corton, C., Bista, I., ... Jarvis, E. D. (2021). Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biology*, 22(1), 120. <https://doi.org/10.1186/s13059-021-02336-9>
- Forsythe, E. S., Nelson, A. D. L., & Beilstein, M. A. (2020). Biased gene retention in the face of introgression obscures species relationships. *Genome Biology and Evolution*, 12(9), 1646–1663. <https://doi.org/10.1093/gbe/evaa149>
- Francis, C. M., Borisenko, A. V., Ivanova, N. V., Eger, J. L., Lim, B. K., Guillén-Servent, A., Kruskop, S. V., Mackie, I., & Hebert, P. D. N. (2010). The role of DNA barcodes in understanding and conservation of mammal diversity in Southeast Asia. *PLoS One*, 5(9), e12575. <https://doi.org/10.1371/journal.pone.0012575>
- GIGA. (2014). The global invertebrate genomics Alliance (GIGA): Developing community resources to study diverse invertebrate genomes. *Journal of Heredity*, 105(1), 1–18. <https://doi.org/10.1093/jhered/est084>
- Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., Spear, S. F., McKee, A., Oyler-McCance, S. J., Cornman, R. S., Laramie, M. B., Mahon, A. R., Lance, R. F., Pilliod, D. S., Strickler, K. M., Waits, L. P., Fremier, A. K., Takahara, T., Herder, J. E., & Taberlet, P. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, 7(11), 1299–1307. <https://doi.org/10.1111/2041-210X.12595>
- Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., & Shabalov, I. (2014). MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Research*, 42(D1), D699–D704. <https://doi.org/10.1093/nar/gkt1183>
- Hajibabaei, M., Porter, T. M., Robinson, C. V., Baird, D. J., Shokralla, S., & Wright, M. T. G. (2019). Watered-down biodiversity? A comparison of metabarcoding results from DNA extracted from matched water and bulk tissue biomonitoring samples. *PLoS One*, 14(12), e0225409. <https://doi.org/10.1371/journal.pone.0225409>
- Hartmann, N., Reichwald, K., Wittig, I., Dröse, S., Schmeisser, S., Lück, C., Hahn, C., Graf, M., Gausmann, U., Terzibas, E., Cellerino, A., Ristow, M., Brandt, U., Platzer, M., & Englert, C. (2011). Mitochondrial DNA copy number and function decrease with age in the short-lived fish *Nothobranchius furzeri*: Decline of mitochondrial function in aging fish. *Aging Cell*, 10(5), 824–831. <https://doi.org/10.1111/j.1474-9726.2011.00723.x>
- Hauck, L. L., Weitemier, K. A., Penaluna, B. E., Garcia, T. S., & Cronn, R. (2019). Casting a broader net: Using microfluidic metagenomics to capture aquatic biodiversity data from diverse taxonomic targets. *Environmental DNA*, 1(3), 251–267. <https://doi.org/10.1002/edn3.26>
- Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512), 313–321. <https://doi.org/10.1098/rspb.2002.2218>
- 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. Series B: Biological Sciences*, 270(suppl_1), S96–S99. <https://doi.org/10.1098/rsbl.2003.0025>
- Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology*, 2(10), e312. <https://doi.org/10.1371/journal.pbio.0020312>
- i5K Consortium. (2013). The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity*, 104(5), 595–600. <https://doi.org/10.1093/jhered/est050>
- Jensen, M. R., Sigsgaard, E. E., Liu, S., Manica, A., Bach, S. S., Hansen, M. M., Møller, P. R., & Thomsen, P. F. (2021). Genome-scale target capture of mitochondrial and nuclear environmental DNA from water samples. *Molecular Ecology Resources*, 21(3), 690–702. <https://doi.org/10.1111/1755-0998.13293>
- Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., & Li, D. Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, 21(1), 241. <https://doi.org/10.1186/s13059-020-02154-5>
- Kane, N., Sveinsson, S., Dempewolf, H., Yang, J. Y., Zhang, D., Engels, J. M. M., & Cronk, Q. (2012). Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany*, 99(2), 320–329. <https://doi.org/10.3732/ajb.1100570>
- Kerr, K. C. R., Lijtmaer, D. A., Barreira, A. S., Hebert, P. D. N., & Tubaro, P. L. (2009). Probing evolutionary patterns in neotropical birds through DNA barcodes. *PLoS One*, 4(2), e4379. <https://doi.org/10.1371/journal.pone.0004379>
- Langlois, V. S., Allison, M. J., Bergman, L. C., To, T. A., & Helbing, C. C. (2021). The need for robust qPCR-based eDNA detection assays in environmental monitoring and species inventories. *Environmental DNA*, 3(3), 519–527. <https://doi.org/10.1002/edn3.164>
- Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences*, 116(45), 22651–22656. <https://doi.org/10.1073/pnas.1911714116>
- Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2020). Reply to Locatelli et al.: Evaluating species-level accuracy of GenBank metazoan sequences will require experts' effort in each group. *Proceedings of the National Academy of Sciences*, 117(51), 32213–32214. <https://doi.org/10.1073/pnas.2019903117>
- Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., ... Zhang, G. (2018). Earth BioGenome project: Sequencing life for

- the future of life. *Proceedings of the National Academy of Sciences*, 115(17), 4325–4333. <https://doi.org/10.1073/pnas.1720115115>
- Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: From gene to genome: Plant identification using DNA barcodes. *Biological Reviews*, 90(1), 157–166. <https://doi.org/10.1111/brv.12104>
- Lim, N. K. M., Tay, Y. C., Srivathsan, A., Tan, J. W. T., Kwik, J. T. B., Baloglu, B., Meier, R., & Yeo, D. C. J. (2016). Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal Society Open Science*, 3(11), 160635. <https://doi.org/10.1098/rsos.160635>
- Locatelli, N. S., McIntyre, P. B., Therkildsen, N. O., & Baetscher, D. S. (2020). GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proceedings of the National Academy of Sciences*, 117(51), 32211–32212. <https://doi.org/10.1073/pnas.2007421117>
- Lynggaard, C., Bertelsen, M. F., Jensen, C. V., Johnson, M. S., Frøslev, T. G., Olsen, M. T., & Bohmann, K. (2022). Airborne environmental DNA for terrestrial vertebrate community monitoring. *Current Biology*, 32(3), 701–707. <https://doi.org/10.1016/j.cub.2021.12.014>
- Markle, D. F., & Tomelleri, J. R. (2016). *A guide to freshwater fishes of Oregon*. Oregon State University Press.
- Meiklejohn, K. A., Damaso, N., & Robertson, J. M. (2019). Assessment of BOLD and GenBank—Their accuracy and reliability for the identification of biological materials. *PLoS One*, 14(6), e0217084. <https://doi.org/10.1371/journal.pone.0217084>
- Meusnier, I., Singer, G. A., Landry, J. F., Hickey, D. A., Hebert, P. D., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, 9(1), 214. <https://doi.org/10.1186/1471-2164-9-214>
- Miao, W., Song, L., Ba, S., Zhang, L., Guan, G., Zhang, Z., & Ning, K. (2020). Protist 10,000 genomes project. *The Innovation*, 1(3), 100058. <https://doi.org/10.1016/j.xinn.2020.100058>
- Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., 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. *Royal Society Open Science*, 2(7), 150088. <https://doi.org/10.1098/rsos.150088>
- Moyle, P. B., Buckmaster, N., & Su, Y. (2023). Taxonomy of the speckled dace species complex (Cypriniformes: Leuciscidae, Rhinichthys) in California, USA. *Zootaxa*, 5249(5), 501–539. <https://doi.org/10.11646/zootaxa.5249.5.1>
- Nagy, Z. T. (2010). A hands-on overview of tissue preservation methods for molecular genetic analyses. *Organisms, Diversity and Evolution*, 10(1), 91–105. <https://doi.org/10.1007/s13127-010-0012-4>
- Nakazawa, M. (2021). *fmsb: Functions for medical statistics book with some demographic data*. <https://CRAN.R-project.org/package=fmsb>
- NCBI: National Center for Biotechnology Information. (2021). *About RefSeq*. [Internet]. <https://www.ncbi.nlm.nih.gov/refseq/about/>
- Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., Green, R. E., & Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, 18(5), 927–939. <https://doi.org/10.1111/1755-0998.12895>
- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, 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., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1), D733–D745. <https://doi.org/10.1093/nar/gkv1189>
- Olson, Z. H., Briggler, J. T., & Williams, R. N. (2012). An eDNA approach to detect eastern hellbenders (*Cryptobranchus a. Alleganiensis*) using samples of water. *Wildlife Research*, 39(7), 629. <https://doi.org/10.1071/WR12114>
