Enabling robust environmental DNA assay design with "unikseq" for the identification of taxon-specific regions within whole mitochondrial genomes

Michael J. Allison, René L. Warren, M. Louie Lopez, Neha Acharya-Patel, Jacob J. Imbery, Lauren Coombe, Cecilia L. Yang, Inanc Birol, & Caren C. Helbing

2023

Allison, M. J., Warren, R. L., López, M. L., Acharya-Patel, N., Imbery, J. J., Coombe, L., Yang, C. L., Birol, İ., & Helbing, C. C. (2023). Enabling robust environmental DNA assay design with "unikseq" for the identification of taxon-specific regions within whole mitochondrial genomes. *Environmental DNA*, *5*(5), 1032–1047. https://doi.org/10.1002/edn3.438

Received: 23 August 2022 | Revised: 17 May 2023 | Accepted: 30 May 2023

DOI: 10.1002/edn3.438

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences

WILEY

SPECIAL ISSUE ARTICLE

Enabling robust environmental DNA assay design with "unikseq" for the identification of taxon-specific regions within whole mitochondrial genomes

Michael J. Allison[1] | René L. Warren[2] | M. Louie Lopez[1] | Neha Acharya-Patel[1] | Jacob J. Imbery[1] | Lauren Coombe[2] | Cecilia L. Yang[2] | Inanc Birol[2,3] | Caren C. Helbing[1]

[1]Department of Biochemistry & Microbiology, University of Victoria, Victoria, British Columbia, Canada

[2]Canada's Michael Smith Genome Sciences Centre at BC Cancer, Vancouver, British Columbia, Canada

[3]Department of Pathology & Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence
Caren C. Helbing, Department of Biochemistry & Microbiology, University of Victoria, Victoria, BC V8P 5C2, Canada.
Email: chelbing@uvic.ca

Funding information
Genome British Columbia, Grant/Award Number: 312ITD; Genome Canada; Genome Québec; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: CGS-D3 and CGS-M

Abstract

Environmental DNA (eDNA) is revolutionizing species monitoring in nature. At the heart of any eDNA approach is the reliance upon sufficient DNA sequence information to satisfy the demands of eDNA assay specificity and sensitivity. The most common source of this information has been restricted to short barcoding regions of the mitochondrial genome (mitogenome) and marker genes. The use of these limited regions for assay design has often resulted in substantial trade-offs in assay performance. With increased accessibility of full mitogenome assemblies, the potential for designing more robust eDNA assays is considerably enhanced. However, this also poses a new challenge to effectively identify suitable regions for assay design using considerably larger sequences. We present *unikseq*, a utility that uses words of length *k* (*k*-mers) to identify unique regions in a reference sequence relative to tolerated (in-group) and not-tolerated (outgroup or non-target) sequence sets, quickly and with low memory that can yield highly specific assays. We illustrate its application within an assay development workflow through use-case examples for the design and validation of four quantitative real-time polymerase chain reaction (qPCR)-based assays selective for American bullfrog (*Rana [Lithobates] catesbeiana*), Burbot (*Lota lota*), Lake trout (*Salvelinus namaycush*), and Quillback rockfish (*Sebastes maliger*). The chosen target species vary in range, habitat, and degree of relatedness to their sympatric species that, consequently, impact eDNA assay design difficulty. We demonstrate the effectiveness of *unikseq* through assay validation and characterization using DNA from voucher specimens, synthetic DNA, and, where possible, field samples, to verify the specificity and sensitivity of the newly designed assays. By facilitating whole mitogenome sequence comparison, the creation of high-performing eDNA assays is substantially enhanced. Having several adjustable parameters for specifying user requirements within *unikseq*, this approach can facilitate the identification of suitable regions for a broad range of applications requiring nucleotide sequence comparisons.

KEYWORDS
eDNA, quantitative real-time polymerase chain reaction, robust assay design, specificity, unique mitogenome regions

1 | INTRODUCTION

Many aspects of environmental monitoring depend on timely, reliable, and affordable methods for the effective detection of the desired taxa. Environmental DNA (eDNA) detection is one such method that has been the subject of intensive research and development over the past decade (Ficetola et al., 2008; Goldberg et al., 2016; Matthias et al., 2021; Veldhoen et al., 2016; Yates et al., 2021). eDNA refers to the genetic material present in any environmental sample such as water, soil, or air. There are many different approaches for eDNA detection, which utilize a variety of assay types, but all rely on the availability of relevant genetic sequences for assay creation (Langlois et al., 2021).

The most frequently used eDNA assays rely upon quantitative real-time polymerase chain reaction (qPCR)-based methods employing a primer pair and a fluorescently labeled probe that specifically amplify DNA from target taxa to infer species presence. The quality of an assay is determined by its ability to exclusively amplify the target DNA (selectivity) and the efficiency at which it does so (sensitivity). For most targeted eDNA applications, an adequately selective assay will demonstrate no amplification of confounding taxa DNA (Thalinger et al., 2021). Confounding or off-target taxa refer to species whose genetic similarity and/or potential for co-occurrence may lead to inadvertent amplification of genetic material and confound eDNA survey results and thus need to be considered in assay design. Sensitivity measures are difficult to set thresholds for, as each context may vary greatly in their requirements, but it is important to achieve as sensitive an assay as possible to have confidence in negative results (Cristescu & Hebert, 2018; Langlois et al., 2021; Matthias et al., 2021; Thalinger et al., 2021).

To satisfy these requirements, one needs aligned DNA sequences that account for the breadth of variation in target taxa, closely related non-target taxa, and other factors in the study region, especially when the target and non-target taxa are closely related and share high sequence identity (Langlois et al., 2021). DNA from multiple target specimens should be used to ensure that haplotype variation does not impact assay performance. Both the methodologies and resources needed for high-quality eDNA assay creation and analysis have improved in recent years, thanks in large part to technological advancements in genomics and bioinformatics (Cristescu, 2019; Williams et al., 2020). Many aspects of assay design can benefit from emerging technological approaches. For example, machine learning has been employed to accurately predict quantitative real-time polymerase chain reaction (qPCR) cross-amplification in barcode regions (Kronenberger et al., 2022), and *k*-mer-based bioinformatics approaches have been used to predict phenotypes such as antibiotic resistance from sequencing data (Aun et al., 2018; Nurmukanova et al., 2022; Simpson et al., 2009; Warren et al., 2019, 2015). One of the most impactful developments has been an increase in sequencing data availability from a large range of taxa that constitute the targets of monitoring efforts as well as confounding taxa. To address these challenges while taking advantage of the recent progress in technology and genetic resources, we have developed *unikseq*, a tool for parsing nucleotide sequence information to identify the optimally selective regions of large sequence datasets for eDNA assay design.

Environmental DNA assays are most often developed using mitochondrial markers, as these regions have suitable genetic variation for species-level differentiation and the number of mitochondrial genome (mitogenome) copies typically far outnumbers nuclear genome copies per cell, thereby improving detection probability in environmental samples (Bylemans et al., 2018; Jensen et al., 2021). Currently, most publicly available animal mitochondrial DNA sequences encode portions of genes used for barcoding purposes and are shorter than 1000 base pairs (bp). The need to consider confounding taxa in assay design often severely limits designs to portions of common barcode genes for which greater taxonomic breadth of sequences exist (Langlois et al., 2021).

Commonly targeted mitochondrial sequences include portions of the *cytochrome c oxidase I* (*mt-co1*), *cytochrome b* (*mt-cyb*), and *ribosomal RNA subunit 2* (*mt-rnr2*) genes. Currently, the most abundant barcoding sequence resources for metazoans are from the *mt-co1* gene, although this gene represents only ~3%–4% of the total nucleotides in an average mitogenome (Leray et al., 2022). While sensitive and specific eDNA assays may be produced using small mitogenome regions, restricting assay design to these regions limits design options and may even make high-quality design impossible (Langlois et al., 2021). Improvements in sequencing technologies and specific large-scale initiatives to increase the number of available complete mitogenome sequences have opened the possibility to design eDNA assays using the aligned sequences of all relevant species' entire mitogenomes (https://www.itrackdna.ca/; https://gen-fish.ca/; http://earthbiogenome.ca/; https://bioscancanada.org/).

While this greatly improves the ability to develop selective assays without sacrificing sensitivity, it presents a new problem. Most animal mitogenomes are between 16,000 and 20,000 bp long, and not all regions are amenable to robust assay design, either in terms of shared identity between sequences or in base pair composition meeting requirements for efficient assays (MacDonald & Sarre, 2017; Wilcox et al., 2013). Navigating entire mitogenomes with existing alignment and analysis tools (i.e., Base-by-base, Bioedit, Geneious Prime, MEGA) is very time-consuming, requires considerable manual analysis, and consequently is prone to human error (Hall, 1999; Tamura et al., 2021; Tu et al., 2018).

Herein we present *unikseq*, a robust tool that allows the non-biased parsing of full mitogenome alignments to target the most appropriate regions more effectively for desired eDNA assay design specificity requirements. *Unikseq* works by comparing a reference sequence to other sequences desired as targets for qPCR amplification ("ingroup") and sequences that are not desired as targets ("outgroup") to produce sequence fragments that are unique to the ingroup sequence set, relative to outgroup or non-target sequences, for targeting with eDNA assays. We present four assays herein that were developed using *unikseq* in combination with an established eDNA assay development workflow which includes rigorous testing using isolated specimen tissue DNA and synthetic DNA to ensure that they are both sensitive and selective enough for the demands of eDNA usage (Figure 1). These four assays were designed to target species that exhibit a range of design challenges, including large numbers of potentially co-occurring species and highly related co-occurring species in each's respective environmental niche. To demonstrate the scalability of *unikseq*, we also show its use on large genome sets using SARS-CoV-2 as an example. We hope that providing this tool to the research community will help improve specificity of species-targeted qPCR-based eDNA assays by harnessing the full potential of mitogenomes.

