Received: 23 April 2022 | Revised: 24 June 2022 | Accepted: 28 June 2022

DOI: 10.1111/1755-0998.13681

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

eDNAssay: A machine learning tool that accurately predicts qPCR cross-amplification

John A. Kronenberger | Taylor M. Wilcox | Daniel H. Mason | Thomas W. Franklin | Kevin S. McKelvey | Michael K. Young | Michael K. Schwartz

National Genomics Center for Wildlife and Fish Conservation, USFS Rocky Mountain Research Station, Missoula, Montana, USA

Correspondence
John A. Kronenberger, National Genomics Center for Wildlife and Fish Conservation, USFS Rocky Mountain Research Station, Missoula, MT, USA.
Email: john.kronenberger@usda.gov

Funding information
U.S. Department of Defense, Grant/Award Number: RC21-5121

Handling Editor: Carla Martins Lopes

Abstract

Environmental DNA (eDNA) sampling is a highly sensitive and cost-effective technique for wildlife monitoring, notably through the use of qPCR assays. However, it can be difficult to ensure assay specificity when many closely related species co-occur. In theory, specificity may be assessed in silico by determining whether assay oligonucleotides have enough base-pair mismatches with nontarget sequences to preclude amplification. However, the mismatch qualities required are poorly understood, making in silico assessments difficult and often necessitating extensive in vitro testing—typically the greatest bottleneck in assay development. Increasing the accuracy of in silico assessments would therefore streamline the assay development process. In this study, we paired 10 qPCR assays with 82 synthetic gene fragments for 530 specificity tests using SYBR Green intercalating dye (n = 262) and TaqMan hydrolysis probes (n = 268). Test results were used to train random forest classifiers to predict amplification. The primer-only model (SYBR Green results) and full-assay model (TaqMan probe-based results) were 99.6% and 100% accurate, respectively, in cross-validation. We further assessed model performance using six independent assays not used in model training. In these tests the primer-only model was 92.4% accurate (n = 119) and the full-assay model was 96.5% accurate (n = 144). The high performance achieved by these models makes it possible for eDNA practitioners to more quickly and confidently develop assays specific to the intended target. Practitioners can access the full-assay model online via eDNAssay (https://NationalGenomicsCenter.shinyapps.io/eDNAssay), a user-friendly tool for predicting qPCR cross-amplification.

KEYWORDS
assay, base-pair mismatches, eDNA, environmental DNA, random forest, specificity

1 | INTRODUCTION

The analysis of DNA that organisms shed into the environment, termed environmental DNA (eDNA), was pioneered by microbiologists in the late 1980s to assess cryptic biodiversity (Ogram et al., 1987; Somerville et al., 1989). It was adopted by macrobiologists in the 2000s to reconstruct palaeocommunities (Willerslev et al., 2003) and detect rare invasive or imperilled species (Ficetola et al., 2008; Thomsen et al., 2012). Environmental DNA analysis has exploded in popularity as a wildlife monitoring tool because it is less invasive and often more sensitive and cost-effective than traditional monitoring techniques (McKelvey

et al., 2016; Rees et al., 2014). Most eDNA surveys can be divided into two categories. The first category is targeted detection of one or a few taxa of interest, typically via quantitative real-time PCR (qPCR; Wilcox et al., 2013) or droplet digital PCR (ddPCR; Nathan et al., 2014) and more recent isothermal techniques such as loop-mediated isothermal amplification (LAMP; Williams et al., 2017) and CRISPR-Cas (Williams et al., 2019). These approaches are favoured when sensitivity—the ability to detect DNA at very low concentrations—is paramount or near-real-time assessments of species presence are required (Bylemans et al., 2019; Harper et al., 2018; Simmons et al., 2016; Wilcox et al., 2020; Wood et al., 2019). The second category is high-throughput assessment of biodiversity within broad taxonomic divisions, typically via metabarcoding (Deiner et al., 2017). Metabarcoding sacrifices sensitivity for the ability to detect many taxa simultaneously (Harper et al., 2018; Wood et al., 2019). An intermediate solution that offers the potential benefits of both approaches while mitigating the costs is high-throughput qPCR (HT-qPCR). This technique involves running thousands of nanolitre-scale reactions simultaneously on a chip to retain the sensitivity of qPCR while increasing throughput (Wilcox et al., 2020) and decreasing per-taxon identification costs when several taxa are of interest. Of these eDNA analysis methods, targeted qPCR is the go-to for many practitioners given its utility for routine monitoring of particular taxa of interest.