- Panova, M., Aronsson, H., Cameron, R. A., Dahl, P., Godhe, A., Lind, U., Ortega-Martinez, O., Pereyra, R., Tesson, S. V. M., Wrangé, A.-L., Blomberg, A., & Johannesson, K. (2016). DNA extraction protocols for whole-genome sequencing in marine organisms. In S. J. Bourlat (Ed.), *Marine genomics* (Vol. 1452, pp. 13–44). Springer. https://doi.org/10.1007/978-1-4939-3774-5_2
- Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, 15(4), 819–830. <https://doi.org/10.1111/1755-0998.12355>
- Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., & Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, 25(2), 527–541. <https://doi.org/10.1111/mec.13481>
- Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, 27(2), 313–338. <https://doi.org/10.1111/mec.14478>
- Prendini, L., Hanner, R., & DeSalle, R. (2002). Obtaining, storing and archiving specimens and tissue samples for use in molecular studies. In: R. DeSalle, G. Giribet, & W. Wheeler (Eds.), *Techniques in molecular systematics and evolution*. (pp. 176–248). Birkhäuser. https://doi.org/10.1007/978-3-0348-8125-8_11
- Ratnasingham, S., & Hebert, P. D. N. (2007). Bold: The barcode of life data system. *Molecular Ecology Notes*, 7(3), 355–364. <https://doi.org/10.1111/j.1471-8286.2007.01678.x>
- Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., ... Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856), 737–746. <https://doi.org/10.1038/s41586-021-03451-0>
- Rimet, F., Aylagas, E., Borja, Á., Bouchez, A., Canino, A., Chauvin, C., Chonova, T., Ciampor, F., Jr., Costa, F. O., Ferrari, B. J. D., Gastineau, R., Goulon, C., Gugger, M., Holzmann, M., Jahn, R., Kahlert, M., Kusber, W. H., Laplace-Tretyure, C., Leese, F., ... Ekrem, T. (2021). Metadata standards and practical guidelines for specimen and DNA curation when building barcode reference libraries for aquatic life. *Metabarcoding and Metagenomics*, 5, e58056. <https://doi.org/10.3897/mbmg.5.58056>
- Rowsey, D. M., & Egge, J. J. (2017). Morphometric analysis of two enigmatic sculpin species, *Cottus gulosus* and *Cottus perplexus* (Scorpaeniformes: Cottidae). *Northwestern Naturalist*, 98(3), 190–202. <https://doi.org/10.1898/NWN16-23.1>
- Salewski, V. (2003). Satellite species in lampreys: A worldwide trend for ecological speciation in sympatry? *Journal of Fish Biology*, 63(2), 267–279. <https://doi.org/10.1046/j.1095-8649.2003.00166.x>
- Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2021). GGally: Extension to “ggplot2”. <https://CRAN.R-project.org/package=GGally>
- Schnell, I. B., Fraser, M., Willerslev, E., & Gilbert, M. T. P. (2010). Characterisation of insect and plant origins using DNA extracted from small volumes of bee honey. *Arthropod-Plant Interactions*, 4(2), 107–116. <https://doi.org/10.1007/s11829-010-9089-0>
- Shaffer, H. B., Toffelmier, E., Corbett-Detig, R. B., Escalona, M., Erickson, B., Fiedler, P., Gold, M., Harrigan, R. J., Hodges, S., Luckau, T. K., Miller, C., Oliveira, D. R., Shaffer, K. E., Shapiro, B., Sork, V. L., & Wang, I. J. (2022). Landscape genomics to enable conservation actions: The California conservation genomics project. *Journal of Heredity*, 113(6), 577–588. <https://doi.org/10.1093/jhered/esac020>
- Sheffield, C. S., Hebert, P. D. N., Kevan, P. G., & Packer, L. (2009). DNA barcoding a regional bee (hymenoptera: Apoidea) fauna and its potential for ecological studies. *Molecular Ecology Resources*, 9, 196–207. <https://doi.org/10.1111/j.1755-0998.2009.02645.x>

- Sims, D., Sudbery, I., Illott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2), 121–132. <https://doi.org/10.1038/nrg3642>
- Sipos, R., Székely, A., Révész, S., & Márialigeti, K. (2010). Addressing PCR biases in environmental microbiology studies. In S. P. Cummings (Ed.), *Bioremediation* (Vol. 599, pp. 37–58). Humana Press. https://doi.org/10.1007/978-1-60761-439-5_3
- Smith, M. A., Poyarkov, N. A., & Hebert, P. D. N. (2008). DNA barcoding: CO1 DNA barcoding amphibians: Take the chance, meet the challenge. *Molecular Ecology Resources*, 8(2), 235–246. <https://doi.org/10.1111/j.1471-8286.2007.01964.x>
- Stadhouders, R., Pas, S. D., Anber, J., Voermans, J., Mes, T. H. M., & Schutten, M. (2010). The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5' nuclease assay. *The Journal of Molecular Diagnostics*, 12(1), 109–117. <https://doi.org/10.2353/jmoldx.2010.090035>