2 | MATERIALS AND METHODS

2.1 | Sequence collection and alignment

Full mitochondrial sequences were acquired for target species and sympatric species, plus dog, cat, and human as their DNA are prevalent in both field and laboratory workspaces. Most sequences used in the present study were available from the National Center for Biotechnology Information (NCBI) GenBank (https://www.ncbi.nlm.nih.gov). For all publicly available mitogenome sequences, FASTA files were collected using an Entrez Direct EFetch script (Kans, 2013). Amphibian, mammalian, and teleost accession numbers were assembled manually, and this list was referenced by the script to acquire corresponding sequences from the NCBI GenBank database. The sequences for all *Sebastes* species excluding *Sebastes aleutianus* and *Sebastes nigrocinctus* were graciously supplied by Dr. Gregory Owens (Kolora et al., 2021). Sequences were then manually sorted into taxonomic class (amphibian, mammal) or infraclass (teleost) and separate alignments for designing each assay were made that included the target species and any potentially co-occurring species from the same taxonomic class or infraclass plus human, dog, and cat to account for the possibility of contamination in the

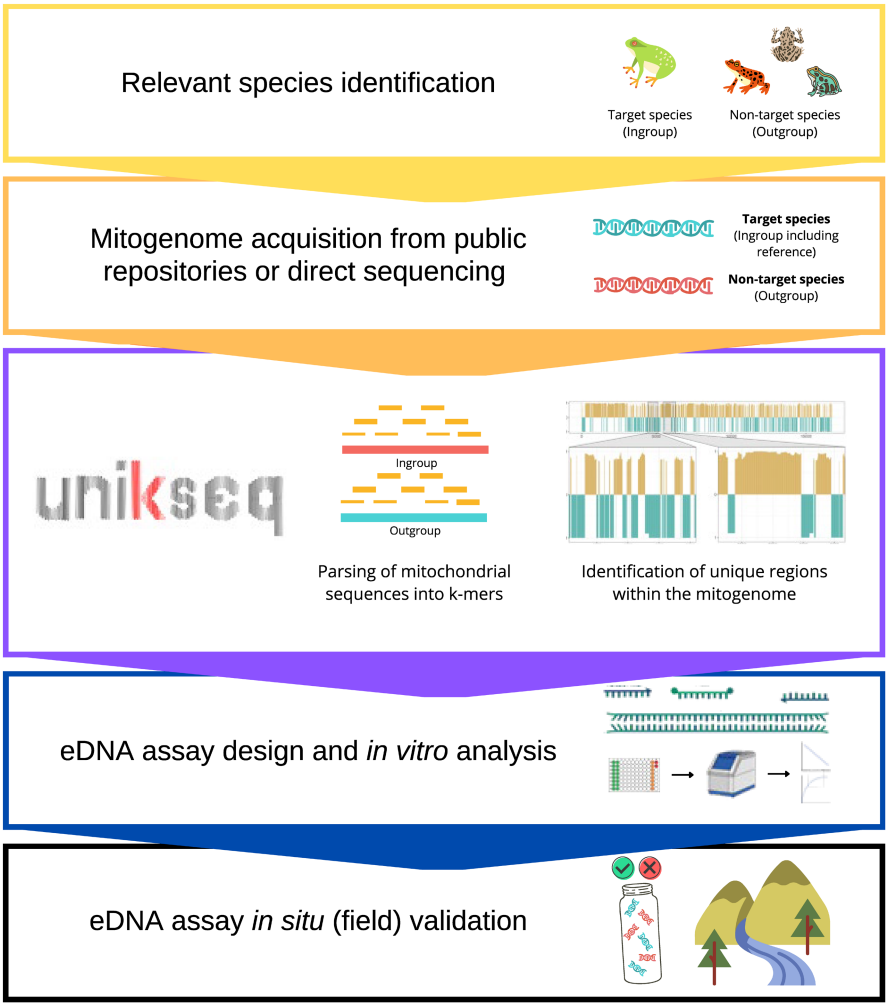

FIGURE 1 The *unikseq*-based assay development workflow. Careful consideration should be put into each step of the assay development pipeline, beginning with identification of any possible factors that may impact the downstream analyses. Special care should also be taken in assessing the quality, source, and orientation of sequences to be used. The *unikseq* script can be used to rapidly parse large amounts of genetic information to deliver sequence meeting a range of specificity requirements. The assay design and validation stages can then proceed without sacrifice of speed or assay quality.

eDNA analysis workflow. The full lists of all mitogenomes used in the present study is shown in Table S1 in Appendix S1. For each target assay, sequence alignments from target and confounding species were created with MAFFT version 7.310 using default settings (Katoh et al., 2019) and visualized using Geneious Prime (version 2202.1.1; Biomatters Ltd). Alignment files were inspected visually to ensure completeness and are available as File S1 in Appendix S1. While only raw sequence files are needed for *unikseq* queries, these alignments are important for the subsequent assay design step to verify specificity of the primer and probe sequences (Figure 1). To help visualize relationships between species for assay development, Neighbor-Joining phylogenetic trees were assembled in Geneious Prime using the Tamura-Nei model, with specific construction approaches detailed in Tamura and Nei (1993).

2.2 | *Unikseq* code

The algorithm begins by first parsing FASTA sequences supplied by the user as "outgroup" (-o option) and "ingroup" (-i option) and extracting every word of length *k* (*k*-mers, -k option) and their reverse complement and storing each in respective two-dimensional hash data structures, keeping track of the *k*-mer occurrence in each FASTA entry for either set. The input in/outgroups are very flexible and can include contiguous/fragmented genome sequences with inconsistent start coordinates, unordered/unoriented contigs—but it is best used with complete mitogenome sequences. This is especially important for outgroup sequence sets, as absence of *k*-mers due to incomplete sequences may result in the identification of "false" unique sequences in the reference sequence under scrutiny. We also note that the in- and outgroup sequences do not need to start at the same position, nor be represented on the same strand since *unikseq* is *k*-mer-based, therefore no specific DNA sequence formatting is required (e.g., no need to adjust the sequence start for mitogenomes) other than supplying FASTA-formatted files, which is the only file requirement. Also, unknown or ambiguous nucleic acid bases (i.e., characters other than A, C, G, T, U) are ignored, and *k*-mers containing them are not considered for analysis. Draft mitogenome assemblies with unresolved sequences (i.e., gaps) may be used, as long as the outgroup sequence set has a certain level of sequence redundancy (i.e., more than one mitogenome supplied as input). The search for stretches of potentially unique sequences in a reference sequence (-r option) relative to outgroup sequences begins with the 5′–3′ extraction of a forward-strand *k*-mer and querying of the above two hash data structures for (1) exclusion of the *k*-mer in the outgroup and (2) inclusion of the *k*-mer in the ingroup, moving the *k*-mer frame over base by base until the entire FASTA sequence is read, and all reference *k*-mers have been interrogated. When a sequence *k*-mer is not found in the outgroup hash data structure, it is deemed unique, and its position and coverage in the outgroup set are tracked. A new unique sequence stretch is initialized if the *k*-mer demarks the beginning of a new unique region, or the last unique base is concatenated onto the growing unique sequence stretch on the 3′-end if a unique *k*-mer follows a *k*-mer previously identified as unique. A record of the *k*-mer's presence in the ingroup hash data structure is also kept for the purpose of plotting and evaluating the overall conservation of the unique sequence stretch in the ingroup (see the -p option described below). This process is repeated until the end of the FASTA reference sequence is reached or until a condition is no longer met, including when the next *k*-mer is found to be non-unique in the outgroup set. To facilitate the detection of unique regions that may be interspersed with non-unique *k*-mers, a leniency parameter (-l option) is used to control the tolerated number of consecutive *k*-mers found in the outgroup. The unique sequence stretch extension is stopped when the number of consecutive reference *k*-mers found in the outgroup set has exceeded that minimum -l threshold. A grace parameter (-m option) can also be used to adjust the tolerance, or uniqueness factor, in the outgroup. This is particularly useful when searching for unique regions in sets of very similar sequences, for instance, as -m describes the maximum proportion of outgroup sequences in the set that are tolerated before the reference *k*-mer is considered unique (default set to 0, indicating no tolerance). The -m and -l parameters work together in controlling the stringency of *unikseq*. When the unique sequence can no longer be extended, it will be written to a FASTA file only when: (1) it is longer than a threshold length (-s option), (2) its constituent *k*-mers have been identified, on average proportion, in at least (-p option) % of ingroup sequences and, overall, (3) those *k*-mers are at least a defined % unique (-u option), as specified by -l and -m. *Unikseq* also outputs a tab-delimited value (tsv) file that tracks, at each coordinate relative to the reference sequence, the proportion of corresponding *k*-mers in the outgroup and ingroup sets, and the data used to generate the butterfly plots presented herein. Instructions and code for generating the butterfly plots in R are available from the GitHub repository at the URL below. The intended use-case of *unikseq* is for identification of unique sequences in mitogenome size or shorter sequences, and thus has not been tested on larger genomes. *Unikseq* is developed in PERL and runs on any system where PERL is installed, requiring no additional libraries, making it cross-platform and versatile. It was tested on both linux servers and consumer mac OS (Catalina) systems. It is distributed under GPLv3 license and available freely at https://github.com/bcgsc/unikseq.