The popularity of targeted qPCR assays highlights the importance of specificity—to be useful, assays need to detect only the target taxon (Wilcox et al., 2013, 2015). It is recommended that specificity be assessed in three stages: in silico by aligning assay oligonucleotides with nontarget sequences, in vitro by testing nontarget DNA, and in situ by testing environmental samples from areas with known presence and absence of the target taxon (Goldberg et al., 2016; Langlois et al., 2021; Thalinger et al., 2021). In silico assessment is typically the first step in the process and also the most ambiguous. The primary objective is to ensure that nontarget sequences have enough base-pair mismatches with primers and hydrolysis probes to prevent amplification and fluorescent signal production. If in silico assessment is inconclusive, DNA templates must either be tested in vitro or the taxonomic or geographical scope of the assay must be revised. However, in vitro testing can cause serious bottlenecks in development because obtaining genomic DNA from diverse (and often very large) suites of species can be logistically difficult and time-consuming—particularly when the needed taxa are rare, can only be sampled at certain times of year, occupy challenging habitats or are protected by law. Another option is to purchase synthetic gene fragments (e.g., Carim et al., 2016), but this can be prohibitively expensive. Given the logistical difficulty of in vitro testing, assay developers may rely too heavily on in silico assessment when declaring their assays to be specific. This is risky because the mismatch characteristics required to preclude amplification are poorly understood.

Several studies have documented the effects of single mismatches in different positions and of different types on amplification efficiency. For primers, the effects of mismatches on amplification are greatest at the 3′ termini and typically decrease with distance from the termini (Lefever et al., 2013; Stadhouders et al., 2010; Wright et al., 2014; but see Bru et al., 2008). Research into different types of mismatches suggests that some purine–purine pairings (AA and AG/GA) and one pyrimidine–pyrimidine pairing (CC) tend to most strongly reduce amplification efficiency (Kwok et al., 1990; Stadhouders et al., 2010; Wright et al., 2014). Our understanding of the effects of multiple mismatches is incomplete. Wright et al. (2014) found two total mismatches to be synergistic, causing reductions in amplification efficiency greater than their sum. Stadhouders et al. (2010) found that two to three primer mismatches completely prevented amplification, but Lefever et al. (2013) identified this threshold at five mismatches, Wilcox et al. (2013) noted some amplification with 10 primer mismatches, and Döring et al. (2019) with 12 primer mismatches. These vastly different results are probably due to complex primer-specific thermodynamic interactions and variations in reaction chemistry and thermal cycling profiles. Regardless, all studies have found a strong negative relationship between primer mismatch number and amplification success (Döring et al., 2019; Lefever et al., 2013; So et al., 2020; Wilcox et al., 2013). Relative to primer mismatches, probe mismatches tend to be more influential (Süß et al., 2009; but see Wilcox et al., 2013). Probes with a minor-groove-binding (MGB) moiety are often used in eDNA assays because they can be short (~15–20 nucleotides) for a given melting temperature, adding specificity as each mismatch represents a greater proportion of total base-pairs (Kutyavin et al., 2000). Mismatches towards the centre of probes may cause greater reduction in fluorescence than those towards the ends (Süß et al., 2009; Yilmaz et al., 2012). For both primers and probes, specificity tends to increase at higher annealing temperatures and at lower melting temperatures relative to the annealing temperature (Rychlik et al., 1990).

With so many interrelated factors, it is no surprise that in silico assessment can be challenging. A number of software programs are capable of generating or testing specific primers, such as PRIMER-BLAST (Ye et al., 2012), which relies on PRIMER3 (Untergasser et al., 2012), as well as DECIPHER (Wright et al., 2014) and MFEPRIMER (Wang et al., 2019). The vast majority of programs use thermodynamic models and simple mismatch heuristics to assess specificity (Noguera et al., 2014). They do not directly incorporate empirical data into their modelling framework, or they do but are not designed for qPCR. As a result, existing programs can seriously underestimate cross-amplification.

Fortunately, the potential of predictive modelling has greatly increased with the advent of machine learning. Supervised machine learning algorithms in particular are useful for classification problems. These algorithms work by associating patterns among user-defined predictor variables with known outcomes, then model performance is assessed either internally (e.g., cross-validation) or using a separate test data set. This modelling framework has quickly gained traction in molecular ecology (Fountain-Jones et al., 2021), including eDNA-based ecological assessments (e.g., Cordier et al., 2018) and primer design (e.g., Cordaro et al., 2021; Döring et al., 2019; Kayama et al., 2021). However, the ability of machine learning to assess qPCR

assay specificity, particularly when hydrolysis probes are involved, has yet to be demonstrated.

Here, we used supervised machine learning to predict qPCR cross-amplification. Our training data set was generated by pairing assay-template mismatch information with empirical qPCR results generated from assays of varying mismatch characteristics. We modelled two different reaction chemistries—SYBR Green intercalating dye and TaqMan hydrolysis probes—as they are likely the most common formats for eDNA assays. A separate model was trained for each. The probe-based model is easily accessible online as eDNAssay (https://NationalGenomicsCenter.shinyapps.io/eDNAssay), a highly accurate and user-friendly tool for eDNA practitioners to assess qPCR assay specificity in silico and potentially reduce the extent of costly in vitro testing.

