Transitioning from environmental genetics to genomics using mitogenome reference databases.

Emily Dziedzic[1]*, Brian Sidlauskas[1], Richard Cronn[2], James Anthony[3], Trevan Cornwell[3], Thomas Friesen[3], Peter Konstantinidis[1], Brooke E. Penaluna[2], Staci Stein[3], Taal Levi[1]

[1] Oregon State University

[2] USDA Forest Service, Pacific Northwest Research Station

[3] Oregon Department of Fish and Wildlife

*Corresponding Author:

Emily Dziedzic

emily.dziedzic@oregonstate.edu

Abstract

Species detection using eDNA is revolutionizing the 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 complete mitogenomic sequence information to evaluate the effectiveness of such data to differentiate, identify and detect taxa. We created the Oregon Biodiversity Genome Project working group to utilize recent advances in sequencing technology to create a database of complete, near error-free mitogenomic sequences for all of Oregon's resident freshwater fishes. So far, we have successfully assembled the complete mitogenomes of 313 specimens of freshwater fish representing 7 families, 55 genera, and 129 (88%) of the 146 resident species and lineages. Our comparative analyses of these sequences illustrate that the short (~150 bp) mitochondrial “barcode” regions typically used for eDNA assays are not consistently diagnostic for species-level identification and that no single region is best for metabarcoding Oregon’s fishes. However, often-overlooked intergenic regions of the mitogenome such as the D-loop have the potential to reliably diagnose and differentiate species. This project provides a blueprint for other researchers to follow as they build regional databases. It also illustrates the taxonomic value and limits of complete mitogenomic sequences, and how current eDNA assays and the “PCR-free” environmental genomics methods of the future can best leverage this information.

Introduction

The use of ambient genetic material—environmental DNA (eDNA)—to detect and identify metazoans in soil (Andersen et al. 2012, Drummond et al. 2015, Pansu et al. 2015), air (Clare et al. 2022, Lyndgaard et al. 2022), marine environments (Port et al. 2015, Yamamoto et al. 2017), and freshwater habitats (Deiner et al. 2016, Lim et al. 2016, Shaw et al. 2016, Valentini et al. 2016) is transforming how we monitor biodiversity. All eDNA detection methods depend on comprehensive reference databases of sequence information for all target, and sympatric, nontarget species in the clade of interest. The oft-cited lack of

41 comprehensive, reliably vouchered sequence information for many species (Schnell et al. 2010, Collins
42 et al. 2013, Bohmann et al. 2014, Porter and Hajibabaei 2018, Cordier et al. 2021) exposes the need to
43 build these reference databases using standardized sample collection, data and specimen curation, and
44 data-sharing protocols (Goldberg et al. 2016).
45
46 Molecular taxonomists have recommended microgenomic methods (e.g. metabarcoding, barcoding, and
47 single-species detection) for decades as a means to work around the limitations of morphology-based
48 identification (Hebert et al. 2003a). These molecular species detection methods rely on diagnostic
49 sequence information from prototypical candidate specimens to ensure that genetic fingerprints
50 "captured" in environmental samples are correctly identified. Though many eDNA-based identification
51 methods rely on short barcode regions (Deiner et al. 2017b), this approach has limitations. For example,
52 gene- and taxon-specific primers introduce a key source of error and bias by design (Yang et al. 2021) in
53 PCR amplification because they select certain DNA sequences over others (Fig 1a) (Deiner et al. 2017a).
54 PCR primer bias is often desirable because it allows preferential amplification of rare target genes and
55 taxa (e.g. metazoan targets in a sample dominated by eubacterial DNA). However, PCR biases due to
56 population variation or species divergence can lead to unwanted loss of information about target taxa.
57 Even minor binding biases among target sequences can affect PCR amplification substantially (Piñol et
58 al. 2015), preventing reliable measurements of species presence and/or relative abundance (Yang et al.
59 2021). In addition, if two or more species have not diverged at the locus targeted by a primer set, the
60 assay will neither diagnose the taxa nor properly assess species presence or abundance (Fig 1a).
61 Incomplete mitogenomic sequence information prevents *in silico* verification that primers will bind to
62 species' DNA or that the captured region will be diagnostic and correctly identify species. In addition,
63 holes in available sequence data and improper taxonomic assignments hinder accurate species
64 identification when querying eDNA metabarcoding results.
65
66 Not even full, reliable mitogenomic sequence information can avoid all issues. For example, hybridization
67 and organelle introgression from secondary contact can obscure the relationships of different species in

an environment. However, comprehensive databases of error-free, taxonomically-verified, full mitogenomic data have the potential to solve issues related to unreliable genetic data and lay the foundation for future environmental genomics technologies.

Cutting-edge environmental genomics methods involve sequencing all the DNA in an environmental sample, an approach known as “shotgun sequencing” (Taberlet et al. 2012 eDNA), “ecogenomics” (Béjà 2004), or “community genomics” (Bragg and Tyson, 2014). Researchers focusing on animals mostly target areas within the mitochondrial genome (“mitogenome”) for eDNA applications because mitochondrial DNA is frequently taxonomically diagnostic (Hebert et al. 2003a, Hebert et al. 2003b), relatively resistant to environmental degradation (Foran 2006), and more easily recovered from degraded samples than lower copy nuclear DNA targets (Hartmann et al. 2011). Once isolated and sequenced, whole mitogenomes can be used for taxonomic assignment in a variety of applications such as 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, reduced quantitative information, and incomplete taxonomic resolution. 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.

While advancements in sequencing technology have made it feasible to generate the voluminous data on which environmental mitogenomics depend, the lack of well-curated genomic sequence databases presents a bottleneck. Such databases are critical to environmental mitogenomics because they allow matching of any mtDNA fragment to complete, taxonomically-verified mitogenomes (Fig 1b). As such, any recovered fragment can yield valuable information on species presence and improved inference about abundance because primer biases are avoided. Local collections have the benefit of being able to

curate full mitogenomic data, providing sequence information for genes and intergenic regions, control error-checking, and identify and resolve taxonomic/genetic inconsistencies through re-sampling and re-validation.

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 is not uniformly linked to taxonomically-verified vouchers. In cases where vouchers do not exist, the link between DNA sequence and taxon is uncertain, and taxonomic identity can't be independently verified (Meiklejohn et al. 2019). In addition, error-checking on GenBank involves screening for contamination and protein coding but is not robust. Quality-checking at GenBank has improved in the decades since its inception (Leray et al. 2019), but sequences in GenBank can be draft quality, may contain errors, and can have incorrect taxonomic assignment (Meiklejohn et al. 2019), particularly at the species or subspecies level (Locatelli et al. 2020). This problem may be especially pronounced for taxon-rich groups like invertebrates (Leray et al. 2020). The Barcode of Life Data System (BOLD) provides an alternative to GenBank with a more rigorous requirement for taxonomic vouchers and a variety of tools to identify data anomalies and low quality records (Ratnasingham and Hebert 2007). However, BOLD skews heavily to information from Cytochrome c oxidase I (COI) due to BOLD's initial development around a single >500 bp barcode region in that single gene. As of this writing, COI sequences represent 80.8% of the data available for phylum Chordata 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). For example, the region Miya et al. (2015) found that was suitable for fish metabarcoding primers—two 20-30 bp conserved regions flanking a hypervariable region—occurred in the 12S mitochondrial gene.

Along with vouchered, error-checked sequence information from multiple mitogenomic loci, intra-specific sampling is also needed to identify taxonomically- or geographically-diagnostic DNA variation.

Redundant sequence data is required to align sequences for multiple individuals both within and among species to test primer-binding specificity and species diagnosability *in silico*. Here again, available reference sequence databases do not meet our needs. (O'Leary et al. 2016), GenBank's curated and well-annotated sequence dataset, undergoes additional rounds of error-checking and provides information for the entire mitogenome. However, as a rule it is non-redundant (About RefSeq 2021) with each species associated with only one complete mitogenome. It also is not comprehensive—RefSeq contained sequence data for 44% of Oregon's freshwater fishes when this study began. 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.

Alternatively, sequencing and assembling mitogenomes has become practical and affordable enough for a small consortium to sequence and assemble hundreds of mitogenomes on a single Illumina Novaseq sequencing lane. This means that little impedes 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. The ideal option for developing management-quality eDNA biodiversity surveys would involve extending the "BOLD model" to create curated reference databases of mitogenome sequences tied to vouchered specimens collected throughout discrete regions. Langlois et al. (2021) echoed this need to expand the range of species with full mitogenomic sequence information. The authors specifically call for full mitochondrial genome sequences for multiple examples per species so that robust, comprehensive sequence alignments can be produced to develop assays that avoid cross-binding of primers to non-target taxa or non-binding of primers to target DNA (Langlois et al. 2021).

Here we provide a roadmap for constructing such a curated mitogenomic reference library using vouchered specimens of freshwater fishes for the state of Oregon, U.S.A. While biodiversity and geographic complexity differs 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 km^2. By curating this

reference database of full mitogenomes, we simultaneously created the taxonomic reference information needed to identify freshwater 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 provides a set of pipelines that can guide other organizations as they develop reference sequence databases for their taxa and regions of interest.

Materials and Methods

Voucher Specimen and Tissue Collection

This effort was motivated by the Oregon Biodiversity Genome Project (OBGP; www.obgp.org), a multi-institution collaboration between scientists and wildlife managers at Oregon State University, the Oregon Department of Fish and Wildlife (ODFW), and the United States Forest Service. The primary objective of the OBGP is to develop a regional genetic reference database to facilitate statewide eDNA monitoring programs for Oregon's resident freshwater fishes. The specific goals of the OBGP (Fig 2a) 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 via a client-server database accessed via a web browser.

The study area encompassed the State of Oregon—the region of interest for our eDNA monitoring program. We collected fishes in Oregon and expanded to a few sites in northern California and Washington State (Fig 2b). We examined historical location records in existing collections such as Oregon State Ichthyology Collection and conferred with local biologists to identify resident fishes and occupied locations. For cases where we knew or suspected that deeply divergent evolutionary lineages existed within the present concept of a species, we aimed to include representatives of all lineages. Biologists from ODFW ultimately identified 146 native and nonnative freshwater fish species and lineages that currently reside in Oregon and strategized collections to span watersheds throughout the

state (Appendix S1). Each sampling kit (Appendix S2 Box S1) 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 (Appendix S2 Box S2), and a field notes sheet (Appendix S2 Box S3) for metadata collection. Collectors anaesthetized and euthanized all fish specimens prior to tissue collection by immersion in an aqueous solution of Tricaine mesylate (MS-222). For collections in 2017, we worked with partners (Appendix S3 collecting_entity) who followed accepted procedures under Oregon State University and USFS IACUC protocols, but an IACUC was not required by all partner institutions. Specimen collection by ODFW was conducted under the agency's statutory management authority and in 2018, 2019, and 2020 ODFW collected specimens for ESA-listed species under National Oceanic and Atmospheric Administration Permit numbers 21780, 22639, and 23527 respectively. Fish under USFWS jurisdiction (i.e. fish that are neither marine nor anadromous) were covered under ODFW's ESA Section 6 Cooperative Agreement with USFWS. Details regarding partner collection permits and authority are listed in Appendix S3. We instructed all partners to collect a minimum of ~0.5 cm^3 of tissue from each specimen, which was then placed in 95% EtOH for DNA extraction and sequencing. Euthanized fish were placed in 10% Formalin to ensure preservation of diagnostic features. When we failed to collect species or redundant examples of species, we augmented in-field collection with tissue samples loaned or gifted from North American ichthyology collections (OS14271, OS18056, OS18057, OS19982, OS19351, OS18993, OS20085, OS20084, OS20081, OS20080, OS20094, OS20088, OS20108, OS14271, OS22282, UW155929, UW158361, UAM:Fish:10376:401245, UAM:Fish:10464:374966, UAM:Fish:10464:374967). The goal of collecting 10 individuals per species was amended to collect three individuals and add specimens only if intraspecific genetic variation was detected in downstream mitogenome identity analyses (See below).

Taxonomic Verification, Accession, and Cataloging

ODFW biologists and partners identified specimens provisionally in the field and Oregon State Ichthyology Collection taxonomists verified and refined those identifications prior to cataloging the

specimens by morphological examination and reference to published keys (Markle and Tomelleri 2016, Wydoski and Whitney 2003). The Oregon State Ichthyology Collection has arranged to accession all vouchers and tissues, with full-bodied voucher specimens being transferred from formalin to isopropyl alcohol for permanent storage. Tissues were stored in 2.0 mL cryotubes at -70°C in 95% EtOH. Accessioning and cataloging were ongoing at the time of writing.

After generating sequence data (See below), we performed distance-based cluster analyses in Geneious to verify morphological identification 10.2.6 using default settings (Global alignment with free end gaps, Cost Matrix of 65% similarity, Tamura-Nei Genetic Distance Model, Neighbor-Joining (NJ) Tree build Method, Gap open penalty of 12, Gap extension penalty of 3). We used the NAD2 gene for Catostomidae, Centrarchidae, Cottidae, Cyprinidae, Ictaluridae, and Salmonidae NJ trees. Because the species of Lampreys (Petromyzontidae) in Oregon possess very similar mitogenomes, we concatenated the NAD4, NAD5, and NAD6 genes in order to increase the length of sequence examined in the search for genetic clusters. In cases of incongruence between morphological and genetic clustering, we revisited the anatomical identifications of the vouchers, investigated the possibility of swapped or contaminated molecular samples, and corrected identifications as needed.