2.3 | Application and validation of *unikseq* outputs

We applied the *unikseq* approach to several target taxa to find assays that displayed both selectivity for the target taxon and sensitivity. Four target taxa at the species level (an amphibian "am" and three teleosts "te"; Table 1) were chosen to present in this work, representing a variety of common application contexts: (1) The American bullfrog (*Rana [Lithobates] catesbeiana*) is an example of a cosmopolitan invasive amphibian species with potential study regions

TABLE 1 Target species and general assay information with details regarding the *unikseq* parameters employed.

Common name	Scientific name	Species code	Assay name	*Unikseq* parameters (default setting)					# of output sequences	Range of output sequence lengths (bp)	Gene used for assay
				k (25)	s (100)	p (25)	l (1)	u (90)	m (0)		
American bullfrog	*Rana [Lithobates] catesbeiana*	am-LICA	eLICA5	25	100	25	1	90	0	44	100–880	*mt-nd5*
Burbot	*Lota lota*	te-LOLO	eLOLO4	25	100	25	1	90	0	40	122–1296	*mt-nd4*
Lake trout	*Salvelinus namaycush*	te-SANA	eSANA1	25	100	25	1	90	25	14	134–796	*mt-nd2*
Quillback rockfish	*Sebastes maliger*	te-SEMA	eSEMA3	25	100	25	3	70	20	6	100–195	*mt-d-loop*

Note: Default values for *unikseq* parameters are indicated in brackets.

throughout the world; (2) Burbot (*Lota lota*) is a freshwater gadiform fish with few closely related co-occurring species (Han et al., 2021); (3) Lake trout (*Salvelinus namaycush*) is a freshwater salmonid with several closely related sympatric species; and (4) Quillback rockfish (*Sebastes maliger*) is a marine species with several dozen closely related co-occurring species. These four species were assigned the species codes am-LICA, te-LOLO, te-SANA, and te-SEMA, respectively (Table 1). Each target assay development is examined in detail by assay type below.

In all use-cases, we collated FASTA sequence sets of mitogenomes from NCBI and Kolora et al. (2021), and further organized the sets as "outgroup," "ingroup," and a reference sequence of interest for use in *unikseq*. We ran *unikseq* (v1.0.0) for each desired target assay to identify sequence sets that were unique to each reference relative to the designated outgroup but tolerated in their corresponding ingroup. In our use-cases, the ingroups contained only sequences from the same species as the reference. Each use-case required specific parameters to be set according to the level of shared identity between ingroup and outgroup sequences (Table 1). In each case, *unikseq* was initially run using default settings. If sequences were generated by default settings, these were used in assay design. If no sequences were identified during the initial query, the uniqueness parameter (-m option) was raised, then the leniency parameter (-l option) was increased, then the percent unique parameter (-u option) was decreased. The final parameters are shown in Table 1. By performing manual adjustments on selected parameters and re-running the script after each change, we retained maximum possible stringency in our queries.

To evaluate the scalability of *unikseq*, we ran it on 1k, 10k, and 100k complete ~30 kbp SARS-CoV-2 genomes downloaded from the GISAID repository (Khare et al., 2021). This was done only to observe performance on complete genome sequences that are similar in size to animal mitogenomes but are available in large numbers compared to the latter. We note that no assays were developed from these queries.

2.3.1 | Assay design and development

Unikseq-identified mitogenome region sequences were systematically searched using Beacon Designer 8.21 (Premier Biosoft International) in conjunction with Geneious Prime to design optimized assays. The longest output sequences were assessed for first and progressively shorter sequences assessed if suitable primers/probe candidates were not found. Assay design followed principles outlined in Langlois et al. (2021) and included running the TaqMan search in Beacon Designer with the following settings and considerations for primers: primer length 18–23 nt, amplicon length 80–400 bp, T_m target 59 ± 5°C, maximum primer pair T_m mismatch 3°C, GC content target 50%, no or minimal primer dimer, 3′ ends of primers must be selective for the target taxon, the 3′ end of primers should preferably have a GC clamp. Probe settings and considerations were probe length 25–30 nt, 5′ end

of probe must be selective for the target taxon, probes cannot have a G at the 5′ end with preference to C followed by a C/G to create a clamp, T_a target 55°C, GC content 50%, and no or minimal hairpin (maximum default cross dimer ΔG = −6 (internal) kcal/mol). The primers/probe combination with the best overall scores were then located on the sequence alignments using Geneious Prime to verify specificity for the target taxon and then used as queries to search against the NCBI non-redundant (nr) sequence database to ensure no other possible confounding taxa DNA would be amplified using the assays.

To determine assay specificity, the primers were first tested against 10 picograms per reaction of target and non-target species total DNA (Table S2 in Appendix S1) with two technical replicates using QIAcuity EvaGreen (QIAGEN Inc.) qPCR assay reagents. DNA amplification thermocycle conditions were consistent with those used previously with an initial activation step of 2 min at 95°C followed by 50 cycles of 15 s denaturation at 95°C, 30 s annealing at 64°C, and 45 s extension at 72°C, carried out on Bio-Rad CFX96 Real-Time PCR Detection Systems (Bio-Rad Laboratories (Canada) Ltd; Hobbs et al., 2020; Matthias et al., 2021; Robinson et al., 2022; Veldhoen et al., 2016). Agarose gel visualization was used to verify the generation of an amplified product (amplicon) of the expected size and absence of amplicon in all non-target species. If no primer sets amplified target species DNA reliably and selectively, new primer candidates would have been designed as described in Langlois et al. (2021). The primers were then tested in combination with designed TaqMan hydrolysis probe candidates on the target DNA plus the non-target species (two technical replicates) using QIAcuity Probe PCR kit (QIAGEN Inc.). If amplification was detected within 50 cycles, the replicate was scored as positive. A primer/probe set was then further tested if it exhibited specificity for the target species. At this stage, isolated total DNA from target and non-target voucher specimens plus human were run with 23 additional technical replicates, for a total of 25 replicates per assay (Matthias et al., 2021).

To empirically determine assay sensitivity, assays were tested using standard curves generated using gBlocks® synthetic DNA from Integrated DNA Technologies matching the respective target amplicons as described previously (Hobbs et al., 2019). Briefly, a 10^7 copies/μL synthetic DNA stock was made containing a working tRNA solution comprised of 10 ng/μL tRNA in TE Buffer pH 8 as a stabilizer (Sigma-Aldrich Canada Co.; Thermo Fisher Scientific Inc.). One μL of this dilution was added to 31 μL of working tRNA solution to produce a temporary stock containing 312,500 copies/μL. This stock was then further serially diluted five-fold with working tRNA solution to produce a range of 10 synthetic DNA working stock concentrations from 31,250 to 0.016 copies/μL. Two μL of each working stock were used in qPCR reactions resulting in a range of 0.032–62,500 copies per qPCR reaction with 25 technical replicates. Additionally, 25 reactions were run with 2 μL UltraPure™ DNAse/RNAse-free distilled water (Invitrogen) as no-template negative controls (NTCs). The generated data in .csv format are obtained from File S5 in Appendix S1.

2.3.2 | Statistical analyses

To highlight the difference in potential for selective assay design between situations where researchers are able to use full mitogenomes and situations where a shorter DNA sequence is used, we performed an in silico comparison of common barcode regions from the *mt-co1*, *mt-cyb*, *mt-rnr2* genes and the three largest continuous output sequences from *unikseq* queries for each target species. We calculated the overall percentage of unique nucleotides in each of the queried regions of sequence alignments using Geneious Prime. The percentage of unique nucleotides is a useful metric for comparing base pair variation between sequences in a region, where a higher percentage indicates greater target taxon assay design potential. This analysis was performed in silico only, and no assays were produced using the barcode regions.

Standard curves were generated for each of the four assays using gBlocks® data to relate C_t values with starting DNA copy number as mentioned in Section 2.3.1 (File S5 in Appendix S1). These were used on binomial data to calculate the limit of blank (LOB), limit of detection (LOD), and limit of quantification (LOQ) using eLowQuant based on a modified Binomial-Poisson distribution model (Lesperance et al., 2021). An additional limit of quantification for continuous data ($LOQ_{continuous}$) was determined as the lowest copy number where there is a ≥95% detection over a defined number of technical replicates. This $LOQ_{continuous}$ serves as the breakpoint defining the computational approaches for determining sample copy number. When sample detections for a defined number of technical replicates is ≥95%, then the equation of the best-fit line is used to estimate the copies per reaction. Where sample detections for a defined number of technical replicates is <95%, then Binomial-Poisson distribution-based copy number estimates were calculated using eLowQuant (Lesperance et al., 2021).

2.3.3 | Field validation

The assays developed for detection of American bullfrog, Burbot, and Quillback rockfish were applied to field samples with known presence of each respective species. Three replicate water samples were taken from an American bullfrog tadpole tank in the Outdoor Aquatics Unit at the University of Victoria. Two water samples were taken by Ecological Logistics & Research Ltd researchers from a site in Crooked Creek, Alberta with known presence of Burbot. Six water samples were taken from the surface and bottom of the Pacific Canada Pavilion tank at the Vancouver Aquarium, which contains Quillback rockfish. Each set of samples also included procedural negative control samples that always resulted in no detection.