2 | MATERIALS AND METHODS

2.1 | Assay development

Specificity was assessed for 10 qPCR assays that were either designed exclusively for this study or were previously published (Table 1). Three assays detect crayfishes in the family Cambaridae, including the virile crayfish (*Faxonius virilis*; assays FVR1 and FVR2) and the red swamp crayfish (*Procambarus clarkii*; PCLA). Another three detect fishes in Cyprinidae, including the grass carp (*Ctenopharyngodon idella*; CID1 and CID2) and silver and bighead carps (*Hypophthalmichthys molitrix* and *nobilis*; HYPO). The remaining four detect the pond slider (*Trachemys scripta*; TSC1, TSC2, TSC3 and TSC4), a turtle in Emydidae. These species are invasive and of high concern within the United States. Nontargets were defined as all confamilials potentially co-occurring with target taxa within the continental United States.

All assays are located within mitochondrial genes: cytochrome *c* oxidase subunit I (*COI*) for the crayfishes and fishes and cytochrome *b* (*cytb*) for the turtle (Table 1). We used GenBank (Sayers et al., 2019) to retrieve all target sequences available and one representative sequence from all nontargets in the database. Sequences were aligned using the ClustalW algorithm (Thompson et al., 1994) in MEGA X (Kumar et al., 2018). Primers and probes were identified from aligned sequences using a combination of DECIPHER (Wright et al., 2012, 2014), implemented in R 4.0.4 (R Core Team, 2021), and visual identification of appropriate gene regions. Primer and probe mismatches were counted iteratively using the BIOSTRINGS package (Pagès et al., 2020) and R script adapted from So et al. (2020) to design assays ranging in mismatch characteristics. Assays were predicted to be free of problematic hairpins and dimers, according to OligoAnalyzer (IDT). For tests of the full assays (primers and probe), primer concentrations were optimized to yield the lowest cycle threshold (C_t) and highest fluorescent signal (ΔR_n) possible (Table 1). Seven-level standard curves with six replicates were analysed to ensure the full assays were able to detect ≤10 DNA copies per reaction.

2.2 | Specificity testing

We paired our 10 assays with 82 templates (five target and 77 nontarget; Table S1) that created a relatively even mismatch distribution (Figure S1). Sequences were synthesized individually as gBlocks Gene Fragments (IDT), DNA concentration was quantified using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific) and aliquots were diluted to 5000 copies per reaction. We expect this DNA concentration to be greater than in the vast majority of environmental samples. For example, of the 1450 eDNA samples in the latest edition of the eDNAtlas (Young et al., 2018; www.fs.fed.us/rm/boise/AWAE/projects/the-aquatic-eDNAtlas-project.html) that were both positive and run with a standard curve, 1,426 (98%) were under 5,000 copies per reaction. Our estimates of specificity are therefore likely to be conservative, as is desirable when false negative errors (when a template is predicted to not amplify but does) are the primary concern.

Specificity testing was conducted for the primers alone (with SYBR Green dye) and the full assay (with TaqMan probes) using StepOnePlus and QuantStudio 3 Real-Time PCR Systems (Thermo Fisher Scientific). We used 262 assay–template combinations to test the primers, although six mismatch profiles were duplicated. Reactions included 10 μL of SYBR Green PCR Master Mix (Thermo Fisher Scientific), 300 nM of each primer and 4 μL (5000 copies) of DNA template, diluted to 20 μL per reaction with sterile water. Melting curves were assessed to ensure the intended amplicons were produced. Any templates that amplified were tested again with hydrolysis probes to increase specificity. (Templates that did not amplify with the primers were assumed to also not amplify after adding a probe, and were therefore not tested a second time.) We used 268 assay–template combinations to test the full assays (primers and probes), although one mismatch profile was duplicated. Reactions included 7.5 μL of TaqMan Environmental Master Mix 2.0 (2×; Thermo Fisher Scientific), optimized primer concentrations (Table 1), 250 nM of probe and 4 μL (5000 copies) of DNA template, diluted to 15 μL per reaction with sterile water. The thermal cycling profile for both SYBR Green and TaqMan reactions was 95°C for 10 min followed by 45 cycles of 95°C for 15 s and 60°C for 1 min. Fluorescence thresholds were held constant at 0.2 ΔR_n to avoid confounding mismatch effects with variation in software-defined threshold values. Reactions were run in triplicate and a test was considered positive if amplification was detected in at least one well.

2.3 | Model training and cross-validation

Our complete training data set was generated by pairing the results of specificity testing with assay-template mismatch information. Predictor variables included the total number of mismatches across both primers, the normalized difference in mismatch number between primers (|forward mismatches − reverse mismatches|/total mismatches), the proportion of mismatches within the first 5 bp on the 3′ ends of primers, the proportion of mismatches on

TABLE 1 Information on the assays used in this study, including the gene targeted, oligonucleotide sequences (F, forward primer; R, reverse primer; P, probe), oligonucleotide lengths, melting temperatures (T_m; as determined using PRIMER EXPRESS), primer concentrations optimized for TaqMan probe-based qPCR, the number of taxa tested via SYBR green and TaqMan chemistries, mean and standard deviation of total mismatches (MM) with templates used in the full-assay model, and references