DNA Extraction and Sequencing

We subsampled tissues into ~1.0 mm^3 volumes and extracted DNA from these subsamples using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) spin-column protocol for animal tissues. To further optimize the lysing process, we crushed tissues in-tube with a micropestle after incubation. We used the Invitrogen dsDNA Broad-Range assay Kit and a Qubit fluorometer (Invitrogen, Carlsbad, CA) to measure DNA concentrations and yield. For each extracted specimen, 100 µL of extract containing 100-2000 ng/µL of DNA was transferred to a 0.65 mL Bioruptor microtube and sonicated (30 s on, 90 s off; 6 cycles) to ~300 bp in length using the manufacturer's protocol using a Bioruptor Pico sonication system (Diagenode, Denville, NJ). We prepared libraries for next generation sequencing for the first two sequencing runs according to manufacturers' instructions using the NEBNext Ultra II DNA

Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) (Appendix S3 library_prep). Oregon State University's Center for Quantitative Life Sciences performed library preparation for the final two runs using the plexWell 96 Kit (SeqWell, Beverly, MA) (Appendix S3 library_prep). Paired-end (2 x 150 bp) sequencing was performed on all samples at multiplexing levels between 50 to 71 samples/lane (Appendix S3 spl) using an Illumina HiSeq 3000 at the Center for Quantitative Life Sciences.

Mitogenome Assembly

To capture geographic genetic variation of each resident species across its range within Oregon, we sequenced the first collected representative of each species and subsequently sequenced specimens collected from separate watersheds. We stored gzipped fastq sequencing files on 2 x 1TB enterprise NL-SAS hard drives, and performed mitogenome assemblies on 4 x 2.30 GHz 16-core processors using 512GB ECC RAM. We targeted the first collected representative of each species for sequencing and maximized geographic distance among subsequent sequenced specimens to capture geographic genetic variation of all species throughout Oregon. Mitochondrial genomes were assembled *de novo* from raw paired reads using SPAdes assembler (versions 3.12.0-3.15.3) (Bankevich et al. 2012) or getOrganelle 1.6.2 or 1.7.5 (Jin et al. 2020). Three mitogenomes were recovered by performing reference-guided filtering with BLAT (Kent 2002) using the complete mitogenome sequences of identical or closely related species prior to SPAdes assembly. We resolved one mitogenome by first mapping reads in Geneious 10.2.6 to the noncircular mitochondrial contig produced from SPAdes *de novo* assembly and then reiteratively mapping reads to *de novo* assemblies subsequently produced in Geneious. When *de novo* mitogenome assemblies did not form a single contig with an overlapping splice point, we performed assembly polishing using BWA (Li and Durbin 2009) followed by Pilon (Walker et al. 2014), or polca.sh from MaSuRCA 4.0.5 (Zimin et al. 2013), used Sealer from ABySS 2.3.1 (Paulino et al. 2015) for gap-closing on one sequence. We calculated quality values (QV) of mitogenome contigs with Merqury (Rhie et al. 2020) and mapped reads to assembled mitogenomes to evaluate coverage uniformity using Tablet 1.21.02.08 (Milne et al. 2013). We then performed polishing and reassembled mitogenomes exhibiting coverage anomalies in an attempt to resolve assembly errors. We annotated all

mitochondrial sequences using a combination of MITOS[2] WebServer (Al Arab et al. 2017, Donath et al. 2019) and Geneious using annotations from identical or closely-related species. Details on pipelines used for individual sequences can be found in the Supplemental Information (Appendix S3).

Mitogenome Variability

To analyze intra- and interspecies mitogenome variability, assembled mitogenomes from all species were aligned with MUSCLE multiple sequence alignment (Edgar 2004) in Geneious 10.2.6 using default parameters. After reciprocal rounds of morphological examination and molecular clustering were complete, we first 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 first used the R package SPIDER (Brown et al. 2012) to perform a sliding window analysis on the master alignment to locate areas with the highest density of taxonomically diagnostic nucleotides (TDN)—defined as windows where a nucleotide is fixed within species and different or unaligned in all other species. In addition, to identify genes with high variability, we plotted variability with heat maps, parallel coordinate plots, and radar charts using Superheat (Barter and Yu 2018), GGally (Schloerke et al. 2021), and fmsb (Nakazawa 2021) packages in R 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 treated described subspecies as full species to calculate mean percent identities for summaries. In order to calculate intraspecies, intrafamily/interspecies, and interfamily/interspecies mitogenome identities and the proportional relationships between these identities, families needed to have mitogenomic sequence information for multiple species and multiple specimens for each species. Seven families satisfied these requirements and were used for this comparative analysis: Catostomidae, Centrarchidae, Cyprinidae, Salmonidae, Ictaluridae, Cottidae, and Petromyzontidae.

Data Sharing

Mitogenome data generated for this project have been deposited in GenBank under the Oregon Biodiversity Genome Project BioProject. Accession numbers and sequence data are included in Supplemental Information (Appendix S3 genbank_accession and sequence). Data is also available at www.obgp.org/downloads and will be made accessible via the obgpdb.org client-server database. As of the time of writing, linked voucher, tissue, and DNA extract accessioning into the Oregon State Ichthyology Collection were ongoing.

Results

Voucher Specimen and Tissue Collection

Thus far, we have collected 625 specimens representing 129 fish species or members of species complexes collected from >240 sites. Twelve additional tissue samples of four species were acquired from natural history collections. Of the 133 collected species, 120 represent the original 146 fish species identified by ODFW 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 1 species endemic to western Washington state (Olympic Mudminnow, *Novumbra hubbsi*) and 1 newly identified lineage of Paiute Sculpin (*Cottus beldingii* ssp.) from the John Day River Basin in central Oregon. Specimen collections from within Oregon are ongoing as of the time of writing.

Taxonomic Verification

Our distance-based cluster analyses (Appendix S2 Figures S1-S4) in combination with laboratory-based taxonomic verification led to refinement of correction of 31 field identifications (9.9%) (Appendix S3 taxonomic_assessments). Seventeen were not evaluated by taxonomists and were reassigned based on NJ clustering alone; sixteen of these specimens were originally identified to the wrong species in the field, and one was identified to the wrong genus (Fig 3). In seven cases, NJ clustering revealed disagreements with taxonomic assignment even after laboratory examination, likely due to a mismatch between morphological identification and mitogenomic inheritance. We assigned eight specimens from cottid species complexes (*Cottus gulosus*/*perplexus*, n=7; *Cottus beldingii*/*confusus*, n=1) based on NJ

clustering after morphological evaluation because they could not be confidently identified by taxonomists due to the inadequacy of existing tools to discriminate these taxa.

Mitogenome Sequencing and Assembly

In total, 313 assembled mitogenomes representing 129 collected species and lineages were used for downstream analysis (Table 1). Nearly all 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). Petromyzontidae GC content ranged from 37.90% to 38.70% (mean 38.05%) (Appendix S3 gc_content). All remaining mitogenomes had GC content between 38.90% and 49.50% (mean 45.14%). Mitogenome sizes ranged from 16098 to 17185 bp in length (mean 16590). All but 17 assembled mitogenomes had error-free contigs when measured with k=31 using Merqury. Those assembled mitogenomes with errors had QVs between 40.7507 and 57.0952 (Appendix S3 contig_qv) indicating 1 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 that was not sufficient to exclude mitogenomes from downstream analyses (Appendix S3 assembly_notes).

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 analyzed gene regions were as follows: COI 7.446, CytB 9.549, NAD1 13.381, NAD5 17.161, NAD4 18.710, NAD2 20.065, 12S 20.977, 16S 25.976 (Fig 4b). The highest concentrations of TDNs occurred in the D-loop and the intergenic region between the NAD2 and COI genes (Fig 4a). Sliding window analyses of aligned taxonomic subsets of the entire mitogenome for Catostomids, Lamprey, Cyprinids, Salmonids, Cottids, and Centrarchids suggested that the density of TDN varies by taxonomic group (Fig

5) with mean TDN/$w_{150}i_{20}$ of 10.656, 12.104, 19.316, 20.209, 21.024, and 26.082 for these families, respectively.

Heat maps illustrate the degrees of similarity between families in different gene regions with greater identity represented by darker colors and lower percentage identity by bright yellow. These heat maps show that sequence identity among families is higher (i.e. there is greater similarity) in the COI gene versus other coding and noncoding gene regions (Fig 6a), which is concordant with the results from the sliding window analysis. Gene regions with the highest apparent contrast between intrafamily and interspecies interfamily percent identity overall are NAD2, NAD5, and 16S. Despite apparent differences in intrafamily identity, our analysis demonstrated that divergence in all mitochondrial gene regions and the D-loop can differentiate taxa at the family level (Fig 6a).

Zooming in to species differences within families, after reiterative morphologic assessment and NJ clustering all species were overall more similar to members of their own species than to members of other species within the same family, as expected (Appendix S4 Table S1). Overall, among species within the same family, there was greatest difference in percent identity (illustrated by the lowest proportional relationship between intra- and interspecies percent identities within each family) in the NAD2 region (mean 0.870), followed by the D-loop (0.882), NAD5 and NAD4 (0.885) genes, with 12S (0.963) and 16S (0.952) being the least differentiable gene regions (Fig 7). Intraspecies mean percent identities for all genes (12S and 16S rRNA and all coding genes >690 bp) and families analyzed ranged from 99.251 to 99.825% (Appendix S4 Table S2) illustrating that all regions analyzed, including the D-loop, are highly conserved within species (Fig 8). The most conserved genes were 12S, 16S, and COX2 regions, with lowest mean values found in the NAD2 and D-loop nevertheless still exceeding 99% identity (Appendix S4 Table S2). Radar charts of mean percent intraspecies, intrafamily interspecies, and interfamily interspecies identity (Fig 8) illustrate that different genes are more conserved among species within certain families than others and indicate that all genes appear to be sufficiently divergent to diagnose familial lineages. For example, species in Catostomidae vary little in sequences from rRNA and

all three COX genes, while the 12S and 16S genes are 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 contain diverged interspecies sequences. Mitochondrial gene sequences were the least diverged among Petromyzontidae species with the proportional relationship between intraspecies and intrafamily interspecies identity ranging from 0.982 to 0.997 (mean 0.989) (Appendix S4 Table S1), indicating they are almost identical. Full mitogenomes were highly conserved within species (mean 99.503% identity) and had sufficient divergence between species in the same family to suggest they would be diagnostic at the species level for Oregon fishes (Fig 9).

Discussion

We demonstrate a cooperative, affordable (wet and dry lab costs per mitogenome assembly ~$200), and feasible pipeline for constructing mitogenomic databases. The workflow begins with the collection of reference specimens and progresses through taxonomic verification, permanent 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.

One of the most daunting aspects of a project of this nature can be how to begin, but the steps for carrying out a similar endeavor in a region of interest for particular target taxa are straightforward: 1. Using historical collection data and local knowledge, determine all resident species and their ranges, 2. Break up the region into manageable subregions for sampling, 3. Based on the results from steps 1 and 2, 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 when possible, 4. Sequence and assemble specimens as they accumulate, measuring intraspecies sequence variability to inform continued collection. We have catalogued the pipelines we used for our bottom-up development of an eDNA

biodiversity reference collection and sequence database and provide our roadmap here (Appendix S2 Figure S5). This bottom-up approach harnesses the expertise, knowledge, and resources of researchers within their region of interest. This is essential as these individuals possess the intimate knowledge of species and geography needed to strategize and carry out collections as well as provide taxonomic expertise.

Because projects of this scale require moderate financial support and substantial human effort, we strongly recommend assessing available resources before launching a new endeavor. We were only able to complete this project on a relatively low budget because individuals donated considerable amounts of their time and because collaborating institutions provided us with access to fish biologists collecting throughout the state of Oregon, genetic laboratory facilities, and genetic sequencing at reduced costs. The workflow also depends on taxonomic expertise and experience identifying specimens within difficult families, namely those featuring many morphologically similar species, undescribed cryptic species and/or species complexes. Finally, access to the infrastructure and archival capacity of a natural history collection is vital, because the voucher specimens must be cataloged properly and preserved in perpetuity for the science to be verifiable and repeatable (Prendini et al. 2002, Astrin et al. 2013, Buckner et al. 2021).

Our efforts reaped the sequence data needed to analyze mitochondrial genetic variability among Oregon's freshwater fishes and gauge our capacity to differentiate species. Our analysis showed that mitochondrial sequences at every level, from individual genes to the entire mitogenome, are sufficiently conserved within species to provide reliable identifications. However, sequences must also diverge sufficiently among taxa to differentiate them. While whole mitogenomes and all individual genes can easily assign specimens to families, we found that not all mtDNA regions can differentiate closely-related species.

Despite their ubiquitous use, COI sequences are not necessarily the most diagnostic. Previous analyses by Hebert et al. (2003a and b) examined the use of COI for species differentiation by quantifying sequence divergence, and using NJ analyses and multidimensional scaling to assign species, using successful lepidopteran species assignment to extrapolate the suitability of their COI barcode to all animal taxa. Their arguments in favor of the COI as the core of a global bioidentification system for animals were logical, albeit speculative (Hebert et al. 2003a), and as we can see from our analyses, a comparative assessment of taxonomically diagnostic nucleotides among target taxonomic groups across all mitochondrial genes is also needed to evaluate diagnosability (Appendix S4 Figure S8). Although Hebert et al. (2003a) were optimistic about using COI for barcoding, their recommended ~650 bp segment of COI exceeds the length feasible for Illumina high throughput sequencing for barcoding and species differentiation (Meusnier et al. 2008). Hebert et al. (2003a) also 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 et al. 2003b). 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, Young et al. 2019; fish: Zemlak et al. 2009; birds: Herbert 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 taxonomically informative. It is unclear, however, if this is equated with being taxonomically diagnostic. In addition, these examinations did not comparatively evaluate other mitochondrial regions and used longer barcode regions inappropriate for eDNA assays using the more powerful Illumina sequencers. Shorter stretches of the COI have been isolated more recently for eDNA metabarcoding primarily for arthropods (Braukmann et al. 2019, Elbrecht et al. 2019, Hardulak et al. 2020) with examinations of vertebrates being exploratory (Valdez-Moreno et al. 2019) or having limited success diagnosing species (Hleap et al. 2021). Clearly, a more complete evaluation of the comparative diagnosability of different parts of the mitogenome is needed for a broad range of taxa. Here, we demonstrate that for Oregon's freshwater fishes, genic and intergenic regions other than the COI appear to better differentiate species in certain taxa.