3 | RESULTS

3.1 | *Unikseq* outputs

We benchmarked *unikseq* on both Linux and mac OS systems using the test data supplied with *unikseq* on GitHub (v1.0 -k 25 -r CEMA.fa

-i shark.fa -o teleost.fa -s 100 -p 25 -l 1 -u 90), which comprises 189 and 868 target and non-target mitogenome sequences for shark.fa and teleost.fa, respectively. The application used 2.5 GB RAM and ran in 49.1 s (wall clock time) on a single CPU thread on a MacBook Pro (2.6 GHz 6-Core Intel Core i7 chipset with 16GB RAM onboard) running mac OS (Catalina v10.15.7). On a server-class CentOS Linux 7 system with 144 Intel(R) Xeon(R) Gold 6254, 3.10 GHz CPUs with 3TB RAM, the same GitHub test sample ran in 31.5 s (wall clock time), using a single thread and required 2.5 GB RAM. To demonstrate the scalability of *unikseq*, we ran *unikseq* queries on 1k, 10k and 100k complete ~30 kbp SARS-CoV-2 genomes downloaded from the GISAID repository and show how the tool runs in linear time, with its average peak memory usage tracking with the number of input sequences (Table S3 and Figure S1 in Appendix S1) (Khare et al., 2021). All other *unikseq* queries using the experimental data presented in the present study used a maximum of 6 GB of RAM on Linux systems and took <1 min to run. *Unikseq* queries included all available mitogenome sequences for species in the same taxonomic group that were potentially co-occurring, representing a real-life application of the tool for eDNA assay design (File S1 in Appendix S1).

Unikseq outputs include: (1) FASTA file containing all identified unique sequences that can be used for assay design; (2) tsv file listing all the *k*-mers used for the analysis that can be used for constructing butterfly graphs; and (3) a text file for the detailed summary of the *unikseq* run. All *unikseq* query information for the four species is shown in Table 1 and *unikseq* output files are available as File S6–S9 in Appendix S1.

3.2 | Use-Case #1: American bullfrog

Designing qPCR-based eDNA assay for the American bullfrog is challenging due to its cosmopolitan geographic distribution. Herein, we used five mitogenome sequences as the ingroup to represent different geographic populations of the target taxon. Given its generic habitat preference, an extensive list of sympatric taxa consisting of 182 mitogenomes (For phylogenetic tree, see Figure S2 in Appendix S1) was considered in designing primers and probe to ensure high specificity of the assay. With *unikseq*, a total of 44 unique regions (ranging from 100 to 880 bp) within the American bullfrog mitogenome were identified (Table 1). The *unikseq*-identified regions compared to the corresponding *mt-co1* gene region show a higher percentage of unique nucleotides (Figure 2) attesting to a greater likelihood for meeting the specificity criterion required for a suitable eDNA assay design. From these recommended sequences, a specific region of the mitogenome encoding for the NADH dehydrogenase 5 (*mt-nd5*) gene generated a suitable primer pair and probe with the necessary specificity and sensitivity performance as described in Section 2.3.1.

The assay designed for this region, eLICA5 (Table 2 and Figure S2 in Appendix S1), demonstrates high specificity for American bullfrog in silico and when tested in vivo against 14 amphibian, human, dog, and cat gDNA samples (Table S2 in Appendix S1). Moreover, the eLICA5

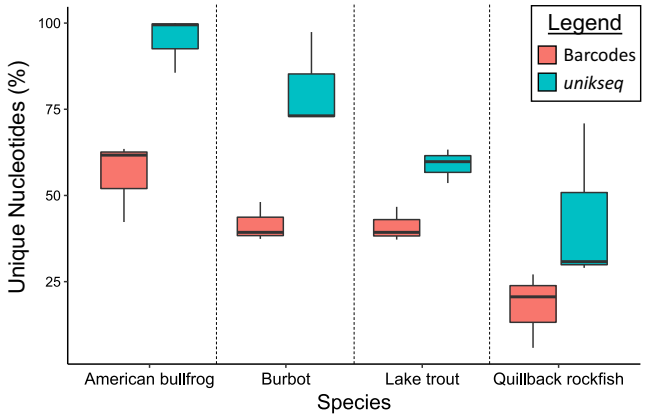

FIGURE 2 Comparison of the percentage of unique nucleotides between common species barcodes and *unikseq* outputs. Common barcode regions of *mt-co1*, *mt-cyb*, and *mt-rnr2* for each species appearing in the alignments shown in Figures S3, S6, S8, and S11 in Appendix S1 were assessed using Geneious Prime to determine the percentage of unique nucleotides as described in Section 2.3.2. The three largest continuous *unikseq* output sequences from each species (File S6–S9 in Appendix S1) were then assessed in the same way. The boxplots with medians are indicated for the collective barcodes (red bars) or *unikseq* outputs (blue bars) by species.

assay exhibits high sensitivity having 0.6 (95% confidence interval [CI]: 0.4–1.5) and 2.4 (95% CI: 1.5–5.5) copies/reaction for LOD and LOQ values, respectively (Table 3 and Figure S3 in Appendix S1). For technical replicates with >95% detections where a linear regression of C_t values relative to copy number is appropriate, the amplification efficiency is 96% with an equation of the line of: $y = -3.4271x + 37.352$ (Table 3 and Figure S4 in Appendix S1). Samples from tanks containing American bullfrog tadpoles tested positive with extremely high copy numbers, and the corresponding negative control containing distilled water was negative using eLICA5 (Table 4).

3.3 | Use-Case #2: Burbot

Burbot (*L. lota*) was chosen as an example of a species with few closely related sympatric taxa in its native environments. We used the mitogenome sequences of three Burbot and 56 potentially co-occurring species (for phylogenetic tree, see Figure S5 in Appendix S1) in a *unikseq* query (Figure 4) and obtained 40 unique regions ranging from 122 to 1296 bp long (Table 1). Several candidate assays were designed using the sequences returned by *unikseq*, however the assay targeting a region of the mitochondrially encoded NADH dehydrogenase 4 (*mt-nd4*) gene, named eLOLO4 (Table 2 and Figure S6 in Appendix S1), was chosen due to its specificity and superior sensitivity (Table 3 and Figure S7 in Appendix S1).

While it was feasible to create eDNA assays targeted to the *mt-co1* barcode region (Figure 4a,b), this sequence was still suboptimal compared with other regions identified by *unikseq* based on the percentage of nucleotide differences present in the alignment (Figure 2), and when known performance parameters are considered

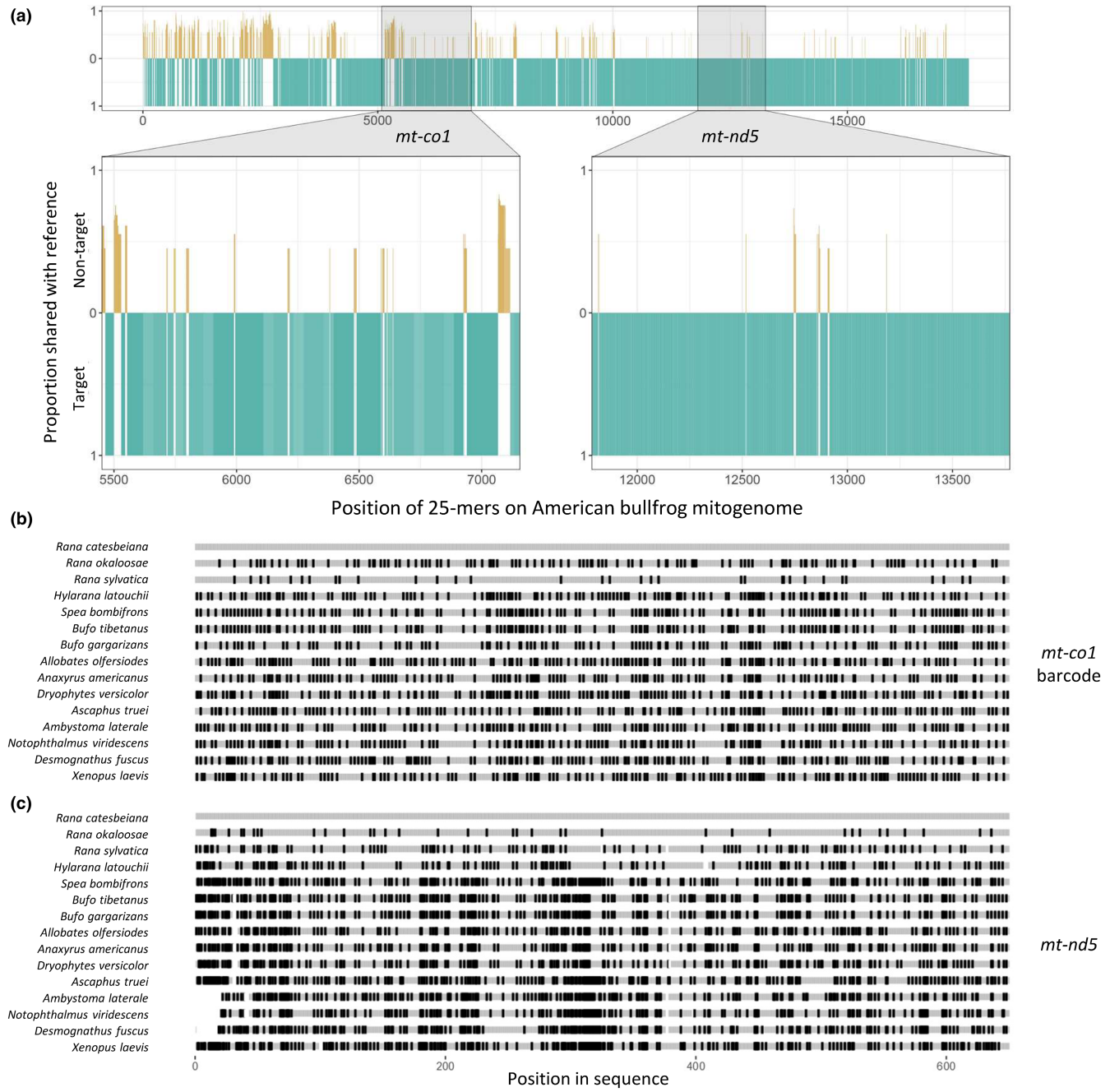

FIGURE 3 (a) American bullfrog butterfly plot with focus on the *mt-co1* barcode and *unikseq*-identified *mt-nd5* gene regions. American bullfrog mitogenomes were processed by *unikseq* and the respective *k*-mers were compared to mitogenomes from outgroup species using a sliding 25-mer window across the entire of the reference mitogenome. The butterfly bar plots show the proportion of 25-mers that have shared identity with outgroup species (gold bars) or are unique to the American bullfrog (blue bars). These blue locations are good candidates for robust qPCR test design. (b) Alignments of American bullfrog with other anuran mitogenome sequences at the *mt-co1* barcode and (c) the *unikseq*-identified region of the *mt-nd5* gene, shortened from 880 to 652 bp within the design region for comparison. While all sequences in the queries were used for assay designs, a subset of species are shown for illustration purposes.