Assay	Species	Gene	Oligonucleotide sequence (5′–3′)	Base-pairs	T_m (°C)	F:R (nM)	SYBR, TaqMan tests	Mean (*SD*) MM	References
Assays for model training and cross-validation									
FVR1	Virile crayfish (*Faxonious virilis*)	*COI*	F: AGAGGAATAGTCGAAAGAGGAGTAGGT	27	58.6	300: 900	29, 29	9.66 (3.04)	Parsley et al. (2020) (primers) and this study (probe)
			R: AGACACCCCTGCTAAATGTAACG	23	58.3				
			P: FAM–TGAACAGTGTATCCTCCTCT–MGB–NFQ	20	70.0				
FVR2	Virile crayfish (*F. virilis*)	*COI*	F: CCTGATATGGCTTTTCCTCGTATAAAT	27	59.9	900:900	29, 29	6.66 (2.69)	This study
			R: AAATACCTAAATCTACTGATGCCCCTG	27	60.0				
			P: FAM–ATCCTCCTCTTGCTTCT–MGB–NFQ	17	69.0				
PCLA	Red-swamp crayfish (*Procambarus clarkii*)	*COI*	F: CTTTGACTTTATTATTAACTAGGGGTATAGTTGAG	35	59.3	300:900	29, 29	9.14 (3.51)	This study
			R: AATAGCAGAAGCTAAAGGAGGATAAACA	28	59.3				
			P: FAM–AGTTGGAACAGGATGGAC–MGB–NFQ	18	70.0				
CID1	Grass carp (*Ctenopharyngodon idella*)	*COI*	F: GAGTTTCTGACTTCTACCCCCTTCT	25	58.5	100:900	32, 32	9.78 (2.70)	Wilson et al. (2014)
			R: CTGTTCACCCTGTTCCAGCTC	21	58.4				
			P: FAM–TCCTCCTACTATTAGCCTCT–MGB–NFQ	20	69.0				
CID2	Grass carp (*C. idella*)	*COI*	F: ATAGCATTCCCACGAATAAACAACA	25	59.6	900:600	58, 58	8.93 (3.49)	This study
			R: CTACGGATGCTCCTGCGTG	19	59.0				
			P: FAM–AGGGTGAACAGTTTACC–MGB–NFQ	17	70.0				
HYPO	Silver/bighead carp (*Hypophthalmichthys molitrix/nobilis*)	*COI*	F: TTTCCTTCTACTACTAGCCTCTTCTGGT	28	58.8	600:900	32, 32	8.84 (2.63)	This study
			R: AATTGTTAGGTCTACGGATGCTCCT	25	59.5				
			P: FAM–CGGAACAGGATGAACAG–MGB–NFQ	17	68.0				
TSC1	Pond slider (*Trachemys scripta*)	*Cytb*	F: CTCAGTAGACAACGCAACCCTG	22	58.5	900:900	15, 16	7.69 (3.58)	This study
			R: CAGTTTCATGTAGGAAAAGTAGGTGTACTAA	31	58.8				
			P: FAM–ACTTTATCCTACCATTTGC–MGB–NFQ	19	69.0				
TSC2	Pond slider (*T. scripta*)	*Cytb*	F: CCAGAATGATACTTTCTTTTCGCCTA	26	60.0	900:900	12, 13	3.54 (2.00)	This study
			R: TCGGAATTGGGTTGTTCGTT	20	58.9				
			P: FAM–AGTACTTGCCCTACTACTC–MGB–NFQ	19	70.0				
TSC3	Pond slider (*T. scripta*)	*Cytb*	F: TAGTACAATGAATCTGAGGCGGATT	25	59.1	900:900	12, 16	4.44 (2.49)	This study
			R: CCTGTTGGGTTGTTTGATCCA	21	59.3				
			P: FAM–TGCCATTATAGGTCTAACATT–MGB–NFQ	21	70.0				

TABLE 1 (Continued)