We found multiple gene regions and the D-loop had high interspecies genetic distance and concentrations of TDNs within families, suggesting they would be diagnostic to the species level. To "capture" these regions with primers, the job is straightforward for single-species qPCR assays as the goal is to capture the target species and no other species. Areas with highest intrafamily distance, high intraspecies identity, and high mean concentrations of TDNs are likely the best candidates for this application—for Oregon's freshwater fishes this would be the NAD1, NAD2, NAD4, NAD5 genes and the D-loop (Table 2). Metabarcoding primers, in contrast with single-species qPCR assays, need to capture sequences from a broad range of taxa—different families, or even different orders, classes, or phyla—so there need to be shared regions (typically between 18 and 27 bases long) that can permit the binding of primers and avoid species dropout (Fig 1a). Essentially, a "Goldilocks" zone is needed for metabarcoding: a region with sufficient genetic divergence to differentiate species, but not to the degree that shared regions are unavailable for primer binding. The hairpin-loop structure of both rRNA regions likely makes them appropriate for metabarcoding despite low overall intrafamily/interspecies variability. This might explain why the most commonly used fish metabarcoding primers are found in the 12S region (Miya et al. 2015), despite this region's high intrafamily/interspecies identity relative to other regions. Complementary hairpin regions are conserved while loops introduce mutations, and our analysis shows that 12S and 16S regions contain TDN "spikes" (Table 2) likely representing clusters of TDNs in loop regions sandwiched by conserved hairpin areas. Although the 16S rRNA region had these "spikes" across more families than other genes (Table 2), our analysis suggests that for metabarcoding, no single region is best for all families of resident freshwater fish in Oregon, and primer-binding requirements further restrict which areas can be used.

For both single-species and metabarcoding assays, if a single, sufficiently-diagnostic mitochondrial region cannot be successfully captured by primers, 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 differentiate closely-related congeners (Dysthe et al. 2018). As an alternative, the variability of full

mitogenomes is significant and can more readily distinguish between species than single genes even among relatively conserved taxa. Species that are difficult to distinguish morphologically and often confound taxonomists also appear to be difficult to resolve genetically. The use of full mitogenomes or a combination of strategically-valuable mitochondrial genes derived from full mitogenomes may make it possible to discern even the most challenging-to-identify species such as individuals from the Cottus gulosus/Cottus perplexus complex. It is important to note, however, that difficulty differentiating species may not be the fault of the chosen genetic region. Failure to diagnose a species may be the biological reality rather than a fault of the method. For example, difficulties with cottid identification may be due to insufficiently diverged lineages or indistinct morphology (Rowsey and Egge 2017), and lamprey taxa may be oversplit (April et al. 2011) based on life cycle rather than actual genetic divergence. That said, due to being inherited matrilineally, mitogenomic information on its own cannot distinguish hybrid species and nuclear genetic information is needed to untangle the genetic complexities of introgression as a result of hybridization—the likely culprit behind difficulties with catostomid species differentiation (Dowling et al. 2016)—and secondary contact.

One important consideration is that the sequencing and assembly pipeline we used works well for fishes, whose mitogenomes are known to contain fewer repeats, insertions, and deletions compared to other vertebrates (Formenti et al. 2021), but this may not be appropriate for other taxa. It would be sensible to have an understanding of the structure and makeup of the mitochondrial and nuclear genomes of target taxa prior to curating mitogenomic sequences to ensure that wet and dry laboratory pipelines can successfully resolve mitogenomes. This may mean increased sequencing depth with short read sequencers for species with large nuclear genomes, or using combined long-read and short-read sequencing as demonstrated by the Formenti et al. (2021) to combat issues with insertions, deletions, and repeats known cause problems in the sequencing and assembly pipeline (Tørresen et al. 2019).

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 provides a useful

497 genetic reference for species identification and the genetic information needed to develop primers for
498 single-species and metabarcoding assays. Arguably more importantly, it furnishes researchers with the
499 data needed to move away from microgenomics such as barcoding or metabarcoding and into capture
500 enrichment (Wilcox et al. 2018) or PCR-free environmental genomics. As previously mentioned,
501 transitioning to mitogenomics allows for PCR-free approaches that by definition solve the problems
502 associated with PCR amplification biases (Fig 1a) (Piñol et al. 2015) making accurate quantification of
503 relative species abundance in a sample a real possibility (Yang et al. 2021). Full mitogenomic data also
504 permits greater taxonomic resolution and frees us from a reliance on short sequences that are
505 inconsistently diagnostic across taxa. Additionally, compiling such a sequence database simultaneously
506 expands the global repository of available genetic data, and creates a genetic archive to test published
507 qPCR primers in silico against local species for binding efficiency and species differentiation, and to
508 query metabarcoding sequencing results.
509
510 In anticipation of future applications, many organizations are assembling whole nuclear genomes.
511 Example consortia include the “Bat1K” (Teeling et al. 2018) and “1000 Fungal Genomes” (Grigoriev et al.
512 2014) projects sequencing 1000 species of bats and fungi respectively, the “i5k” consortium sequencing
513 5000 arthropod genomes (i5k Consortium), the “10KP” and “P10K” projects sequencing 10,000 plant
514 (Cheng et al. 2018) and protist (Miao et al. 2020) species, and the “GIGA” project (GIGA Community of
515 Scientists 2014) dedicated to sequencing invertebrate genomes. Even more ambitious projects are
516 efforts to sequence genomes from representatives of every vertebrate species (Vertebrate Genomes
517 Project (VGP; Rhie et al. 2021), and all of Earth's eukaryotic biodiversity by 2027 (Earth Biogenome
518 Project; Lewin et al. 2018). These large-scale, expensive, and top-down efforts will create nuclear
519 genome reference databases for many species in the coming decades, but their global focus is unlikely
520 to provide comprehensive genetic information for a specific geographic region or taxonomic groups of
521 interest in the near future. For example, as of the time of writing the VGP has produced 110 nuclear
522 assemblies and published data for 125 mitogenomes (Formenti et al. 2021). This is an invaluable
523 contribution but lacks the redundant sequence information for species required to develop new eDNA

assays and test pre-existing ones *in silico*. In addition, although some of these efforts, such as the VGP, have created clearinghouses to provide easy access to their assemblies, data for some projects can only be found through individual publications, making it less accessible. Nevertheless, creating reference databases of full nuclear genomes for all life should be our ultimate goal. At present, however, the costs and complications associated with nuclear genome sequencing are out of reach for many small research labs, so it is fortunate that so much can be gleaned from complete mitogenomes on their own.

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. Widely ranging mitogenomic databases would expand eDNA monitoring potential to more regions. A repository of vouchered samples and full mitogenomic information as described here not only provides 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 simultaneously lays the groundwork to compile all the available genetic information for freshwater fishes or other taxa in a region of interest.

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, United States Fish and Wildlife Service, and the California Department of Fish and Wildlife for providing valuable contributions toward our collection goals. In addition, we greatly appreciate the individual contributions of Mark Buettner, Nolan Banish, Bruce Hansen, Paul Divine, and Dave Hering toward collections. Laura Hauck and Lucas Longway provided invaluable assistance with laboratory training and Shawn Clements had the foresight and drive to get this project off the ground. Many thanks to the University of Washington, the University of Alaska Museum, and the Oregon State

University Ichthyology Collection for providing us with vouchered tissue samples. We also thank Dave Markel and Alvaro Cortes 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.

Author Contributions

ED, RC, BEP, TL, and PK conceived the project. JA, TC, TF, BS, BEP, and SS strategized and carried out specimen collection. ED analyzed the data. ED, TL, BS, RC, BEP, JA, PK, and TF wrote the paper.

References

About RefSeq [Internet]. 2021. Bethesda (MD): National Center for Biotechnology Information; [cited 3 February 2021]. Available from: https://www.ncbi.nlm.nih.gov/refseq/about/

Al Arab M, Höner zu Siederdissen C, Tout K, Sahyoun AH, Stadler PF, Bernt M. 2017. Accurate annotation of protein-coding genes in mitochondrial genomes. Molecular Phylogenetics and Evolution. 106:209–216. doi:10.1016/j.ympev.2016.09.024.

Andersen K, Bird KL, Rasmussen M, Haile J, Breuning-Madsen H, Kjaer KH, Orlando L, Gilbert MTP, Willerslev E. 2012. Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. Molecular Ecology. 21(8):1966–1979. doi:10.1111/j.1365-294X.2011.05261.x.

April J, Mayden RL, Hanner RH, 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. doi:10.1073/pnas.1016437108.

Arulandhu AJ, Staats M, Hagelaar R, Voorhuijzen MM, Prins TW, Scholtens I, Costessi A, Duijsings D, Rechenmann F, Gaspar FB, et al. 2017. Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. GigaScience. 6(10). doi:10.1093/gigascience/gix080. [accessed 2021 Nov 15]. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/gix080/4085313.

Astrin J, Zhou X, Misof B. 2013. The importance of biobanking in molecular taxonomy, with proposed definitions for vouchers in a molecular context. ZK. 365:67–70. doi:10.3897/zookeys.365.5875.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 19(5):455–477. doi:10.1089/cmb.2012.0021.

Barter RL, 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. doi:10.1080/10618600.2018.1473780.

Béjà O. 2004. To BAC or not to BAC: Marine ecogenomics. Current Opinion in Biotechnology. 15(3):187–190. doi:10.1016/j.copbio.2004.03.005.

Benson D. 1996. GenBank. Nucleic Acids Research. 24(1):1–5. doi:10.1093/nar/24.1.1.

Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, de Bruyn M. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution. 29(6):358–367. doi:10.1016/j.tree.2014.04.003.

Bragg L, Tyson GW. 2014. Metagenomics using next-generation sequencing. In: Paulsen IT, Holmes AJ, editors. Environmental Microbiology. Vol. 1096. Totowa, NJ: Humana Press. (Methods in Molecular Biology). p. 183–201. [accessed 2022 Feb 2]. http://link.springer.com/10.1007/978-1-62703-712-9_15.

Braukmann TWA, Ivanova NV, Prosser SWJ, Elbrecht V, Steinke D, Ratnasingham S, de Waard JR, Sones JE, Zakharov EV, Hebert PDN. 2019. Metabarcoding a diverse arthropod mock community. Mol Ecol Resour. 19(3):711–727. doi:10.1111/1755-0998.13008.

Brown SDJ, Collins RA, Boyer S, Lefort M, Malumbres-Olarte J, Vink CJ, Cruickshank RH. 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. doi:10.1111/j.1755-0998.2011.03108.x.

Buckner JC, Sanders RC, Faircloth BC, Chakrabarty P. 2021. The critical importance of vouchers in genomics. eLife. 10:e68264. doi:10.7554/eLife.68264.

Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux P-M, Li F-W, Melkonian B, Mavrodiev EV, Sun W, et al. 2018. 10KP: A phylodiverse genome sequencing plan. GigaScience. 7(3). doi:10.1093/gigascience/giy013. [accessed 2022 Feb 2]. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giy013/4880447.

Clare EL, Economou CK, Bennett FJ, Dyer CE, Adams K, McRobie B, Drinkwater R, Littlefair JE. 2022 Jan. Measuring biodiversity from DNA in the air. Current Biology.:S096098222101650X. doi:10.1016/j.cub.2021.11.064.

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. Nucleic Acids Res. 44(D1):D67–D72. doi:10.1093/nar/gkv1276.

Collins RA, Armstrong KF, Holyoake AJ, Keeling S. 2013. Something in the water: Biosecurity monitoring of ornamental fish imports using environmental DNA. Biol Invasions. 15(6):1209–1215. doi:10.1007/s10530-012-0376-9.

Cordier T, Alonso-Sáez L, Apothéloz-Perret-Gentil L, Aylagas E, Bohan DA, Bouchez A, Chariton A, Creer S, Frühe L, Keck F, et al. 2021. Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. Mol Ecol. 30(13):2937–2958. doi:10.1111/mec.15472.

Curd EE, Gold Z, Kandlikar GS, Gomer J, Ogden M, O'Connell T, Pipes L, Schweizer TM, Rabichow L, Lin M, et al. 2019. Anacapa Toolkit : An environmental DNA toolkit for processing multilocus metabarcode datasets. Yu D, editor. Methods Ecol Evol. 10(9):1469–1475. doi:10.1111/2041-210X.13214.

Cywinska A, Hunter FF, Hebert PDN. 2006. Identifying Canadian mosquito species through DNA barcodes. Med Vet Entomol. 20(4):413–424. doi:10.1111/j.1365-2915.2006.00653.x.

de Santana CD, Parenti LR, Dillman CB, Coddington JA, Bastos DA, Baldwin CC, Zuanon J, Torrente-Vilara G, Covain R, Menezes NA, et al. 2021. The critical role of natural history museums in advancing eDNA for biodiversity studies: A case study with Amazonian fishes. Sci Rep. 11(1):18159. doi:10.1038/s41598-021-97128-3.

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P. 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. Biol Lett. 10(9):20140562. doi:10.1098/rsbl.2014.0562.

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, Vere N de, et al. 2017a. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology. 26(21):5872–5895. doi:10.1111/mec.14350.

Deiner K, Fronhofer EA, Mächler E, Walser J-C, Altermatt F. 2016. Environmental DNA reveals that rivers are conveyor belts of biodiversity information. Nat Commun. 7(1):12544. doi:10.1038/ncomms12544.

Deiner K, Renshaw MA, Li Y, Olds BP, Lodge DM, Pfrender ME. 2017b. Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. Methods Ecol Evol. 8(12):1888–1898. doi:10.1111/2041-210X.12836.

Donath A, Jühling F, Al-Arab M, Bernhart SH, Reinhardt F, Stadler PF, Middendorf M, Bernt M. 2019. Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. Nucleic Acids Research. 47(20):10543–10552. doi:10.1093/nar/gkz833.