(e.g., GC content, nucleotide base runs). The best performing Burbot assay, eLOLO4 (Table 2 and Figure S6 in Appendix S1), was verified to be specific using gDNA from six Burbot target voucher specimens and 14 co-occurring teleost species as well as dog, cat, and human (Table S2 in Appendix S1). Using serially diluted synthetic DNA the LOD is 0.3 copies/reaction (95% CI: 0.3–0.6), the LOQ is 1.3 copies/reaction (95% CI: 0.9–2.1; Table 3 and Figure S7 in Appendix S1). The amplification efficiency is 85% for technical replicates with >95% detection where a linear regression of C_t values relative to copy number is appropriate with an equation of the line of: $y = -3.7514x + 40.665$ (Table 3 and Figure S7 in Appendix S1). eDNA samples from Crooked Creek with known Burbot presence tested positive using eLOLO4 (Table 4), and samples from sites without Burbot were negative (data not shown).

3.4 | Use-Case #3: Lake trout

As a freshwater fish with several potential co-occurring species from the same genus, the lake trout (*S. namaycush*) poses a challenge for robust eDNA assay design. We used six lake trout mitogenomes in *unikseq* queries to account for any intraspecies genetic variation. Mitochondrial genomes of 60 potentially co-occurring freshwater teleost species were included in the *unikseq* query outgroup and sequence alignments (for phylogenetic tree, see Figure S5 in Appendix S1). Running *unikseq* on these sequences (Figure 5) yielded 14 unique regions within the lake trout mitogenomes ranging between 134 and 3796 bp (Table 1). Similarly to the two use-cases presented above, the *unikseq*-identified regions had a higher percentage of unique nucleotides compared to the corresponding *mt-co1* gene region (Figure 2) from which suitable eDNA assays could be created. Several candidate assays were designed in these regions, and, of those, the assay named eSANA1 that targeted a region of the mitogenome encoded NADH dehydrogenase 2 (*mt-nd2*) gene performed best in in vitro validations (Table 2 and Figure S8 in Appendix S1).

The eSANA1 assay demonstrates complete specificity for lake trout when tested alongside 15 relevant teleost, human, and dog DNA samples (Table S2 in Appendix S1). In tests using synthetic DNA, eSANA1 showed high sensitivity with a LOD of 1 copy/reaction (95% CI: 0.7–1.8) and LOQ of 3.9 copies/reaction (95% CI: 2.7–6.7; Table 3 and Figure S9 in Appendix S1). The amplification efficiency for technical replicates with >95% detection where a linear regression of C_t values relative to copy number is appropriate is 73% with an equation of the line of: $y = -4.204x + 42.288$ (Table 3 and Figure S9 in Appendix S1).

3.5 | Use-Case #4: Quillback rockfish

Rockfish, which includes the *Sebastes* genus, is comprised of over 100 species, the majority of which are from the North Pacific Ocean. In British Columbia, Canada, there are 36 *Sebastes* species that often reside together in the same environments. The close phylogenetic and sympatric relationships within this genus provide an extremely difficult set of circumstances to overcome when designing an eDNA assay for one specific species.

We used two mitogenomes of the Quillback rockfish (*S. maliger*) and compared them with 78 other *Sebastes* species and 19 other fish species that co-occur geographically (For phylogenetic tree, see Figure S10 in Appendix S1). The *unikseq* results demonstrate the difficulty in finding unique stretches of DNA sequence (Figure 6), six unique regions were generated ranging from 100 to 195 bp long (Table 1). Like the use-cases above, the *unikseq*-identified regions had a higher percentage of unique nucleotides compared to the corresponding *mt-co1* gene region (Figure 2) from which suitable eDNA assays could be created. From those regions the best assay candidates were in the *mt-d-loop* region (Figure 6). From in silico and in vitro validation results, the eSEMA3 assay performed much better than other tested candidates (Table 2 and Figure S11 in Appendix S1). The eSEMA3 assay showed complete specificity for the target species when tested against tissue-derived DNA samples of 21 other *Sebastes* species, and 21 other co-occurring taxa including human, cat, and dog (Table S2 in Appendix S1). Finally, in tests using synthetic DNA, eSEMA3 shows an LOD of 1.8 copies/reaction (95% CI: 1.1–3.5) and LOQ of 4.2 copies/reaction (95% CI: 2.8–7.8; Table 3 and Figure S12 in Appendix S1). The amplification efficiency is 99% for technical replicates with >95% detections where a linear regression of C_t values relative to copy number is appropriate with an equation of the line of: $y = -3.5022x + 38.512$ (Figure S12 in Appendix S1). All eDNA samples from the Vancouver Aquarium Pacific Canada Pavilion which contains Quillback rockfish tested positive using eSEMA3 (Table 4).

4 | DISCUSSION

Reliable detection of eDNA from environmental samples demands reproducible, sensitive, and accurate analyses. Achieving this is only

TABLE 2 Primer and probe sequences of all assays developed in the present paper.

Species	Gene	Assay name	Forward primer (5′ → 3′)	Reverse primer (5′ → 3′)	Probe (5′ → 3′)
American bullfrog	*mt-nd5*	eLICA5	ATACACCGCACTATTACT	AAGAGGACTGATAGGTAAG	FAM-CCTTAACT CAGCCTGACAAC TTATT-ZEN/IB
Burbot	*mt-nd4*	eLOLO4	CTTGCTGCTGTATTACTAA	ATAAACTATTTCCTTGGAGAG	FAM-ATGAGTCGTAT TATACCATAACC GCCTAGT-ZEN/IB
Lake trout	*mt-nd2*	eSANA1	GGGCTTATCCTGTCTACATG	CCCAGGGATAGAAGCACTA	FAM-TGACTCTT CCTTAATTATCG CATTA-ZEN/IB
Quillback rockfish	*mt-d-loop*	eSEMA3	CGAAGGTATTACATAAAGCA	GAGTGTTTGTGTTAGGTCTTA	FAM-CCAACAATCAT TTATAAGGACTG AGCGAAT-ZEN/IB

Note: All probes used FAM fluorophores quenched with ZEN/Iowa Black™ FQ (ZEN/IB).

TABLE 3 Limits of detection and quantification for binary and continuous data for each of the four validated assays.

End-use case	Target taxon	Assay name	Binary data						Continuous data				
			LOD (c/rxn)[a]	LOD 95% CI lower	LOD 95% CI upper	LOQ (c/rxn)	LOQ 95% CI lower	LOQ 95% CI upper	$LOQ_{continuous}$[b] (c/rxn)	Slope	Y-intercept	R^2 value	% Efficiency
1	American bullfrog	eLICA5	0.6	0.4	1.5	2.4	1.5	5.5	4	−3.427	37.352	0.9984	96
2	Burbot	eLOLO4	0.3	0.3	0.6	1.3	0.9	2.1	4	−3.751	40.665	0.9985	85
3	Lake trout	eSANA1	1	0.7	1.8	3.9	2.7	6.7	20	−4.204	42.288	0.9874	73
4	Quillback rockfish	eSEMA3	1.8	1.1	3.5	4.2	2.8	7.8	20	−3.502	38.512	0.9985	99

Note: Full descriptions of statistical analyses and R script for dealing with binary data using eLowQuant are found in Lesperance et al. (2021).

[a]Copies/reaction.

[b]Lowest copies/reaction tested with ≥95% hits.

possible with proper study design and high quality eDNA assay design, validation, and implementation (Cristescu & Hebert, 2018; Goldberg et al., 2016; Hobbs et al., 2020; Veldhoen et al., 2016). To facilitate this, we have developed *unikseq*, a tool capable of parsing large sequence datasets with high efficiency and flexibility to address the many challenges faced in designing robust eDNA assays.

Assay design begins with the identification of pertinent confounding taxa and attainment of verified sequences with strict attention to any possible errors in the available sequence information (Langlois et al., 2021; Mioduchowska et al., 2018). The accuracy of DNA sequences should be closely examined, as misidentified sequences will have significant impacts on *unikseq* outputs (Stavrou et al., 2018). Special care at these first stages of assay design will vastly reduce the occurrence of errors later in the workflow. With the rapid growth in the eDNA field, it is important to have powerful in silico tools that can make assay design more rigorous and reproducible. As the number of available inter- and intraspecies mitogenome sequences increases, the need for identifying suitable assay design regions will become increasingly apparent.

We developed the *unikseq* utility for the identification of potentially unique sequence stretches in a reference and user-defined "ingroup" sequence set, relative to a user-defined "outgroup" sequence set. For the ingroup, the user can designate closely related population and/or species that can be tolerated in terms of sequence uniqueness, thus easily accounting for haplotypes that might be present in other geographic populations or closely related taxa. Sequences identified by *unikseq* can then be used for the downstream purpose of primers/probe qPCR design narrowing the search space considerably and identifying unique regions that are likely to generate more specific qPCR amplification signals. With *unikseq*, entire mitogenomes can be simultaneously processed, which substantially increases the sequence possibilities compared to that of earlier eDNA assay design paradigms often limited to mitogenome barcodes. With substantial advances in whole genome shotgun sequencing technologies including the increased sequence throughput and lower costs, we anticipate that full mitogenome sequences for a wide range of taxa will become increasingly available in public repositories.