Assay	Species	Gene	Oligonucleotide sequence (5′–3′)	Base-pairs	T_m (°C)	F:R (nM)	SYBR, TaqMan tests	Mean (*SD*) MM	References
TSC4	Pond slider (*T. scripta*)	*Cytb*	F: CATACAAAGACCTCCTAGGCATCATT	26	59.8	600:900	14, 14	8.36 (4.44)	This study
			R: TATCTGGGTCTCCTAAGAGATTTGGA	26	59.5				
			P: FAM–ACTAACCCTCCTACTAACTC–MGB–NFQ	20	70.0				
Assays for model testing only									
FLNM	Flannelmouth sucker (*Catostomus latipinnis*)	*ND2*	F: CTACAAGAGCTGGCCAAGCAA	21	59.2	900:600	25, 25	10.12 (3.81)	Daniel Mason (in prep)
			R: CTCGTCATGGGGCGATAGAG	20	59.2				
			P: FAM–ATGCGATAACCCTGACCAT–MGB–NFQ	19	70.0				
SACS	Sacramento/Modoc suckers (*Catostomus occidentalis/microps*)	*ND2*	F: CAACTGACCCTCCTTGCCC	19	59.1	900:900	0, 25	7.93 (3.78)	Daniel Mason (in prep)
			R: AATAGGGCTCTTTGACCAGGC	21	58.3				
			P: FAM–CTAAAGCTCTCATCAGCTACTA–MGB–NFQ	22	70.0				
DSUK	Desert sucker (*Pantosteus clarkii*)	*Cytb*	F: TCTGAACCCTAGTTGCCGACATA	23	59.6	900:600	25, 25	9.29 (4.70)	Daniel Mason (in prep)
			R: GCTAGGATTAGGAATAGGGCAAAGTATAA	29	59.7				
			P: FAM–TGCCATCGGACAAAT–MGB–NFQ	15	69				
RGSU	Rio Grande sucker (*Pantosteus plebeius*)	*Cytb*	F: GCCACGGTGATTACTAACCTTTTG	24	59.8	300:600	25, 25	9.74 (2.56)	Caleb Dysthe (in prep)
			R: GCGGCGACTACAAATGGTAAT	21	57.8				
			P: FAM–TTGCCTACATAAGGGACTG–MGB–NFQ	19	69.0				
CAVI	Snake River bluehead sucker (*Pantosteus virescens*)	*ND1*	F: AGAGCCAATTCGTCCGTCTACA	22	59.7	100:300	21, 21	8.65 (3.75)	Caleb Dysthe (in prep)
			R: CCCAGAGCCCAGAATGGAGTAC	22	60.8				
			P: FAM–CCCATGCCTTATCC–MGB–NFQ	14	69.0				
RZBS	Razorback sucker (*Xyrauchen texanus*)	*ND4*	F: CACGACTGCACATAGCCTGC	20	59.1	600:900	23, 23	9.50 (3.18)	Daniel Mason (in prep)
			R: GTGGGGTAGATAGGGGGTCC	20	58.1				
			P: FAM–TGAACATCCGAAGCCG–MGB–NFQ	16	69.0				

the 3′ termini of primers, the total number of probe mismatches, the proportion of probe mismatches not within the first 5 bp on either end (i.e., in the centre), the proportion of primer and probe mismatches of each type (AA, AG/GA, AC/CA, TT, TG/GT, TC/CT, CC and GG), the mean primer length and probe length, the mean primer melting temperature (T_m), the difference in T_m between primers, and probe T_m (as determined using PRIMER EXPRES 3.0.1; Thermo Fisher Scientific). While random forest classifiers allow for correlations among predictor variables, we chose to limit multicollinearity by modelling mismatch characteristics as proportions of the total, rather than quantities. This is because model performance was essentially the same with both parameterizations (Table S2) and avoiding multicollinearity can produce less biased estimates of variable importance (as described below; Figure S2) (Toloşi & Lengauer, 2011). Furthermore, we did not include GC content in our models even though it can influence oligonucleotide behaviour (McDowell et al., 1998). GC content was strongly correlated with length ($r = -.88$ and $-.66$ for primers and probes, respectively) and therefore provided little additional information. In fact, performance was slightly higher with GC content removed (Table S2). This was also advantageous because removing uninformative parameters reduces model dimensionality (i.e., complexity) and therefore the risk of overfitting limited data sets.

These parameters were used to train two models: (i) a primer-only model with SYBR Green results and primer-specific predictor variables and (ii) a full-assay model with TaqMan probe-based results and all predictor variables. SYBR Green results had a balanced class distribution (124 positive and 138 negative) but TaqMan classes were relatively unbalanced (59 positive and 209 negative). TaqMan classes were therefore balanced using the synthetic minority oversampling technique (SMOTE) in the R package SMOTEFAMILY (Chawla et al., 2002), which synthesizes new instances of the minority class from existing data. This resulted in an effective distribution of 177 positive and 209 negative tests. Our results were very similar with and without oversampling (Table S3; Figure S3).

Both models were run using random forest classifiers in the R package RANDOMFOREST (Liaw & Wiener, 2002), implemented in CARET (Kuhn, 2008) with 1000 decision trees. We chose random forest because in preliminary analysis it outperformed other commonly used classification algorithms: logistic regression, support-vector machine and naïve Bayes (Table S3). Model performance was assessed using 10-fold cross-validation with 10 repeats. All values of *mtry* (the number of random predictor variables considered at each split in a given tree) between 1 and the total number of predictors were assessed and the values imparting the greatest recall (the proportion of positives that were correctly identified) were selected to minimize false negative error rates. Variable importance was estimated by calculating the difference in out-of-bag prediction accuracy between the full models and models with each variable permuted, normalized by the standard error among decision trees and scaled to have a maximum value of 100 (the default method used by CARET; Kuhn, 2008). However, variable importance was not estimated for predictors with an overabundance of zeros (3′ terminus primer mismatches and mismatches of different types) as they may cause resampling bias during cross-validation that skews results.