Dowling TE, Markle DF, Tranah GJ, Carson EW, Wagman DW, May BP. 2016. Introgressive Hybridization and the Evolution of Lake-Adapted Catostomid Fishes. Aravanopoulos FA, editor. PLoS ONE. 11(3):e0149884. doi:10.1371/journal.pone.0149884.

Drummond AJ, Newcomb RD, Buckley TR, Xie D, Dopheide A, Potter BC, Heled J, Ross HA, Tooman L, Grosser S, et al. 2015. Evaluating a multigene environmental DNA approach for biodiversity assessment. GigaSci. 4(1):46. doi:10.1186/s13742-015-0086-1.

Dysthe JC, Franklin TW, McKelvey KS, Young MK, Schwartz MK. 2018. An improved environmental DNA assay for bull trout (Salvelinus confluentus) based on the ribosomal internal transcribed spacer I. Doi H, editor. PLoS ONE. 13(11):e0206851. doi:10.1371/journal.pone.0206851.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32(5):1792–1797. doi:10.1093/nar/gkh340.

Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, Zakharov EV, Hebert PDN, Steinke D. 2019. Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ. 7:e7745. doi:10.7717/peerj.7745.

Foran DR. 2006. Relative Degradation of Nuclear and Mitochondrial DNA: An Experimental Approach*. J Forensic Sci. 51(4):766–770. doi:10.1111/j.1556-4029.2006.00176.x.

Formenti G, Rhie A, Balacco J, Haase B, Mountcastle J, Fedrigo O, Brown S, Capodiferro MR, Al-Ajli FO, et al. 2021. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome Biol. 22(1):120. doi:10.1186/s13059-021-02336-9.

Francis CM, Borisenko AV, Ivanova NV, Eger JL, Lim BK, Guillén-Servent A, Kruskop SV, Mackie I, Hebert PDN. 2010. The Role of DNA Barcodes in Understanding and Conservation of Mammal Diversity in Southeast Asia. Joly S, editor. PLoS ONE. 5(9):e12575. doi:10.1371/journal.pone.0012575.

GIGA Community of Scientists. 2014. The Global Invertebrate Genomics Alliance (GIGA): Developing community resources to study diverse invertebrate genomes. Journal of Heredity. 105(1):1–18. doi:10.1093/jhered/est084.

Goldberg CS, Turner CR, Deiner K, Klymus KE, Thomsen PF, Murphy MA, Spear SF, McKee A, Oyler-McCance SJ, Cornman RS, et al. 2016. Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods Ecol Evol. 7(11):1299–1307. doi:10.1111/2041-210X.12595.

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, et al. 2014. MycoCosm portal: Gearing up for 1000 fungal genomes. Nucl Acids Res. 42(D1):D699–D704. doi:10.1093/nar/gkt1183.

Hardulak LA, Morinière J, Hausmann A, Hendrich L, Schmidt S, Doczkal D, Müller J, Hebert PDN, Haszprunar G. 2020. DNA metabarcoding for biodiversity monitoring in a national park: Screening for invasive and pest species. Mol Ecol Resour. 20(6):1542–1557. doi:10.1111/1755-0998.13212.

Hartmann N, Reichwald K, Wittig I, Dröse S, Schmeisser S, Lück C, Hahn C, Graf M, Gausmann U, Terzibasi E, et al. 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. doi:10.1111/j.1474-9726.2011.00723.x.

Hebert Paul D. N., Cywinska A, Ball SL, deWaard JR. 2003a. Biological identifications through DNA barcodes. Proc R Soc Lond B. 270(1512):313–321. doi:10.1098/rspb.2002.2218.

Hebert Paul D.N., Ratnasingham S, de Waard JR. 2003b. Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc Lond B. 270(suppl_1). doi:10.1098/rsbl.2003.0025. [accessed 2021 Oct 28]. https://royalsocietypublishing.org/doi/10.1098/rsbl.2003.0025.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. 2004. Identification of Birds through DNA Barcodes. Charles Godfray, editor. PLoS Biol. 2(10):e312. doi:10.1371/journal.pbio.0020312.

Hleap JS, Littlefair JE, Steinke D, Hebert PDN, Cristescu ME. 2021. Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Mol Ecol Resour. 21(7):2190–2203. doi:10.1111/1755-0998.13407.

i5K Consortium. 2013. The i5K Initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. Journal of Heredity. 104(5):595–600. doi:10.1093/jhered/est050.

Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z. 2020. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 21(1):241. doi:10.1186/s13059-020-02154-5.

Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, 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. doi:10.3732/ajb.1100570.

Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. Genome Research. 12(4):656–664. doi:10.1101/gr.229202.

Kerr KCR, Lijtmaer DA, Barreira AS, Hebert PDN, Tubaro PL. 2009. Probing Evolutionary Patterns in Neotropical Birds through DNA Barcodes. DeSalle R, editor. PLoS ONE. 4(2):e4379. doi:10.1371/journal.pone.0004379.

Langlois VS, Allison MJ, Bergman LC, To TA, Helbing CC. 2021. The need for robust qPCR-based eDNA detection assays in environmental monitoring and species inventories. Environmental DNA. 3(3):519–527. doi:10.1002/edn3.164.

Leray M, Knowlton N, Ho S-L, Nguyen BN, Machida RJ. 2019. GenBank is a reliable resource for 21st century biodiversity research. Proc Natl Acad Sci USA. 116(45):22651–22656. doi:10.1073/pnas.1911714116.

Leray M, Knowlton N, Ho S-L, Nguyen BN, Machida RJ. 2020. Reply to Locatelli et al.: Evaluating species-level accuracy of GenBank metazoan sequences will require experts' effort in each group. Proc Natl Acad Sci USA. 117(51):32213–32214. doi:10.1073/pnas.2019903117.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, et al. 2018. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci USA. 115(17):4325–4333. doi:10.1073/pnas.1720115115.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. 2015. Plant DNA barcoding: From gene to genome: Plant identification using DNA barcodes. Biol Rev. 90(1):157–166. doi:10.1111/brv.12104.

Lim NKM, Tay YC, Srivathsan A, Tan JWT, Kwik JTB, Baloğlu B, Meier R, Yeo DCJ. 2016. Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. R Soc open sci. 3(11):160635. doi:10.1098/rsos.160635.

Locatelli NS, McIntyre PB, Therkildsen NO, Baetscher DS. 2020. GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. Proc Natl Acad Sci USA. 117(51):32211–32212. doi:10.1073/pnas.2007421117.

Lynggaard C, Bertelsen MF, Jensen CV, Johnson MS, Frøslev TG, Olsen MT, Bohmann K. 2022 Jan. Airborne environmental DNA for terrestrial vertebrate community monitoring. Current Biology.:S0960982221016900. doi:10.1016/j.cub.2021.12.014.

Markle DF, Tomelleri JR. 2016. A guide to freshwater fishes of Oregon. Corvallis: Oregon State University Press.

Meiklejohn KA, Damaso N, Robertson JM. 2019. Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. Fugmann SD, editor. PLoS ONE. 14(6):e0217084. doi:10.1371/journal.pone.0217084.

Meusnier I, Singer GA, Landry J-F, Hickey DA, Hebert PD, Hajibabaei M. 2008. A universal DNA mini-barcode for biodiversity analysis. BMC Genomics. 9(1):214. doi: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. doi:10.1016/j.xinn.2020.100058.

Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics. 14(2):193–202. doi:10.1093/bib/bbs012.

Miya M., Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, Minamoto T, Yamamoto S, Yamanaka H, Araki H, et al. 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. doi:10.1098/rsos.150088.

Nakazawa M. 2021. fmsb: Functions for Medical Statistics Book with some Demographic Data. https://CRAN.R-project.org/package=fmsb.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44(D1):D733–D745. doi:10.1093/nar/gkv1189.

Pansu J, De Danieli S, Puissant J, Gonzalez J-M, Gielly L, Cordonnier T, Zinger L, Brun J-J, Choler P, Taberlet P, et al. 2015. Landscape-scale distribution patterns of earthworms inferred from soil DNA. Soil Biology and Biochemistry. 83:100–105. doi:10.1016/j.soilbio.2015.01.004.

Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, Birol I. 2015. Sealer: A scalable gap-closing application for finishing draft genomes. BMC Bioinformatics. 16(1):230. doi:10.1186/s12859-015-0663-4.

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. Mol Ecol Resour. 15(4):819–830. doi:10.1111/1755-0998.12355.

Port JA, O'Donnell JL, Romero-Maraccini OC, Leary PR, Litvin SY, Nickols KJ, Yamahara KM, Kelly RP. 2016. Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. Mol Ecol. 25(2):527–541. doi:10.1111/mec.13481.

Porter TM, Hajibabaei M. 2018. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Mol Ecol. 27(2):313–338. doi: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: DeSalle R, Giribet G, Wheeler W, editors. Techniques in Molecular Systematics and Evolution. Basel: Birkhäuser Basel. p. 176–248. [accessed 2022 Feb 22]. http://link.springer.com/10.1007/978-3-0348-8125-8_11.

Ratnasingham S, Hebert PDN. 2007. BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org): BARCODING. Molecular Ecology Notes. 7(3):355–364. doi:10.1111/j.1471-8286.2007.01678.x.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 592(7856):737–746. doi:10.1038/s41586-021-03451-0.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21(1):245. doi:10.1186/s13059-020-02134-9.

Rimet F, Aylagas E, Borja Á, Bouchez A, Canino A, Chauvin C, Chonova T, Ciampor Jr F, Costa FO, Ferrari BJD, et al. 2021. Metadata standards and practical guidelines for specimen and DNA curation when building barcode reference libraries for aquatic life. MBMG. 5:e58056. doi:10.3897/mbmg.5.58056.

Rowsey DM, Egge JJ. 2017. Morphometric Analysis of Two Enigmatic Sculpin Species, Cottus gulosus and Cottus perplexus (Scorpaeniformes: Cottidae). Northwestern Naturalist. 98(3):190–202. doi:10.1898/NWN16-23.1.

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 IB, Fraser M, Willerslev E, Gilbert MTP. 2010. Characterisation of insect and plant origins using DNA extracted from small volumes of bee honey. Arthropod-Plant Interactions. 4(2):107–116. doi:10.1007/s11829-010-9089-0.

Seifert KA, Samson RA, deWaard JR, Houbraken J, Levesque CA, Moncalvo J-M, Louis-Seize G, Hebert PDN. 2007. Prospects for fungus identification using COI DNA barcodes, with Penicillium as a test case. Proceedings of the National Academy of Sciences. 104(10):3901–3906. doi:10.1073/pnas.0611691104.

Shaw JLA, Clarke LJ, Wedderburn SD, Barnes TC, Weyrich LS, Cooper A. 2016. Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. Biological Conservation. 197:131–138. doi:10.1016/j.biocon.2016.03.010.

Sheffield CS, Hebert PDN, Kevan PG, Packer L. 2009. DNA barcoding a regional bee (Hymenoptera: Apoidea) fauna and its potential for ecological studies. Molecular Ecology Resources. 9:196–207. doi:10.1111/j.1755-0998.2009.02645.x.

Smith MA, Poyarkov NA, Hebert PDN. 2008. DNA barcoding: COI DNA barcoding amphibians: Take the chance, meet the challenge. Molecular Ecology Resources. 8(2):235–246. doi:10.1111/j.1471-8286.2007.01964.x.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH. 2012. Environmental DNA. Molecular Ecology. 21(8):1789–1793. doi:10.1111/j.1365-294X.2012.05542.x.

Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C, et al. 2015. High-throughput monitoring of wild bee diversity and abundance via mitogenomics. Gilbert M, editor. Methods Ecol Evol. 6(9):1034–1043. doi:10.1111/2041-210X.12416.

Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, Bat1K Consortium. 2018. Bat Biology, Genomes, and the Bat1K Project: To generate chromosome-level genomes for all living bat species. Annu Rev Anim Biosci. 6(1):23–46. doi:10.1146/annurev-animal-022516-022811.

Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, et al. 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. doi:10.1093/nar/gkz841.

Valdez-Moreno M, Ivanova NV, Elías-Gutiérrez M, Pedersen SL, Bessonov K, Hebert PDN. 2019. Using eDNA to biomonitor the fish community in a tropical oligotrophic lake. Doi H, editor. PLoS ONE. 14(4):e0215505. doi:10.1371/journal.pone.0215505.

Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, Bellemain E, Besnard A, Coissac E, Boyer F, et al. 2016. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Mol Ecol. 25(4):929–942. doi:10.1111/mec.13428.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. Wang J, editor. PLoS ONE. 9(11):e112963. doi:10.1371/journal.pone.0112963.

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 26 September 2019.

Wilcox TM, Zarn KE, Piggott MP, Young MK, McKelvey KS, Schwartz MK. 2018. Capture enrichment of aquatic environmental DNA: A first proof of concept. Molecular Ecology Resources. 18(6):1392–1401. doi:10.1111/1755-0998.12928.

Wydoski RS, Whitney RR. 2003. Inland fishes of Washington. 2nd edition, revised and expanded. Bethesda, MD: 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. Sci Rep. 7(1):40368. doi:10.1038/srep40368.

Yang C, Bohmann K, Wang X, Cai W, Wales N, Ding Z, Gopalakrishnan S, Yu DW. 2021. Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. Methods Ecol Evol. 12(7):1252–1264. doi:10.1111/2041-210X.13602.

Young MR, Proctor HC, deWaard JR, Hebert PDN. 2019. DNA barcodes expose unexpected diversity in Canadian mites. Mol Ecol. 28(24):5347–5359. doi:10.1111/mec.15292.

Zemlak TS, Ward RD, Connell AD, Holmes BH, Hebert PDN. 2009. DNA barcoding reveals overlooked marine fishes. Molecular Ecology Resources. 9:237–242. doi:10.1111/j.1755-0998.2009.02649.x.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. Bioinformatics. 29(21):2669–2677. doi:10.1093/bioinformatics/btt476.