Our *unikseq*-based eDNA assay development workflow streamlines the process of selecting appropriate DNA sequences without excluding potentially useful sequence resources in the design process. Indeed, *unikseq* makes it possible to include many more outgroups more easily during assay design including species that are frequent potential confounders such as humans and pets. The *unikseq* script facilitates the selection of mitogenome regions most suitable for eDNA assay design and enables assay design efforts to rapidly focus on those areas. The American bullfrog and Burbot were examples of species with strong assay design potential throughout their mitogenomes. Even so, *unikseq* enabled unbiased selection of promising regions for robust assay design. Due to the extensive invasive range of the American bullfrog, many potentially sympatric species had to be considered with a large range of taxonomic relatedness to the target species. In addition, *unikseq* was especially helpful in enabling assay design for Lake trout and Quillback rockfish,

TABLE 4 Field validations using eDNA samples with known presence of American bullfrog, Burbot, and Quillback rockfish.

Assay used	Location name	Sample collection date	IntegritE-DNA™ frequency (/4)	Target assay frequency (/8)	Estimated total copies/L (±SE)
eLICA5	UVic OAU Tadpole Tank	Aug 10 2022	4	8	2,792,880 ± 400,260
		Aug 10 2022	4	8	2,165,488 ± 327,373
		Aug 10 2022	4	8	2,163,011 ± 311,261
eLOLO4	Crooked Creek, Alberta	Apr 5 2021	4	4	188 ± 103
		Apr 5 2021	4	5	273 ± 138
eSEMA3	Vancouver Aquarium	Jun 21 2022	4	8	2620 ± 322
		Jun 21 2022	4	8	2215 ± 370
		Jun 21 2022	4	8	4136 ± 529
		Jun 21 2022	4	8	4815 ± 495
		Jun 21 2022	4	8	4216 ± 427
		Jun 21 2022	4	8	10,610 ± 1003

Note: Water samples came from locations with known target species present: a bullfrog tadpole tank in the Outdoor Aquatics Unit (OAU) at the University of Victoria, a location in Crooked Creek, Alberta, and a tank from the Pacific Canada Pavilion at the Vancouver Aquarium. Procedural blanks did not have any detections (not shown).

where many closely related co-occurring taxa strictly limit design potential using the mitogenome. Irrespective of the challenge, manually parsing through this quantity of sequence data would have been a gargantuan task, whereas the work was completed by *unikseq* in mere minutes.

It can be challenging and time-consuming to manually identify distinct regions in the mitogenome of species with limited genetic variability. Comparisons of the overall nucleotide heterogeneity in species groups between common barcode regions and *unikseq* output sequences demonstrates clear advantages to assay designs using the tool. Instead of defaulting to the standard barcode sequences for assay design, *unikseq* enables quick identification of distinctive regions that increase the potential for selective assay design. The genetic regions that are most appropriate for assays can vary between contexts, especially among cases with different demands for specificity and resolution of target taxa as demonstrated by our examples.

Application of the *unikseq* utility is not limited to full mitogenomes and works well using both long and short gene sequences and incomplete assemblies. The *unikseq* queries of coronavirus genome data served to demonstrate scalability of the tool exclusively. Each coronavirus genome is roughly twice the size of a typical animal mitogenome, so the amount of data parsed by *unikseq* was several orders of magnitude larger than any of the queries performed for eDNA assay development shown in the present study. The time required for coronavirus runs scaled linearly with the size of the datasets using the CentOS Linux 7 system. This is in contrast to many other tools designed to parse sequence data such as multiple sequence alignment tools, whose run times increase exponentially as the dataset size increases (Zhu et al., 2015). As demonstrated herein with the qPCR assay design queries we presented and for most typical eDNA use cases and field applications, which consider up to 1000 mitogenomes, *unikseq* will run on computers with modest hardware specifications (i.e., <8 GB RAM on a single CPU).

Unikseq has many potential applications that could strengthen existing eDNA assay design pipelines. The taxa-specific targeted assays presented here obviously benefitted from assisted sequence selection through high specificity of designed primers and probe with higher performance characteristics than could be achieved with limited considered sequence. But this could be extended to several other types of studies. While targeted eDNA studies typically target at the species level, *unikseq* search parameter flexibility allows for the potential to find mitogenome regions appropriate for targeting more specific taxon groups (e.g., subspecies) or less specific groups (e.g., genus or family). The workflow could aid searches for intra-taxon assay region appropriateness, for instance in search of useful SNPs for haplotype analyses at a mitogenome-wide scale (Shah et al., 2005). Metabarcode-based investigations could also benefit from new regions that are more suitable for amplifying barcoding regions for a specific taxon.

Herein, we ran *unikseq* on 100s–1000s of animal mitogenome sequences and, to assess its scalability on similar-size genomes, we ran it to completion on up to 100,000 SARS-CoV-2 genomes available from GISAID. But for most eDNA applications, and when comparing hundreds to up to one or a few thousands mitogenomes that one may suspect co-occur in each environment, the relatively low memory footprint and fast run time of *unikseq* should make it suitable to run on most desktop computers. As more sequences are supplied to the program, and particularly to the non-target sequence set, the requirements for unique region identification may become harder to meet. This is especially a concern when designing assays to distinguish between closely related sympatric species, as exemplified in our work on rockfish and bullfrog. Built-in leniency parameters help *unikseq* identify candidate regions within a reference despite them sharing *k*-mers with sequences in the non-target set. A similar option for adjusting the tolerance to haplotype variation with the target (in-group) set can be fine-tuned to identify suitable regions for eDNA assay design. For example, users may opt to increase the leniency (-l)

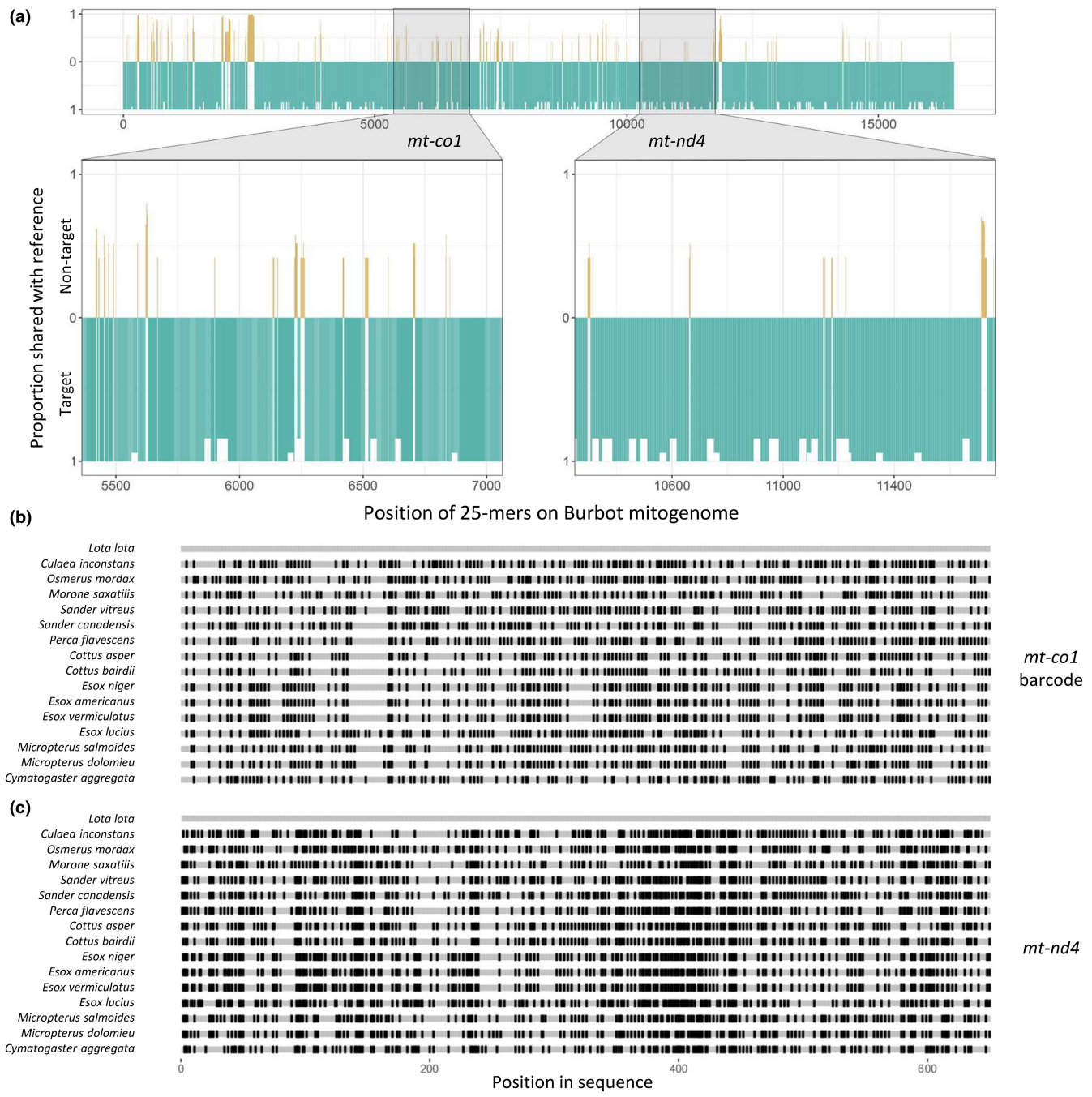

FIGURE 4 (a) Burbot butterfly plot with focus on the *mt-co1* barcode and *unikseq*-identified *mt-nd4* gene regions. (b) Alignments of Burbot with other freshwater teleost mitogenome sequences at the *mt-co1* barcode and (c) the *unikseq*-identified region of the *mt-nd4* gene, extended from 480 to 652 bp for comparison. See the Figure 3 legend for more details.

parameter of *unikseq* to tolerate *k*-mers shared between reference and non-target sequences, and/or decrease the minimum proportion (-p) of target species entries with sequences conserved with that of the reference under evaluation.