2.4 | Assignment threshold optimization

Binary classification models utilize a threshold value to assign class labels to assignment probabilities (typically 0.5). However, this threshold may be moved, or "optimized," to make certain types of errors less likely. In the context of eDNA, false negatives are typically more costly than false positives because incorrectly assigning a template to the "nonamplify" class (a false negative; FN) can be highly consequential, but incorrectly assigning a template to the "amplify" class (a false positive; FP) only requires additional in vitro testing. However, for false negative error risk to be removed entirely, the assignment threshold must be 0, which would necessitate in vitro testing of all taxa. Thus, some false negative error risk must be tolerated for our models to be useful. To account for this, we calculated the total cost of all FN:FP cost ratios at integers 1:1 and 100:1, where total cost = (FN rate × FN cost) + (FP rate × FP cost). For a given cost ratio, the threshold with the lowest total cost is optimal.

2.5 | Testing of independent assays

In addition to cross-validation, we evaluated model performance via six independent assays not used in model training (Table 1). These assays target fishes of conservation concern in the family Catostomidae: flannelmouth sucker (*Catostomus latipinnis*; assay FLNM), Sacramento and Modoc suckers (*C. occidentalis* and *C. microps*; SACS), Rio Grande sucker (*C. plebeius*; RGSU), desert sucker (*Pantosteus clarkii*; DSUK), Snake River bluehead sucker (*P. virescens*; CAVI) and razorback sucker (*Xyrauchen texanus*; RZBS). Nontargets included select confamilial species. All nontarget sequences available on GenBank were input into our models and each assay was tested in vitro to assess the accuracy of mean assignment probabilities (Table S4). Genomic DNA was extracted from tissues using DNeasy Blood and Tissue Kits (Qiagen) and diluted to 0.1 ng μl^{-1}, as determined using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific). Reactions were conducted as described above, except that they were performed singly and default fluorescence thresholds were used.

3 | RESULTS

3.1 | Assay specificity

We tested a total of 530 assay–template combinations in this study: 262 with primers alone and SYBR Green intercalating dye and 268 with primers and TaqMan hydrolysis probes. The 10 qPCR assays we used to train our models were highly variable in mismatch number, type and position, as well as oligonucleotide length and melting

temperature (Table 1). Total primer mismatches (across the forward and reverse primers) ranged from zero to 12 and probe mismatches from zero to six (Figure S1). Amplification was detected via SYBR Green chemistry with a maximum of 10 total primer mismatches and via TaqMan chemistry with a maximum of eight total primer mismatches, three probe mismatches, and nine mismatches in both primers and probes.

3.2 | Primer-only model performance

Random forest classifiers were successful at predicting qPCR camplification. In 10-fold cross-validation with 10 repeats, the primer-only model had an accuracy of 0.996 (95% confidence interval [CI]: 0.979–1.000; Figure 1a) and an area under the ROC curve (AUC) of 0.951 (95% CI: 0.943–0.959; Figure 2) in cross-validation. The mean assignment probability was 0.072 (range, 0.000–0.550) for negative tests and 0.926 (range, 0.572–1.000) for positive tests (Figure 3a). The one template that was incorrectly predicted had an assignment probability of 0.55. The optimal threshold was 0.50 for the 1:1 FN:FP cost ratio and dropped to a low of 0.02 by the 10:1 cost ratio (Figure 4). As such a low threshold would not result in a useful model, we limited thresholds to ≥0.2 and the optimal threshold dropped to 0.2 by the 15:1 cost ratio. Of the predictor variables that were well represented in the data set (without an overabundance of zeros), total primer mismatches had the most predictive power, followed by proportion of mismatches on the 3′ end (Figure 5). The remaining variables had relatively low importance. The model performed less well at predicting the specificity of six independent assays, with an accuracy of 0.924 across 119 tests (Figure 1c). Incorrect predictions had mean assignment probabilities between 0.17 and 0.70 (Figure 3a; Table S5). SYBR Green results from the SACS assay were not included due to ambiguous melting curves that precluded determination of specificity.

3.3 | Full-assay model performance

The full-assay model had an accuracy of 1.000 (95% CI: 0.991–1.000; Figure 1b) and an AUC of 0.992 (95% CI: 0.990–0.994; Figure 2) in cross-validation. The mean assignment probability was 0.079 (range, 0.000–0.463) for negative tests and 0.929 (range, 0.610–1.000) for positive tests (Figure 3b). The optimal assignment threshold was 0.54 for the 1:1 FN:FP cost ratio and dropped to a low of 0.28 by the 25:1 cost ratio (Figure 4). The most important predictor variable was total primer mismatches, followed by probe mismatches, proportion of primer mismatches on the 3′ ends and then proportion of probe mismatches in the centre (Figure 5). The remaining variables had relatively low importance. The model performed nearly as well at predicting the specificity of six independent assays, with an accuracy of 0.965 across 144 tests (Figure 1d). Incorrect predictions had mean assignment probabilities between 0.36 and 0.55 (Figure 3b; Table S5).