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8), 2045–2050. <https://doi.org/10.1111/j.1365-294X.2012.05470.x>
- Tang, M., Hardman, C. J., Ji, Y., Meng, G., Liu, S., Tan, M., Yang, S., Moss, E. D., Wang, J., Yang, C., Bruce, C., Nevard, T., Potts, S. G., Zhou, X., & Yu, D. W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, 6(9), 1034–1043. <https://doi.org/10.1111/2041-210X.12416>
- Teeling, E. C., Vernes, S. C., Dávalos, L. M., Ray, D. A., Gilbert, M. T. P., Myers, E., & Bat1K Consortium. (2018). Bat biology, genomes, and the Bat1K project: To generate chromosome-level genomes for all living bat species. *Annual Review of Animal Biosciences*, 6(1), 23–46. <https://doi.org/10.1146/annurev-animal-022516-022811>
- Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, 47(21), 10994–11006. <https://doi.org/10.1093/nar/gkz841>
- Trevisan, B., Alcantara, D. M. C., Machado, D. J., Marques, F. P. L., & Lahr, D. J. G. (2019). Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. *PeerJ*, 7, e7543. <https://doi.org/10.7717/peerj.7543>
- Trigodet, F., Lolans, K., Fogarty, E., Shaiber, A., Morrison, H. G., Barreiro, L., Jabri, B., & Eren, A. M. (2022). High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Molecular Ecology Resources*, 22(5), 1786–1802. <https://doi.org/10.1111/1755-0998.13588>
- 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(15), e115. <https://doi.org/10.1093/nar/gks596>
- Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J. M., ... Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, 25(4), 929–942. <https://doi.org/10.1111/mec.13428>
- Vladykov, V. D., & Kott, E. (1979). Satellite species among the holarctic lampreys (Petromyzonidae). *Canadian Journal of Zoology*, 57(4), 860–867. <https://doi.org/10.1139/z79-106>
- Waugh, J. (2007). DNA barcoding in animal species: Progress, potential and pitfalls. *BioEssays*, 29(2), 188–197. <https://doi.org/10.1002/bies.20529>
- Wilcox, T. M., Zarn, K. E., Piggott, M. P., Young, M. K., McKelvey, K. S., & Schwartz, M. K. (2018). Capture enrichment of aquatic environmental DNA: A first proof of concept. *Molecular Ecology Resources*, 18(6), 1392–1401. <https://doi.org/10.1111/1755-0998.12928>
- Wu, J. H., Hong, P. Y., & Liu, W. T. (2009). Quantitative effects of position and type of single mismatch on single base primer extension. *Journal of Microbiological Methods*, 77(3), 267–275. <https://doi.org/10.1016/j.mimet.2009.03.001>
- Wydoski, R. S., & Whitney, R. R. (2003). *Inland fishes of Washington* (2nd ed., rev.expanded). American Fisheries Society in association with University of Washington Press.
- Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Minamoto, T., & Miya, M. (2017). Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Scientific Reports*, 7(1), 40368. <https://doi.org/10.1038/srep40368>
- Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., Ding, Z., Gopalakrishnan, S., & Yu, D. W. (2021). Biodiversity soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods in Ecology and Evolution*, 12(7), 1252–1264. <https://doi.org/10.1111/2041-210X.13602>
- Young, M. K., Smith, R., Pilgrim, K. L., Isaak, D. J., McKelvey, K. S., Parkes, S., Egge, J., & Schwartz, M. K. (2022). A molecular taxonomy of Cottus in western North America. *Western North American Naturalist*, 82(2), 307–345. <https://doi.org/10.3398/064.082.0208>
- Young, M. R., Proctor, H. C., deWaard, J. R., & Hebert, P. D. N. (2019). DNA barcodes expose unexpected diversity in Canadian mites. *Molecular Ecology*, 28(24), 5347–5359. <https://doi.org/10.1111/mec.15292>
- Zemlak, T. S., Ward, R. D., Connell, A. D., Holmes, B. H., & Hebert, P. D. N. (2009). DNA barcoding reveals overlooked marine fishes. *Molecular Ecology Resources*, 9, 237–242. <https://doi.org/10.1111/j.1755-0998.2009.02649.x>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Dziedzic, E., Sidlauskas, B., Cronn, R., Anthony, J., Cornwell, T., Friesen, T. A., Konstantinidis, P., Penaluna, B. E., Stein, S., & Levi, T. (2023). Creating, curating and evaluating a mitogenomic reference database to improve regional species identification using environmental DNA. *Molecular Ecology Resources*, 23, 1880–1904. <https://doi.org/10.1111/1755-0998.13855>