Overall, *unikseq* unleashes the full capacity of mitogenomes for assay design with high efficiency. This greatly reduces the compromises between the specificity and sensitivity of eDNA assays while improving the likelihood of high quality eDNA assay design by unbiased selection of unique mitogenome regions without the need to manually parse through whole alignments. As a tool for harnessing the expanding potential of genetic resources, we anticipate broad application of *unikseq* for eDNA assay design across the tree of life.

AUTHOR CONTRIBUTIONS

Michael J. Allison, M. Louie Lopez, Neha Acharya-Patel and Jacob J. Imbery: Data acquisition and analysis, manuscript preparation. René L. Warren: Model concept, code development, data analysis, manuscript preparation. Lauren Coombe and Cecilia L. Yang: Data

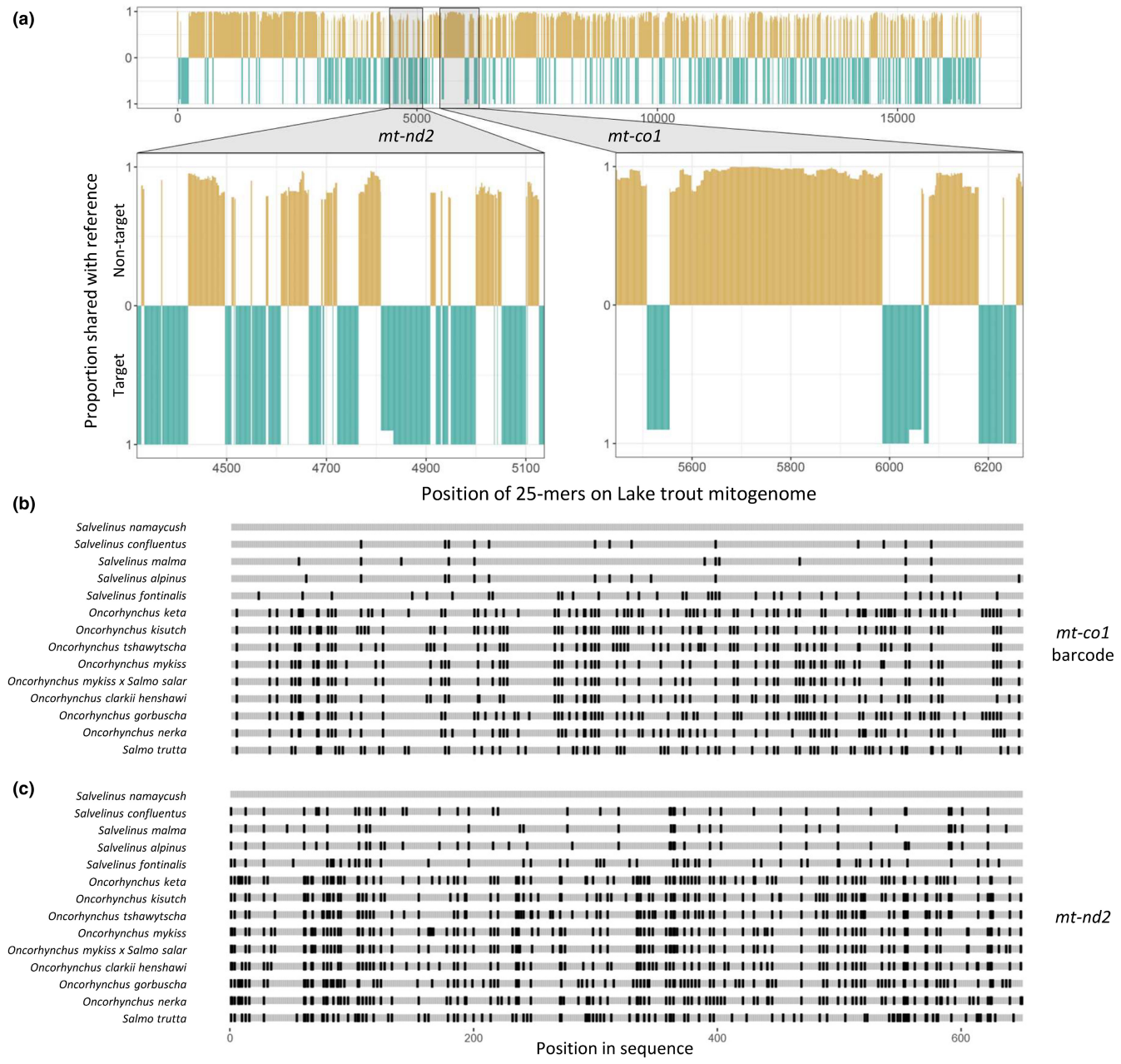

FIGURE 5 (a) Lake trout butterfly plot with focus on the *mt-co1* barcode and *unikseq*-identified *mt-nd2* gene regions. (b) Alignments of Lake trout with other salmonid mitogenome sequences at the *mt-co1* barcode and (c) the *unikseq*-identified region of the *mt-nd2* gene, truncated from 1590 to 652 bp for comparison. See the Figure 3 legend for more details.

analysis, manuscript preparation. Inanc Birol and Caren C. Helbing: Model concept, data analysis, manuscript preparation, secured funding.

ACKNOWLEDGMENTS

The authors thank Dr. G. Owens at the University of Victoria for generously providing access to rockfish mitogenome sequences. We also thank G. Rudman from ELR Ltd. and Government of Yukon, Department of Highways and Public Works – Transportation Division for provision of samples and providing some funding for assay development. We also thank E. Crichton, A. Dema, Y. Ren, E. Groenwold, and M. Bonderud for their excellent technical assistance. The present work was funded in part by Genome Canada, Genome British Columbia, and Genome Québec large-scale applied research project #312ITD to CCH and IB. MJA and JJI are recipients of Natural Sciences and Engineering Research Council of Canada (NSERC) CGS-D3 and CGS-M scholarships, respectively.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

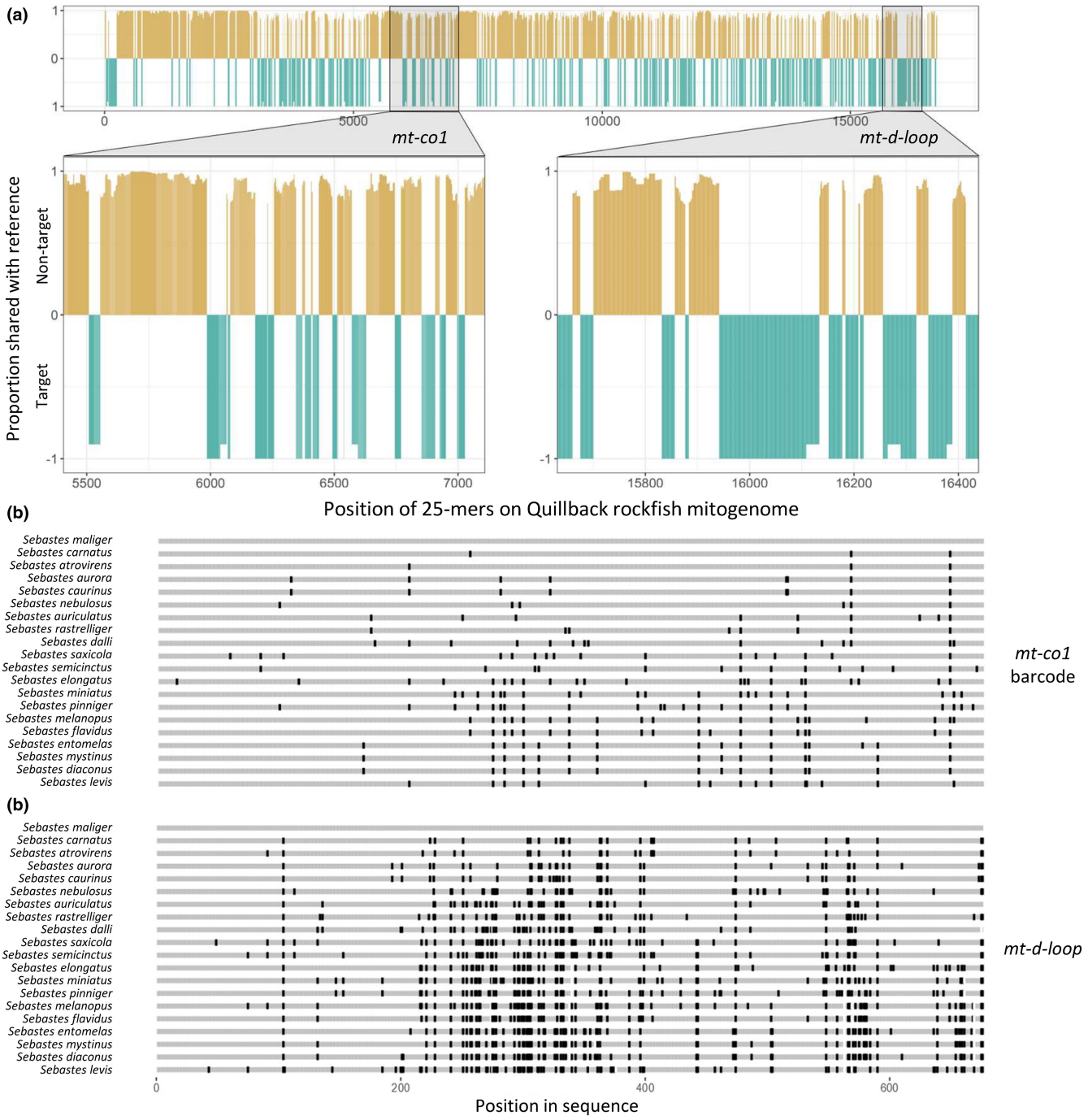

FIGURE 6 (a) Quillback rockfish butterfly plot with focus on the *mt-co1* barcode and *unikseq*-identified *mt-d-loop* gene regions. (b) Alignments of Quillback rockfish with other *Sebastes* spp. at the *mt-co1* barcode and (c) the *unikseq*-identified region of the *mt-d-loop* gene, extended from 195 to 652 bp for comparison. See the Figure 3 legend for more details.