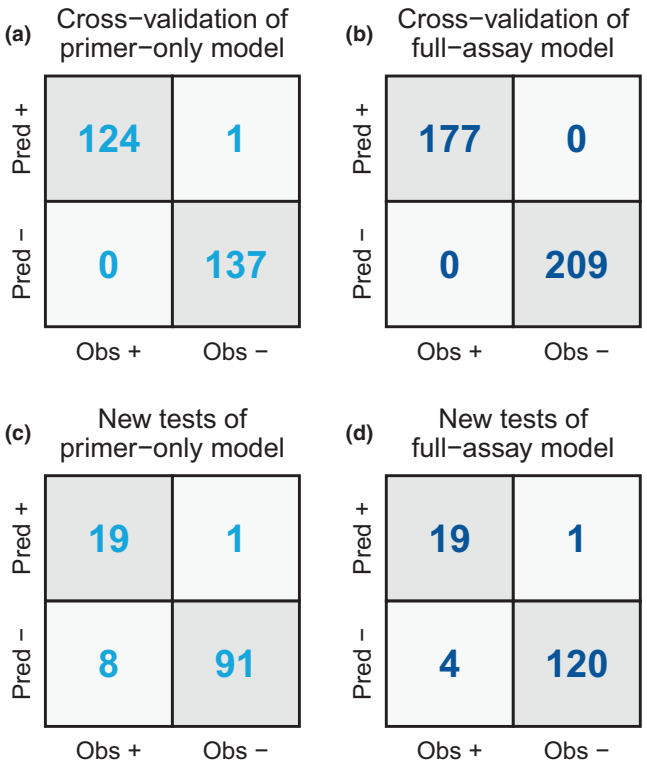

FIGURE 1 Confusion matrices from cross-validation of the primer-only model (SYBR green results; a) and full-assay model (TaqMan probe-based results; b), and from six independent assays tested with the primer-only model (c) and full-assay model (d). Pred +, test predicted to be positive; Pred −, test predicted to be negative; Obs +, test observed to be positive; Obs −, test observed to be negative

4 | DISCUSSION

The goal of our study was to more effectively evaluate qPCR assay specificity in silico and thereby minimize the in vitro testing bottleneck experienced by many assay developers. We did not detect amplification beyond 10 total primer mismatches via SYBR Green chemistry or eight total primer mismatches, three probe mismatches, and nine mismatches in both primers and probes via TaqMan chemistry. These results corroborate those of Wilcox et al. (2013) and Döring et al. (2019), who noted only rare or slight amplification at 10 and 12 primer mismatches, respectively. As the number of mismatches increased, the likelihood of amplification decreased, as found in other studies (Döring et al., 2019; Lefever et al., 2013; So et al., 2020; Stadhouders et al., 2010; Süß et al., 2009; Wilcox et al., 2013; Wright et al., 2014; Yilmaz et al., 2012). However, it is notable that primer mismatches were more important than probe mismatches in the full-assay model (Figure 5). This seems to contradict evidence that probe mismatches have greater impacts then primer mismatches (e.g., Süß et al., 2009), but aligns with observations of Wilcox et al. (2013). Different conclusions may relate to whether single or multiple mismatches are compared, with single probe mismatches being more impactful, but primers often having greater numbers of mismatches with a stronger collective impact. In

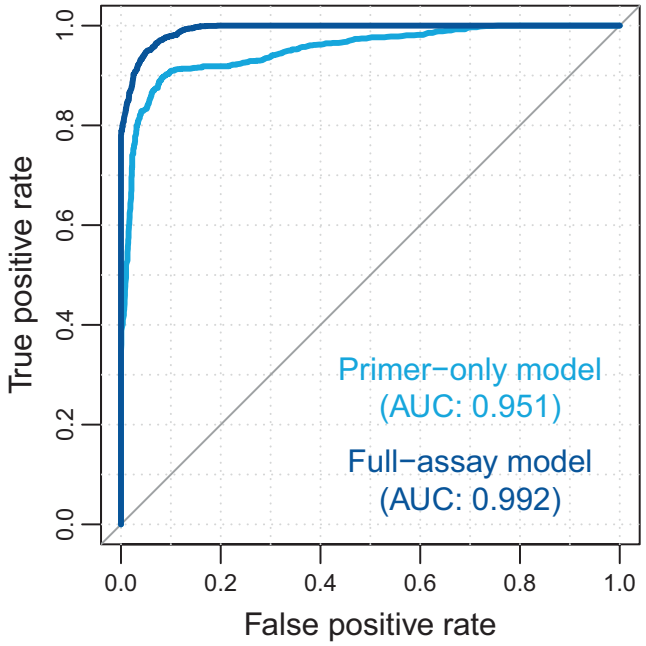

FIGURE 2 Receiver operating characteristic (ROC) curves from the primer-only model (SYBR green results) and full-assay model (TaqMan probe-based results), along with estimates of area under the ROC curve (AUC). Results are from 10-fold cross-validation of random forest classifiers with 10 repeats