Data Availability Statement

Appendix S3 contains all sequence data and details regarding the methodological processes used to derive the sequence data used in downstream analysis for this paper.

Table 1. Assembled Mitogenome Taxa Counts: OBGP specimens with assembled mitogenomes are grouped according to taxonomic designation with counts for each taxonomic level.

OBGP ID	Family		Genus		Species		Subspecies/Lineage	
OBGP-2019-237	Acipenseridae	3	*Acipenser*	3	*medirostris*	1		
OBGP-2018-244					*transmontanus*	2		
OBGP-2018-267								
OBGP-2019-023	Atherinopsidae	2	*Atherinops*	2	*affinis*	2		
OBGP-2019-230								
OBGP-2018-006	Catostomidae	61	*Catostomus*	48	*bondi*	3		
OBGP-2018-023								
OBGP-2019-038								
OBGP-2017-206					*columbianus*	1		
OBGP-2017-199					*macrocheilus*	2		
OBGP-2018-203								
OBGP-2017-020					*microps*	8		
OBGP-2017-021								
OBGP-2017-022								
OBGP-2017-090								
OBGP-2017-091								
OBGP-2017-092								
OBGP-2017-103								
OBGP-2017-104								
OBGP-2017-007					*occidentalis*	9	*lacusanserinus*	5
OBGP-2017-008								
OBGP-2017-009								
OBGP-2017-011								
OBGP-2018-226								
OBGP-2017-093								
OBGP-2017-096								
OBGP-2017-100								
OBGP-2017-101								
OBGP-2017-179					*rimiculus*	6	Jenny Creek	3
OBGP-2017-183								
OBGP-2017-184								
OBGP-2018-215							Klamath	2
OBGP-2019-156								
OBGP-2018-264							Rogue	1
OBGP-2017-232					*snyderi*	10		
OBGP-2017-233								
OBGP-2017-234								
OBGP-2017-235								
OBGP-2017-236								
OBGP-2017-237								
OBGP-2017-251								
OBGP-2017-252								
OBGP-2017-253								
OBGP-2017-254								
OBGP-2018-092					*tahoensis*	1		
OBGP-2019-142					*tsiltcoosensis*	6	Coos	2
OBGP-2019-148								
OBGP-2019-145							Coquille	1
OBGP-2017-216							Siuslaw	1
OBGP-2017-064							Umpqua	2
OBGP-2017-147								
OBGP-1993-001					*warnerensis*	2		
OBGP-2018-186								
OBGP-2017-302			*Chasmistes*	2	*brevirostris*	2		
OBGP-2017-304								
OBGP-2017-230			*Deltistes*	11	*luxatus*	11		
OBGP-2017-231								
OBGP-2017-249								
OBGP-2017-306								
OBGP-2017-309								
OBGP-2017-310								
OBGP-2017-311								
OBGP-2017-313								
OBGP-2017-314								
OBGP-2017-315								
OBGP-2017-316								
OBGP-2017-297	Centrarchidae	24	*Archoplites*	2	*interruptus*	2		
OBGP-2017-312								
OBGP-2017-178			*Lepomis*	10	*cyanellus*	1		
OBGP-2017-275					*gibbosus*	5		
OBGP-2017-320								
OBGP-2017-360								
OBGP-2018-042								
OBGP-2018-159								
OBGP-2018-172					*gulosus*	1		
OBGP-2017-277					*macrochirus*	3		
OBGP-2017-381								
OBGP-2018-174								
OBGP-2017-063			*Micropterus*	7	*dolomieu*	2		
OBGP-2017-151								
OBGP-2017-238					*salmoides*	5		
OBGP-2017-241								
OBGP-2017-276								
OBGP-2017-287								
OBGP-2019-057								
OBGP-2017-001			*Pomoxis*	5	*nigromaculatus*	5		
OBGP-2017-050								
OBGP-2017-308								
OBGP-2017-349								
OBGP-2018-177								
OBGP-2018-111	Clupeidae	2	*Alosa*	2	*sapidissima*	2		
OBGP-2018-179								
OBGP-2018-105	Cobitidae	2	*Misgurnus*	2	*anguillicaudatus*	2		
OBGP-2018-185								
OBGP-2017-326	Cottidae	46	*Cottus*	45	*aleuticus*	2		
OBGP-2017-350								
OBGP-2017-012					*asper*	5		
OBGP-2017-220								
OBGP-2017-269								
OBGP-2017-272								
OBGP-2017-273								
OBGP-2017-148					*beldingii*	6	ssp.	3
OBGP-2018-320								
OBGP-2018-321								
OBGP-2017-203								
OBGP-2018-156								
OBGP-2018-287								
OBGP-2018-036					*bendirei*	1		
OBGP-2017-351					*confusus*	1		
OBGP-2016-004					*gulosus*	4		
OBGP-2017-056								
OBGP-2017-084								
OBGP-2017-188								
OBGP-2017-218					*klamathensis*	3		
OBGP-2017-246								
OBGP-2017-247								
OBGP-2018-127					*marginatus*	2		
OBGP-2018-138								
OBGP-2017-132					*perplexus*	10		
OBGP-2017-134								
OBGP-2017-138								
OBGP-2017-140								
OBGP-2017-192								
OBGP-2017-193								
OBGP-2017-270								
OBGP-2017-285								
OBGP-2017-318								
OBGP-2017-346								
OBGP-2019-138					*pitensis*	2		
OBGP-2019-150								
OBGP-2019-160					*polyporus*	1		
OBGP-2017-212					*princeps*	1		
OBGP-2017-141					*rhotheus*	5		
OBGP-2017-201								
OBGP-2017-288								
OBGP-2018-012								
OBGP-2019-178								
OBGP-2017-162					*tenuis*	2		
OBGP-2017-171								
OBGP-2019-025			*Enophrys*	1	*bison*	1		
OBGP-2017-200	Cyprinidae	79	*Acrocheilus*	2	*alutaceus*	2		
OBGP-2017-207								
OBGP-2017-239			*Carassius*	2	*auratus*	2		
OBGP-2017-240								
OBGP-2018-047			*Cyprinus*	2	*carpio*	2		
OBGP-2018-288								
OBGP-2017-170			*Gila*	3	*coerulea*	3		
OBGP-2017-244								
OBGP-2017-245								
OBGP-2018-221			*Hesperoleucus*	2	*symmetricus*	2		
OBGP-2018-223								
OBGP-2017-327			*Mylocheilus*	1	*caurinus*	1		
OBGP-2017-370			*Notemigonus*	1	*crysoleucas*	1		
OBGP-2017-032			*Oregonichthys*	5	*crameri*	2		
OBGP-2017-137								
OBGP-2016-002					*kalawatseti*	3		
OBGP-2017-099								
OBGP-2018-232								
OBGP-2017-176			*Pimephales*	3	*promelas*	3		
OBGP-2018-033								
OBGP-2018-216								
OBGP-2017-019			*Ptychocheilus*	5	*oregonensis*	2		
OBGP-2017-195								
OBGP-2017-135					*umpquae*	3	Siuslaw	1
OBGP-2017-154								
OBGP-2018-089								
OBGP-2017-054			*Rhinichthys*	26	*cataractae*	4	'Millicoma Dace'	1
OBGP-2017-014								
OBGP-2017-175								
OBGP-2018-242								
OBGP-2016-001					*evermanni*	2		
OBGP-2018-184								
OBGP-2017-330					*falcatus*	2		
OBGP-2018-100								
OBGP-2016-005					*osculus*	16	Black Lined	5
OBGP-2017-016								
OBGP-2017-017								
OBGP-2017-279								
OBGP-2017-290								
OBGP-2017-086							Closed Basin	2
OBGP-2018-019								
OBGP-2018-189							Foskett Spring	2
OBGP-2018-190								
OBGP-2017-166							Klamath	2
OBGP-2017-172								
OBGP-2018-057							Malheur Stream	2
OBGP-2018-061								
OBGP-2018-045							Stinking Lake Spring	1
OBGP-2017-158							Western	2
OBGP-2017-202								
OBGP-2018-069					*umatilla*	2		
OBGP-2018-122								
OBGP-2017-033			*Richardsonius*	14	*balteatus*	12	*balteatus*	5
OBGP-2017-197								
OBGP-2017-268								
OBGP-2017-359								
OBGP-2018-039								
OBGP-2018-005							*hydrophlox*	2
OBGP-2018-048								
OBGP-2016-006							Siuslaw	5
OBGP-2017-065								
OBGP-2017-278								
OBGP-2019-137								
OBGP-2019-149								
OBGP-2018-028					*egregius*	2		
OBGP-2018-094								
OBGP-2011-001			*Siphateles*	12	*alvordensis*	2		
OBGP-2019-212								
OBGP-2018-066					*bicolor*	8	*eurysoma*	1
OBGP-2017-177							Klamath	2
OBGP-2019-136								
OBGP-2018-007							*obesus*	2
OBGP-2018-026								
OBGP-2017-366							Silver Lake Basin	1
OBGP-2017-002							*thalassinus*	2
OBGP-2017-003								
OBGP-2009-001					*boraxobius*	2		
OBGP-2009-002								
OBGP-2019-223			*Tinca*	1	*tinca*	1		
OBGP-2017-055	Embiotocidae	3	*Cymatogaster*	1	*aggregata*	1		
OBGP-2019-013			*Embiotoca*	1	*lateralis*	1		
OBGP-2019-021			*Phanerodon*	1	*furcatus*	1		
OBGP-2019-224	Esocidae	1	*Esox*	1	*lucius*	1		
OBGP-2017-329	Fundulidae	2	*Fundulus*	2	*diaphanus*	2		
OBGP-2018-098								
OBGP-2017-136	Gasterosteidae	1	*Gasterosteus*	1	*aculeatus*	1		
OBGP-2017-185	Ictaluridae	10	*Ameiurus*	7	*natalis*	3		
OBGP-2017-305								
OBGP-2017-307								
OBGP-2017-018					*nebulosus*	4		
OBGP-2017-257								
OBGP-2018-046								
OBGP-2019-170								
OBGP-2018-292			*Ictalurus*	2	*punctatus*	2		
OBGP-2018-293								
OBGP-2019-194			*Noturus*	1	*gyrinus*	1		
OBGP-2019-032	Osmeridae	2	*Hypomesus*	1	*pretiosus*	1		
OBGP-2019-222			*Thaleichthys*	1	*pacificus*	1		
OBGP-2018-178	Oxudercidae	1	*Rhinogobius*	1	*brunneus*	1		
OBGP-2017-221	Percidae	4	*Perca*	3	*flavescens*	3		
OBGP-2017-242								
OBGP-2017-243								
OBGP-2018-268			*Sander*	1	*vitreus*	1		
OBGP-2017-383	Percopsidae	1	*Percopsis*	1	*transmontana*	1		
OBGP-2016-007	Petromyzontidae	13	*Entosphenus*	12	*lethophagus*	3		
OBGP-2017-024								
OBGP-2017-025								
OBGP-2017-027					*minimus*	1		
OBGP-2017-248					*similis*	3		
OBGP-2017-250								
OBGP-2019-143								
OBGP-2017-030					*tridentatus*	5		
OBGP-2017-325								
OBGP-2019-058								
OBGP-2019-167								
OBGP-2019-168								
OBGP-2019-010			*Lampetra*	1	*richardsoni*	1		
OBGP-2019-027	Pholidae	2	*Apodichthys*	1	*flavidus*	1		
OBGP-2019-029			*Pholis*	1	*ornata*	1		
OBGP-2019-026	Pleuronectidae	3	*Parophrys*	1	*vetulus*	1		
OBGP-2017-053			*Platichthys*	2	*stellatus*	2		
OBGP-2018-181								
OBGP-2017-342	Poeciliidae	2	*Gambusia*	2	*affinis*	2		
OBGP-2019-060								
OBGP-2018-065	Salmonidae	47	*Oncorhynchus*	27	*clarkii*	8	*henshawi*	3
OBGP-2018-068								
OBGP-2018-096								
OBGP-2017-149							*lewisi*	2
OBGP-2017-155								
OBGP-2017-332								
OBGP-2017-348								
OBGP-2018-240								
OBGP-2017-258					*keta*	2		
OBGP-2017-259								
OBGP-2017-271					*kisutch*	3		
OBGP-2018-038								
OBGP-2018-248								
OBGP-2017-015					*mykiss*	9	*gairdneri*	3
OBGP-2017-194								
OBGP-2017-198								
OBGP-2016-003							*irideus*	3
OBGP-2017-013								
OBGP-2017-061								
OBGP-2017-180							*newberrii*	3
OBGP-2017-255								
OBGP-2017-256								
OBGP-2019-190					*nerka*	2		
OBGP-2019-191								
OBGP-2017-052					*tshawytscha*	3		
OBGP-2017-356								
OBGP-2019-056								
OBGP-2017-062			*Prosopium*	2	*williamsoni*	2		
OBGP-2017-196								
OBGP-2017-167			*Salmo*	3	*trutta*	3		
OBGP-2017-227								
OBGP-2017-229								
OBGP-2011-002			*Salvelinus*	15	*confluentus*	6		
OBGP-2014-001								
OBGP-2017-223								
OBGP-2019-269								
OBGP-2019-272								
OBGP-2019-273								
OBGP-2017-168					*fontinalis*	6		
OBGP-2017-226								
OBGP-2017-228								
OBGP-2017-368								
OBGP-2018-020								
OBGP-2018-064								
OBGP-2019-211					*namaycush*	3		
OBGP-2019-227								
OBGP-2019-228								
OBGP-2019-030	Sebastidae	1	*Sebastes*	1	*caurinus*	1		
OBGP-2019-008	Umbridae	1	*Novumbra*	1	*hubbsi*	1		

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. TDN "spikes", where max(TDN) > 2 * mean(TDN) are in bold. Proportional relationship between mean within-family intraspecies and interspecies identity (id_prop) is also listed.