DATA AVAILABILITY STATEMENT

Raw data are available upon request to the corresponding author.

ORCID

Caren C. Helbing https://orcid.org/0000-0002-8861-1070

REFERENCES

Aun, E., Brauer, A., Kisand, V., Tenson, T., & Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Computational Biology*, *14*(10), e1006434.

Bylemans, J., Furlan, E. M., Gleeson, D. M., Hardy, C. M., & Duncan, R. P. (2018). Does size matter? An experimental evaluation of the relative abundance and decay rates of aquatic environmental DNA. *Environmental Science & Technology*, *52*(11), 6408–6416.

Cristescu, M. E. (2019). Can environmental RNA revolutionize biodiversity science? *Trends in Ecology & Evolution*, *34*(8), 694–697.

Cristescu, M. E., & Hebert, P. D. N. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. *Annual Review of Ecology, Evolution, and Systematics*, *49*(1), 209–230.

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*(4), 423–425.

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., Spear, S. F., McKee, A., Oyler-McCance, S. J., Cornman, R. S., Laramie, M. B., Mahon, A. R., Lance, R. F., Pilliod, D. S., Strickler, K. M., Waits, L. P., Fremier, A. K., Takahara, T., Herder, J. E., … Gilbert, M. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307.

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Han, Z., Liu, M., Liu, Q., Zhai, H., Xiao, S., & Gao, T. (2021). Chromosome-level genome assembly of burbot (*Lota lota*) provides insights into the evolutionary adaptations in freshwater. *Molecular Ecology Resources*, *21*(6), 2022–2033.

Hobbs, J., Adams, I. T., Round, J. M., Goldberg, C. S., Allison, M. J., Bergman, L. C., Mirabzadeh, A., Allen, H., & Helbing, C. C. (2020). Revising the range of Rocky Mountain tailed frog, *Ascaphus montanus*, in British Columbia, Canada, using environmental DNA methods. *Environmental DNA*, *2*(3), 350–361.

Hobbs, J., Round, J. M., Allison, M. J., & Helbing, C. C. (2019). Expansion of the known distribution of the coastal tailed frog, *Ascaphus truei*, in British Columbia, Canada, using robust eDNA detection methods. *PLoS One*, *14*(3), 16.

Jensen, M. R., Sigsgaard, E. E., Liu, S., Manica, A., Bach, S. S., Hansen, M. M., Moller, P. R., & Thomsen, P. F. (2021). Genome-scale target capture of mitochondrial and nuclear environmental DNA from water samples. *Molecular Ecology Resources*, *21*(3), 690–702.

Kans, J. (2013). *Entrez direct: E-utilities on the Unix command line. Entrez programming utilities help*. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK179288/

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166.

Khare, S., Gurry, C., Freitas, L., Schultz, M. B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R. T., Yeo, W., Team, G. C. C., & Maurer-Stroh, S. (2021). GISAID's role in pandemic response. *China CDC Weekly*, *3*(49), 1049–1051.

Kolora, S. R. R., Owens, G. L., Vazquez, J. M., Stubbs, A., Chatla, K., Jainese, C., Seeto, K., McCrea, M., Sandel, M. W., Vianna, J. A., Maslenikov, K., Bachtrog, D., Orr, J. W., Love, M., & Sudmant, P. H. (2021). Origins and evolution of extreme life span in Pacific Ocean rockfishes. *Science*, *374*(6569), 842–847.

Kronenberger, J. A., Wilcox, T. M., Mason, D. H., Franklin, T. W., McKelvey, K. S., Young, M. K., & Schwartz, M. K. (2022). eDNAssay: A machine learning tool that accurately predicts qPCR cross-amplification. *Molecular Ecology Resources*, *8*, 2994–3005.

Langlois, V. S., Allison, M. J., Bergman, L. C., To, T. A., & Helbing, C. C. (2021). The need for robust qPCR-based eDNA detection assays in environmental monitoring and species inventories. *Environmental DNA*, *3*(3), 519–527.

Leray, M., Knowlton, N., & Machida, R. J. (2022). MIDORI2: A collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. *Environmental DNA*, *4*(4), 894–907.

Lesperance, M. L., Allison, M. J., Bergman, L. C., Hocking, M. D., & Helbing, C. C. (2021). A statistical model for calibration and computation of detection and quantification limits for low copy number environmental DNA samples. *Environmental DNA*, *3*, 970–981.

MacDonald, A. J., & Sarre, S. D. (2017). A framework for developing and validating taxon-specific primers for specimen identification from environmental DNA. *Molecular Ecology Resources*, *17*(4), 708–720.

Matthias, L., Allison, M. J., Maslovat, C. Y., Hobbs, J., & Helbing, C. C. (2021). Improving ecological surveys for the detection of cryptic, fossorial snakes using eDNA on and under artificial cover objects. *Ecological Indicators*, *131*, 108187.

Mioduchowska, M., Czyz, M. J., Goldyn, B., Kur, J., & Sell, J. (2018). Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"? *PLoS One*, *13*(6), e0199609.

Nurmukanova, V., Stetsenko, I., Say, A., Ayginin, A., Mikhaylov, I., Matsvay, A., & Shipulin, G. (2022). Application of a novel k-mer primer design algorithm for detecting antibiotic resistance determinants. Centre for Strategic Planning and Management of Biomedical Health Risks (pp. 428).

Robinson, C. L. K., Bergman, L. C., Allison, M. J., Huard, J., Sutherst, J., & Helbing, C. C. (2022). Application of environmental DNA as a tool for detecting intertidal habitat use by forage fish. *Ecological Indicators*, *142*, 109306.

Shah, N., Teplitsky, M. V., Minovitsky, S., Pennacchio, L. A., Hugenholtz, P., Hamann, B., & Dubchak, I. L. (2005). SNP-VISTA: An interactive SNP visualization tool. *BMC Bioinformatics*, *6*, 292.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123.

Stavrou, A. A., Mixão, V., Boekhout, T., & Gabaldón, T. (2018). Misidentification of genome assemblies in public databases: The case of *Naumovozyma dairenensis* and proposal of a protocol to correct misidentifications. *Yeast*, *35*, 425–429.

Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, *10*(3), 512–526.

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular evolutionary genetics analysis version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027.

Thalinger, B., Deiner, K., Harper, L. R., Rees, H. C., Blackman, R. C., Sint, D., Traugott, M., Goldberg, C. S., & Bruce, K. (2021). A validation scale to determine the readiness of environmental DNA assays for routine species monitoring. *Environmental DNA*, *3*, 823–836.

Tu, S. L., Staheli, J. P., McClay, C., McLeod, K., Rose, T. M., & Upton, C. (2018). Base-by-base version 3: New comparative tools for large virus genomes. *Viruses*, *10*(11), 10110637.

Veldhoen, N., Hobbs, J., Ikonomou, G., Hii, M., Lesperance, M., & Helbing, C. C. (2016). Implementation of novel design features for qPCR-based eDNA assessment. *PLoS One*, *11*(11), e0164907.

Warren, R. L., Coombe, L., Mohamadi, H., Zhang, J., Jaquish, B., Isabel, N., Jones, S. J. M., Bousquet, J., Bohlmann, J., & Birol, I. (2019). ntEdit: Scalable genome sequence polishing. *Bioinformatics*, *35*(21), 4430–4432.

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J., & Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience*, *4*, 35.

Wilcox, T. M., McKelvey, K. S., Young, M. K., Jane, S. F., Lowe, W. H., Whiteley, A. R., & Schwartz, M. K. (2013). Robust detection of rare species using environmental DNA: The importance of primer specificity. *PLoS One*, *8*(3), e59520.

Williams, M. A., Hernandez, C., O'Sullivan, A. M., April, J., Regan, F., Bernatchez, L., & Parle-McDermott, A. (2020). Comparing CRISPR-Cas and qPCR eDNA assays for the detection of Atlantic salmon (*Salmo salar* L.). *Environmental DNA*, *3*(1), 297–304.

Yates, M. C., Cristescu, M. E., & Derry, A. M. (2021). Integrating physiology and environmental dynamics to operationalize environmental DNA (eDNA) as a means to monitor freshwater macro-organism abundance. *Molecular Ecology*, *30*(24), 6531–6550.

Zhu, X., Li, K., Salah, A., Shi, L., & Li, K. (2015). Parallel implementation of MAFFT on CUDA-enabled graphics hardware. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *12*(1), 205–218.

SUPPORTING INFORMATION

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

How to cite this article: Allison, M. J., Warren, R. L., Lopez, M. L., Acharya-Patel, N., Imbery, J. J., Coombe, L., Yang, C. L., Birol, I., & Helbing, C. C. (2023). Enabling robust environmental DNA assay design with "unikseq" for the identification of taxon-specific regions within whole mitochondrial genomes. *Environmental DNA*, *5*, 1032–1047. https://doi.org/10.1002/edn3.438