	Salmonidae			Cyprinidae			Catostomidae			Centrarchidae			Cottidae			Ictaluridae			Petromyzontidae		
	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop
rrnS	**21**	8.05	0.971	**13**	5.29	0.961	7	3.7	0.991	**9**	4.27	0.909	17	11	0.982	24	15.3	0.948	7	4.03	0.995
rrnL	**31**	7.73	0.97	18	9.09	0.926	**15**	5	0.984	**16**	6.41	0.904	**23**	11.3	0.975	30	15.2	0.944	**8**	3.92	0.997
nad1	30	23.5	0.882	**18**	8.95	0.849	24	14.5	0.943	18	9.83	0.827	32	21.9	0.936	48	33	0.868	24	14.5	0.988
nad2	47	35	0.871	19	10.6	0.826	20	12.3	0.933	17	11	0.795	34	26.3	0.925	44	33.8	0.861	32	19.8	0.984
cox1	29	17.6	0.908	12	6.3	0.892	**16**	7.73	0.97	15	7.79	0.853	29	19.3	0.956	41	29.5	0.883	19	11.7	0.991
cox2	**2**	0.536	0.932	13	8.36	0.906	11	7.71	0.974	8	4.5	0.869	24	18.3	0.964	27	17.4	0.911	20	14.4	0.987
cox3	**5**	2.28	0.912	13	7.88	0.89	17	11.2	0.966	12	7.84	0.855	30	17.9	0.955	39	27.9	0.896	18	13	0.989
nad4	35	25.7	0.885	**21**	8	0.846	22	15.5	0.948	18	10.7	0.809	18	10.3	0.937	44	34.4	0.86	26	16.1	0.987
nad5	43	23.3	0.889	19	9.92	0.832	22	12	0.953	**19**	8.94	0.812	37	21.3	0.939	51	34	0.862	25	16.1	0.985
cytb	33	21.3	0.894	12	7.58	0.864	18	12.4	0.947	18	10.5	0.826	27	19.5	0.951	41	31.9	0.873	18	12.5	0.989
dloop	**83**	15	0.852	**83**	28.4	0.82	13	7.54	0.965	**37**	8.93	0.834	**107**	41.7	0.938	46	26.5	0.885	**46**	15.6	0.989

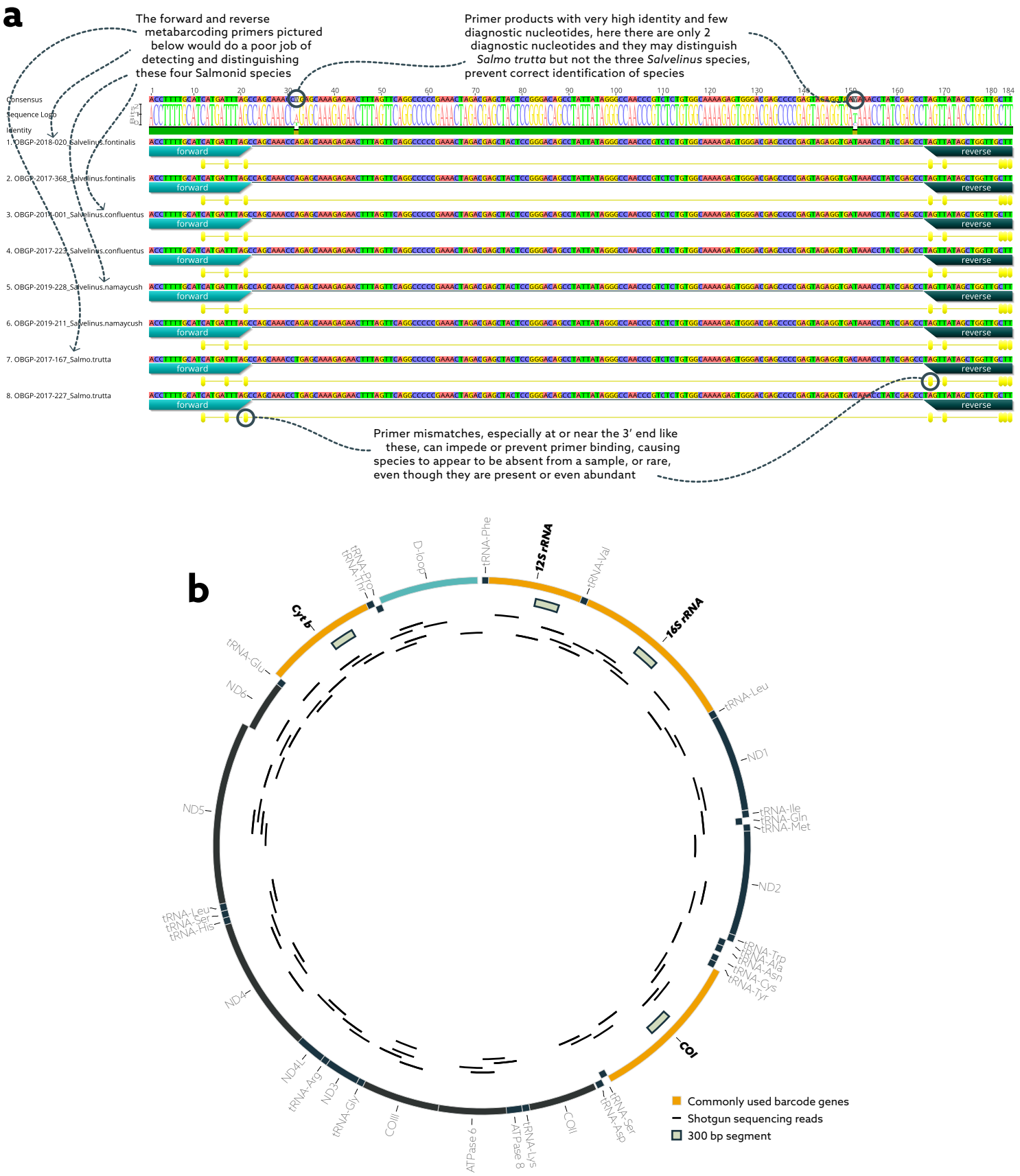

FIGURE 1
Primer Issues and Mapping Shotgun Reads. a. The Problems with Primers. A sequence alignment of PCR products produced in silico illustrates some of the issues with primers. b. Mapping Shotgun Reads. Samples that are shotgun-sequenced will produce reads from the entire mitogenome. This graphic 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 (300 bp or less) fragment of genetic information from the mitogenome.

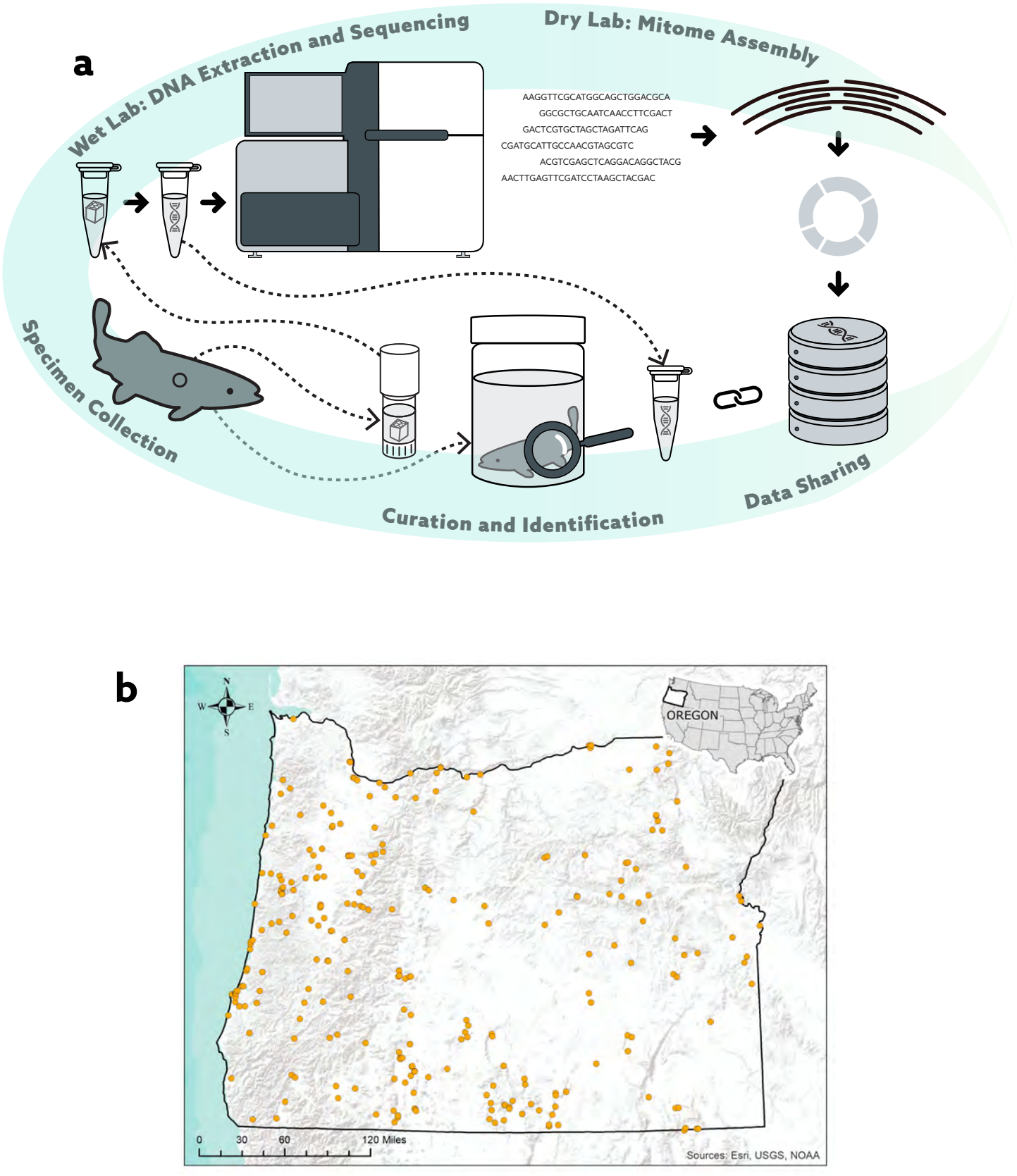

FIGURE 2
Sequence Database Creation. a. Pipeline: Reference Sequence Database Development. Specimens are collected and euthanized in MS-222. Tissues are sampled and preserved in 95% EtOH prior to immersing full-body vouchers in a 10% Formalin solution. DNA is extracted from subsampled tissues and prepped for shotgun sequencing. The resulting sequencing reads are assembled using a variety of bioinformatics pipelines. Tissues, vouchers specimens, and DNA extracts are accessioned into a natural history collection and linked to sequence data stored on GenBank. b. Map of Study Area. An interactive map can be viewed here.

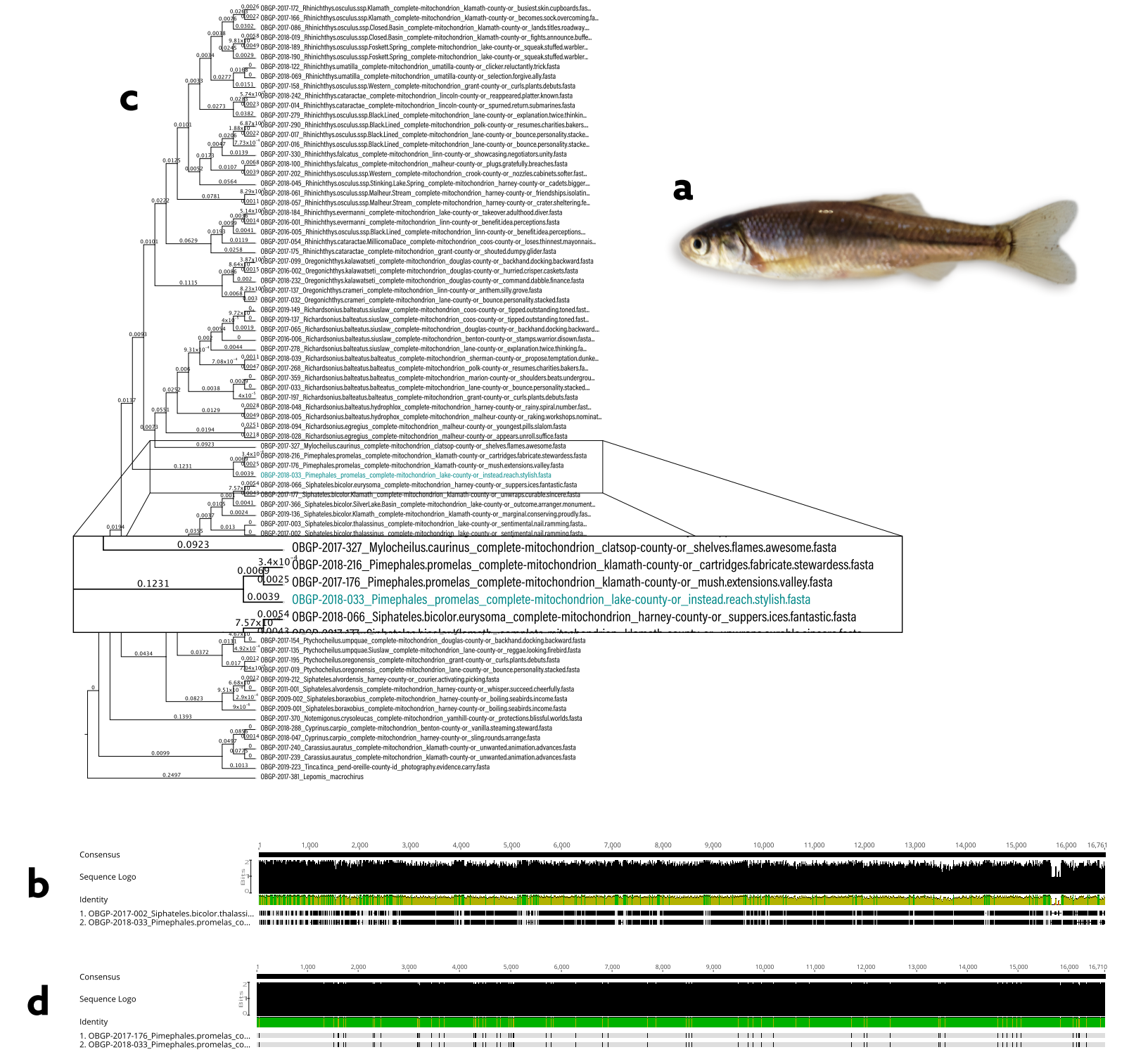

FIGURE 3
Taxonomic Reassignment from NJ Clustering Analyses: Pimephales promelas Example Specimen. OBGP-2018-033 (a) was originally identified provisionally as Tui Chub (Siphateles bicolor thalassinus). An alignment of its full mitogenome with previously identified S. bicolor (b) demonstrated only 83.133% identity with the specimen, indicating that its provisional assignation was incorrect. An NJ clustering analysis of Cyprinids (c) grouped OBGP-2018-033 tightly with Fathead Minnow (Pimephales promelas) suggesting that this was the correct species designation. A full mitogenome alignment with previously identified P. promelas specimen OBGP-2017-176 (d) supported this species reassignment and shared 99.569% identity with the specimen. Note: Pimephales promelas is a non-native species that has been expanding its range in Oregon. Although the species was being targeted where this specimen was collected, field conditions and fish life stage can challenge species determinations. Zoomed-in sequence alignments (b and d) are available in Appendix S4 Figures S6 and S7.

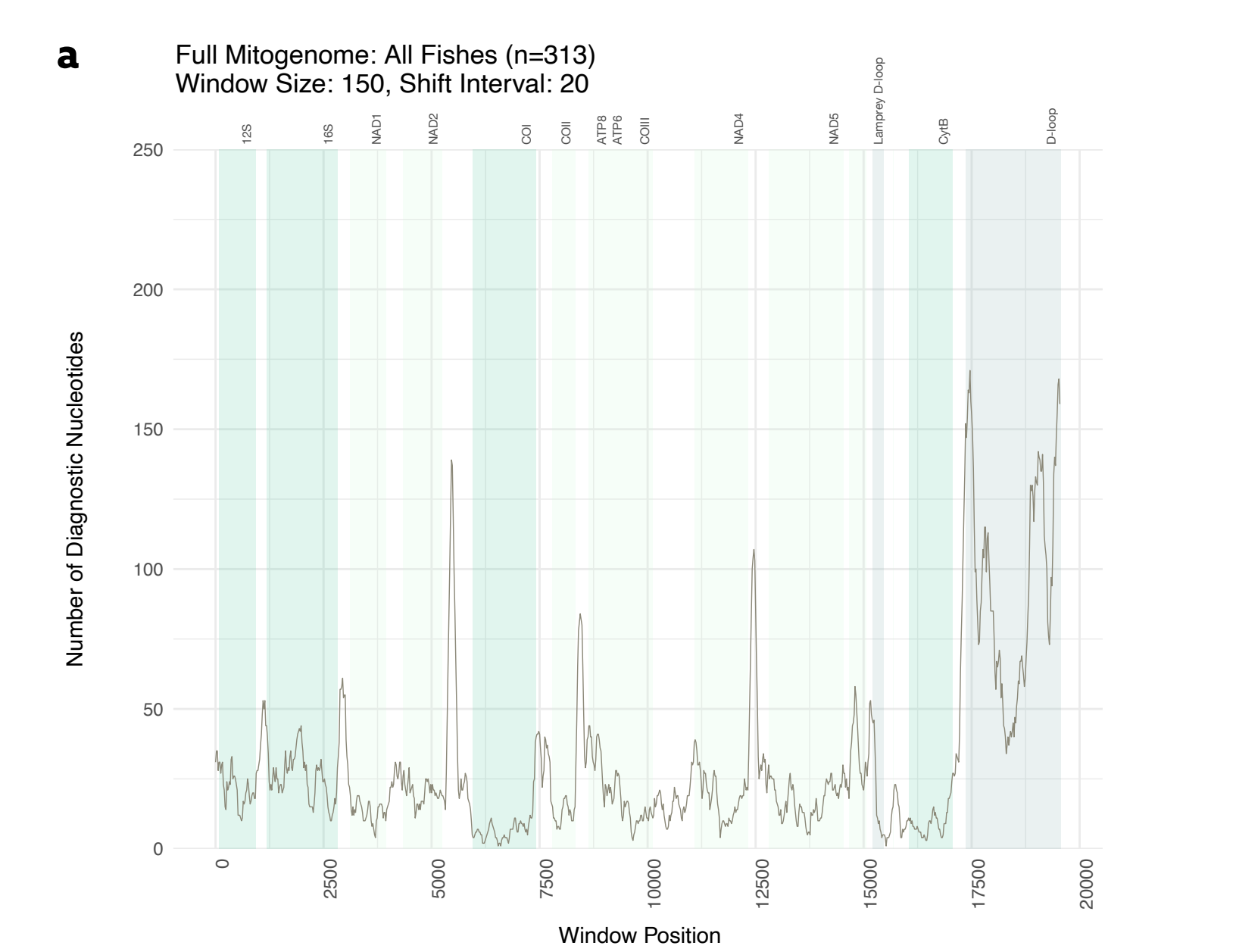

FIGURE 4
Sliding Window Analysis: A window 150 bp in length is placed along the length of an alignment of 313 individuals and shifted at 20 bp intervals. At each interval, the number of diagnostic nucleotides—where a base is shared within a species but is either different or unaligned with other species—within the 150 bp window is counted. a. Full mitogenome. Gene regions and the D-loop are shaded in blue. Areas of highest variability are in noncoding regions. b. Genes. These plots zoom in on a subset of individual genes within the mitogenome to focus on the number of diagnostic nucleotides within 8 barcode regions. Means (TDN/$w_{150}i_{20}$): COI, 7.446; CytB, 9.549; NAD1, 13.381; NAD5, 17.161; NAD4, 18.710; NAD2, 20.065; 12S, 20.977; 16S, 25.976.

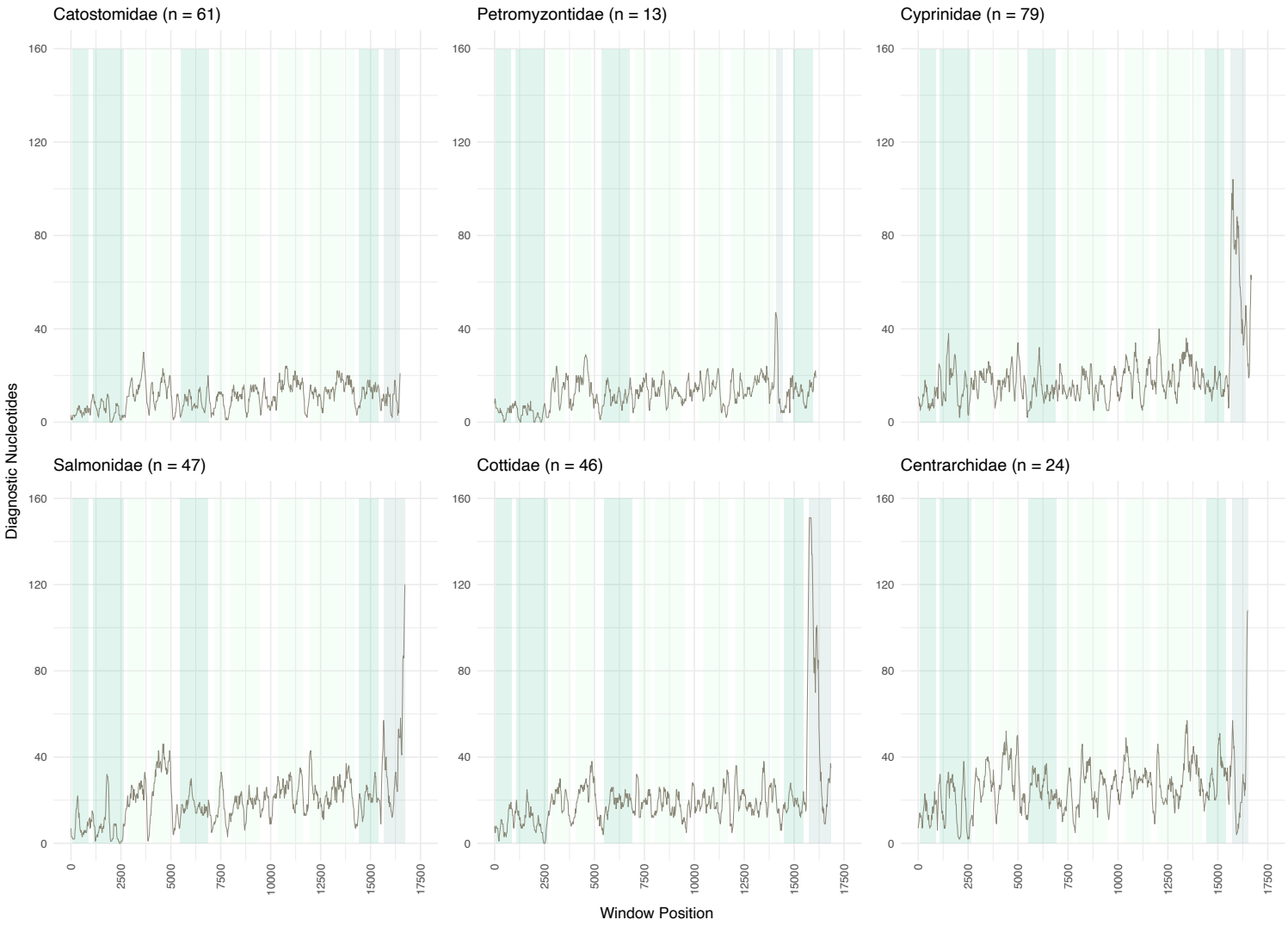

FIGURE 5
Sliding window analysis, by family, full mitogenome. The full alignment of 313 specimens is separated into individual families and a window 150 bp in length was placed along the length of an alignment of each family alignment and moved at 20 bp intervals. At each position, the number of diagnostic 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. Note: Petromyzontidae mitogenome is structured with its control region upstream of the CytB gene. Means (TDN/$w_{150}i_{20}$): Catostomidae, 10.656; Petromyzontidae, 12.104; Cyprinidae, 19.316; Salmonidae, 20.209; Cottidae, 21.024; Centrarchidae, 26.082.

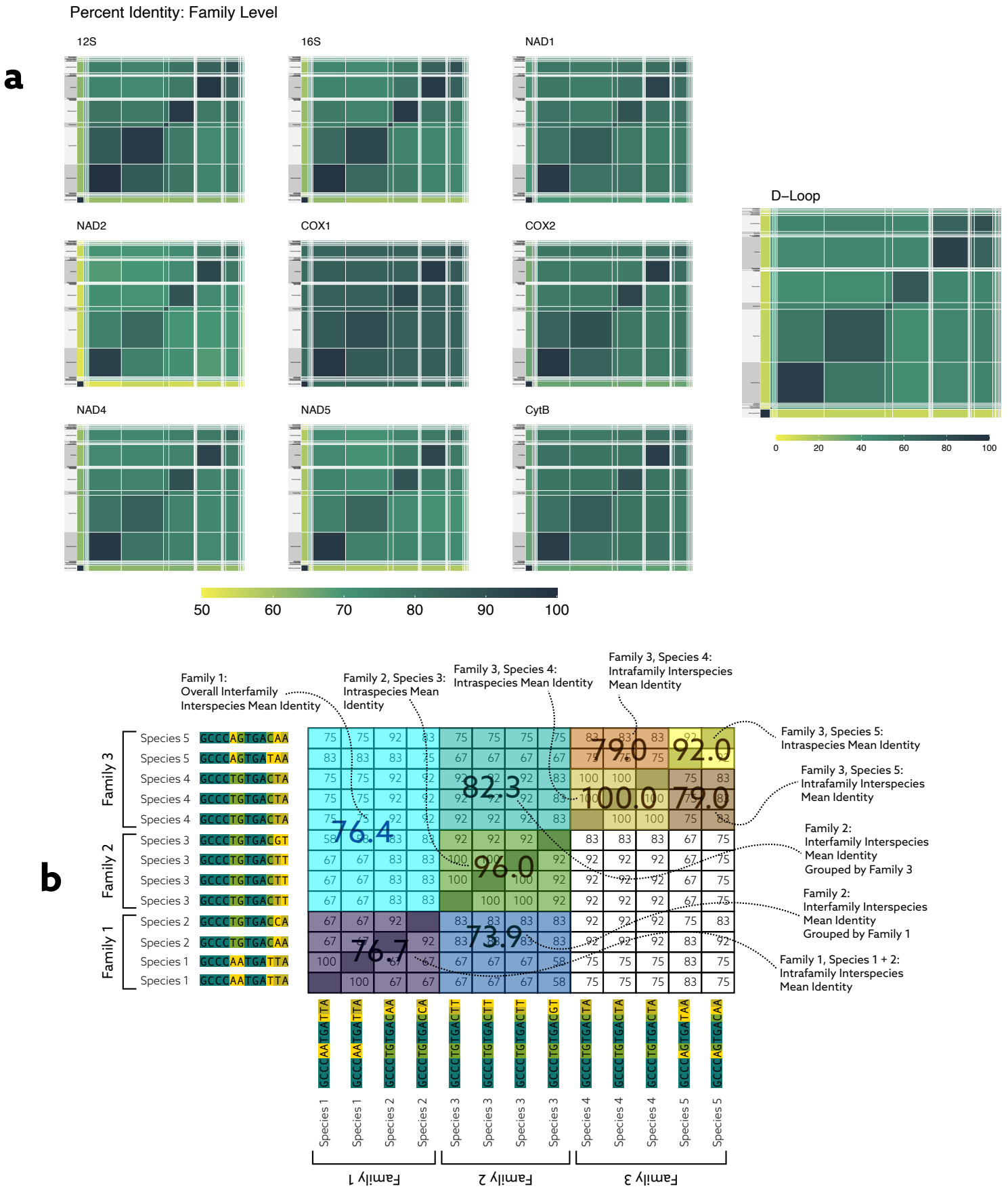

FIGURE 6

a. Percent Identity Heatmaps: Family level. This is a graphical representation of interfamily mitogenome percent identity. Yellow colors 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. Raw numbers can be referenced in Appendix S5. b. Anatomy 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 to itself—remaining blank. Data can be grouped, and mean percent identities can be calculated as depicted here.

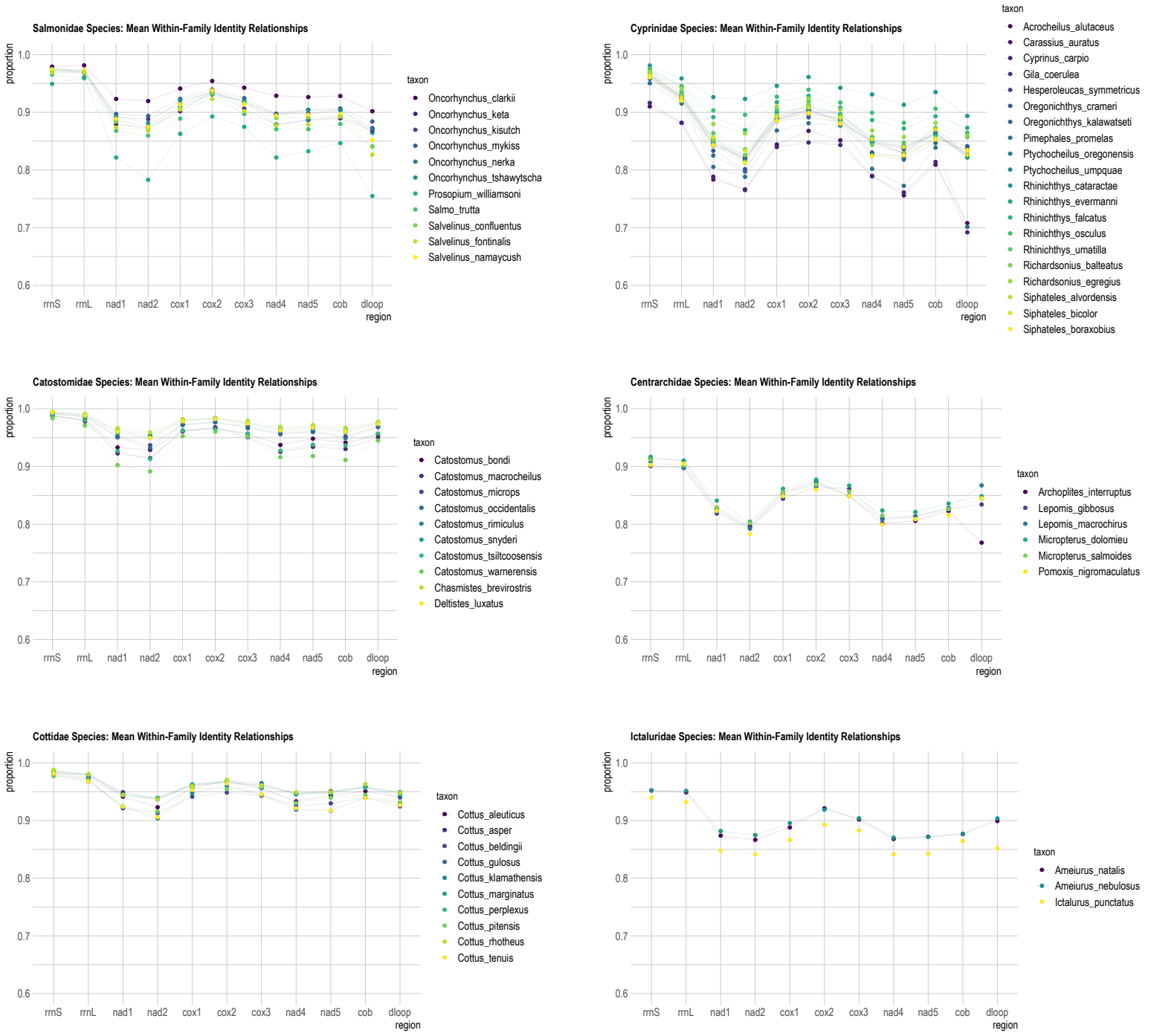

FIGURE 7
Within-Family Interspecies Relationships in Mean Percent Identity: Parallel Coordinate Plots. These plots illustrate the relative genetic distance among versus within species within the 5 plotted families. The mean percent identity is calculated among individuals of the same species and the mean percent identity is then calculated between that species and all other species within a given family. The proportional relationship between these two means is plotted here. Species in some families exhibit greater genetic distance—Cyprinidae and Centrarchidae—between families than others—Cottidae and Catostomidae. See Appendix S4 Table S1 for proportional identity figures.

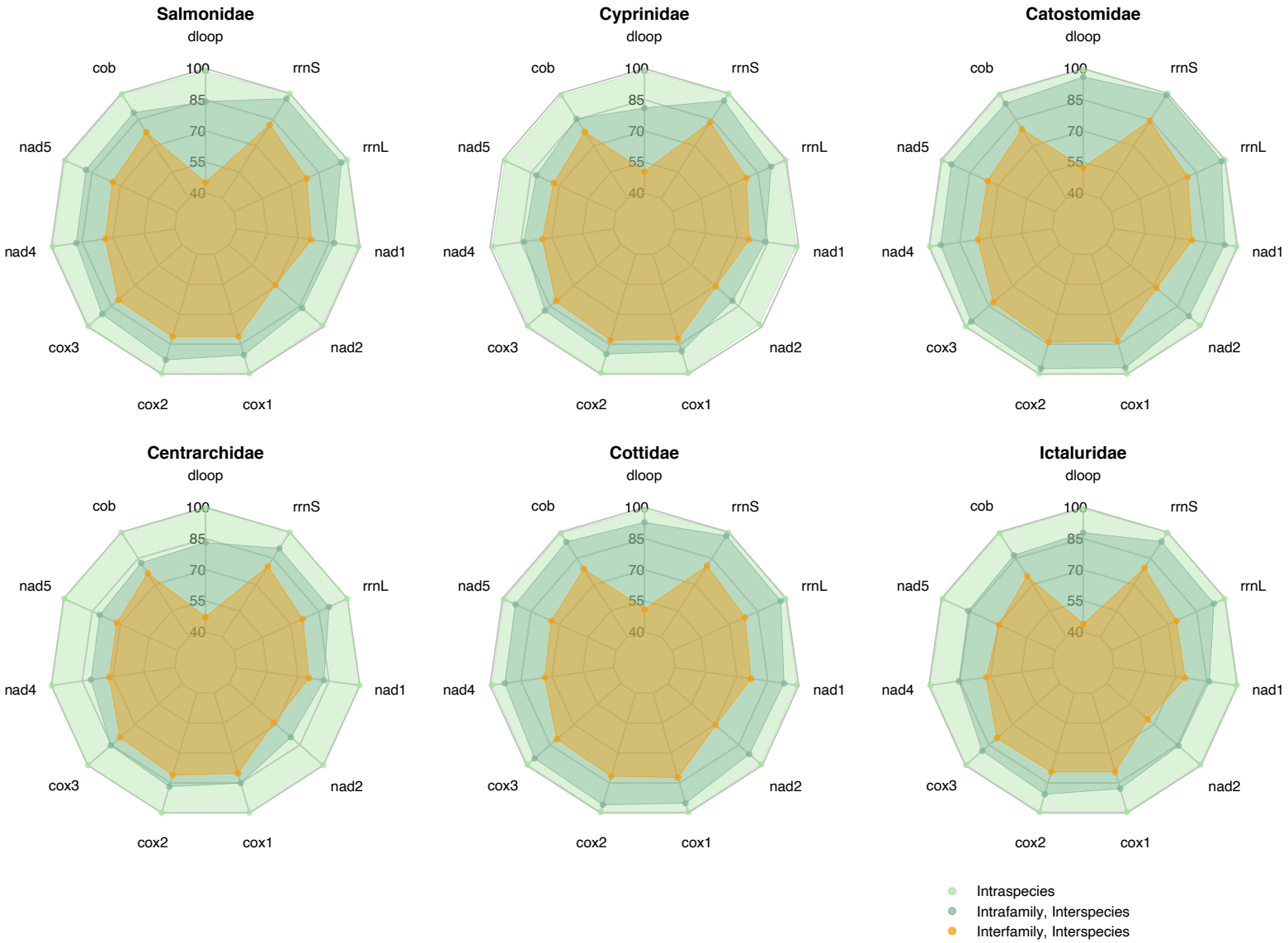

FIGURE 8
Intraspecies, Interspecies within Family, and Interspecies Among Family Percent Identity in a Subset of Families. Mean intraspecies percent identities are calculated and plotted for each gene (pale green), along with interspecies intrafamily percent identities (dark green), and interspecies interfamily percent identities (orange). Interfamily calculations were computed between all families, not just the families depicted. The axes on the radar charts span from 40% 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 was not included in plots due to highly skewed interspecies/interfamily identity.

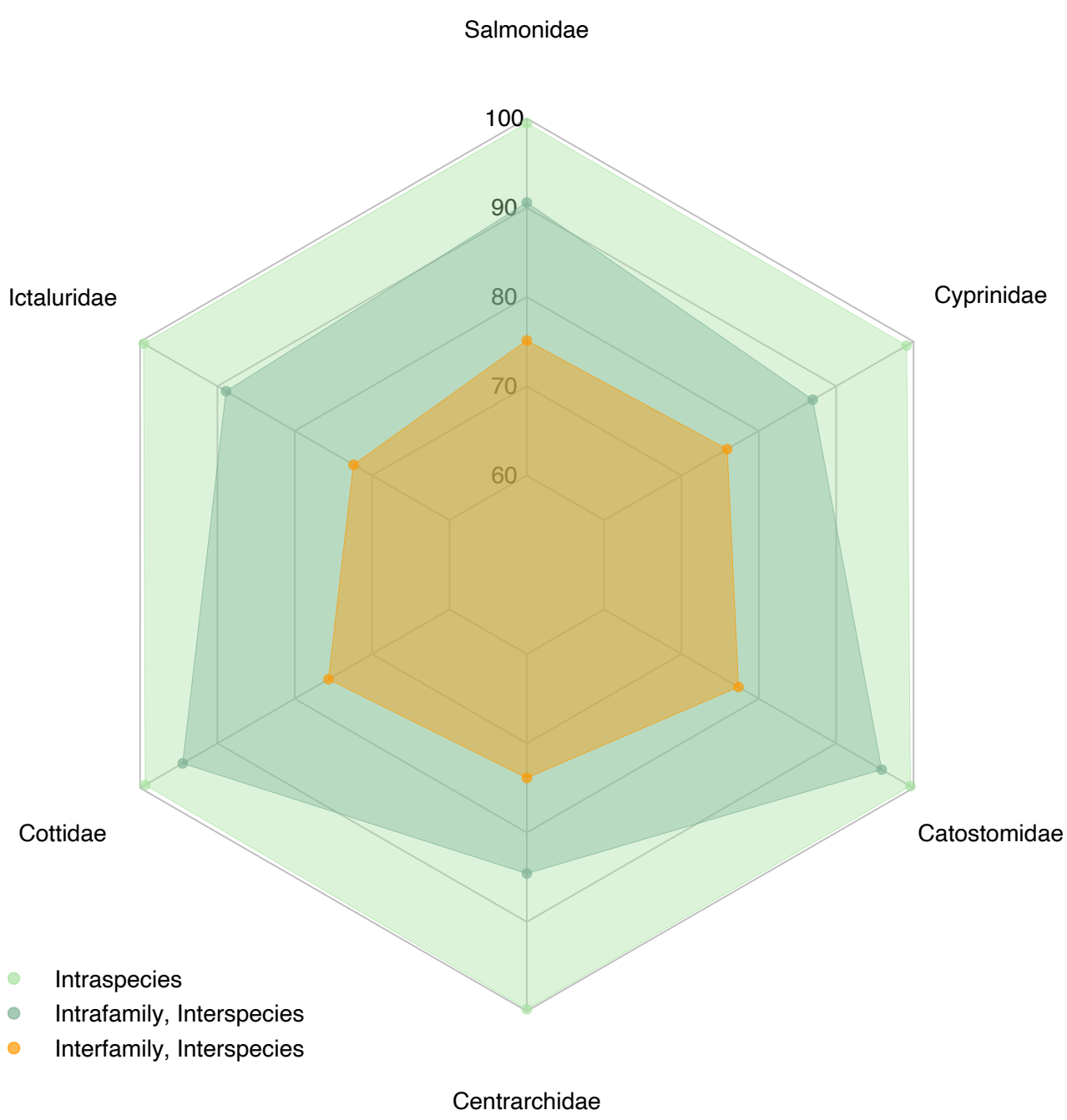

FIGURE 9
Whole Mitogenome Intraspecies, Interspecies within Family, and Interspecies Among Family Percent Identity in a Subset of Families. Mean intraspecies percent identities are calculated and plotted (pale green), along with interspecies intrafamily percent identities (dark green), and and interspecies interfamily percent identities (orange). Interfamily calculations were computed between all families, not just the families depicted. The axes on the radar charts span from 60% at the innermost ring to 100% identity at the outermost ring. Petromyzontidae was not included in plots due to highly skewed interspecies/interfamily identity.