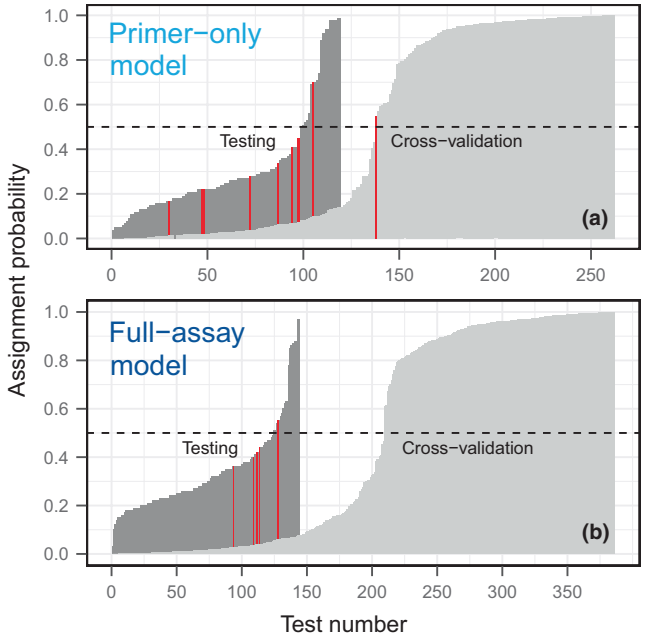

FIGURE 3 Assignment probabilities from the primer-only model (SYBR green results; a) and full-assay model (TaqMan probe-based results; b). Estimates in light grey are from 10-fold cross-validation of random forest classifiers with 10 repeats; estimates in dark grey are from six independent assays tested using genomic DNA. Incorrect predictions are highlighted in red. Horizontal dashed lines separate the "nonamplify" class (<0.5) from the "amplify" class (>0.5)

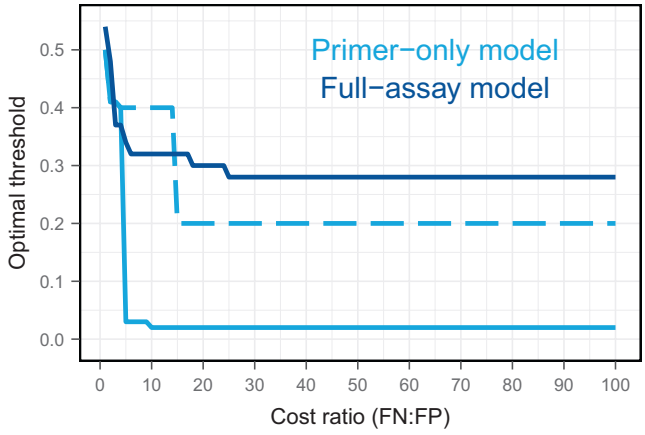

FIGURE 4 Optimal assignment thresholds from the primer-only model (SYBR green results) and full-assay model (TaqMan probe-based results) for a range of cost ratios. The dashed line only includes thresholds ≥0.2. For eDNA assays, the cost of a false negative (FN; a template predicted to not amplify that does) is typically greater than the cost of a false positive (FP; a template predicted to amplify that does not). If false negatives are deemed particularly risky for a given assay, the cost ratio will be high and the optimal threshold low, necessitating additional in vitro testing

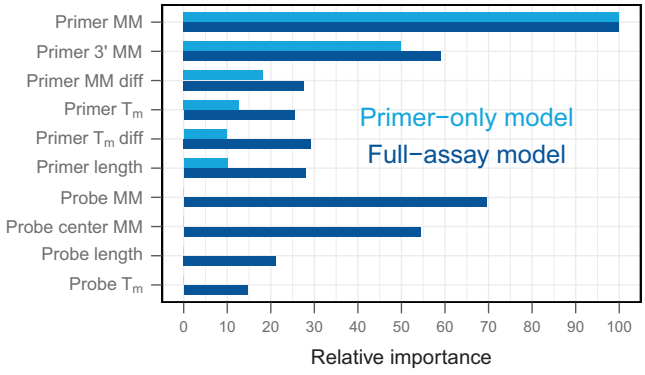

FIGURE 5 Relative importance of predictor variables during cross-validation of the primer-only model (SYBR green results) and full-assay model (TaqMan probe-based results). Importance was determined by comparing the accuracy of the final model to that of models with each variable permuted. Only well-represented predictor variables (without an overabundance of zeros) were estimated to avoid resampling bias. MM, mismatches; T_m, melting temperature

this study, primers clearly had more mismatches (mean of 5.6) than did probes (mean of 2.6).

Mismatch information can be utilized to predict qPCR amplification by pairing empirical results with supervised machine learning classifiers. We trained two random forest classification models: one for the primers alone (SYBR Green results) and one for the full assays (TaqMan probe-based results). In cross-validation, these models were 99.6% and 100% accurate, respectively (Figure 1a,b). In tests of six assays not used in model training, the primer-only model was 92.4% accurate and the full-assay model was 96.5% accurate (Figure 1c,d). These results show improved performance over existing in silico assessment techniques. For example, PRIMER3 (Untergasser et al., 2012)

can be highly accurate at predicting whether candidate primers will amplify the intended sequence, but may fall short when primer specificity is essential. PRIMER-BLAST (Ye et al., 2012) combines PRIMER3 design elements with a BLAST search to assess primer specificity, as determined by the user-defined number of allowable mismatches. However, So et al. (2020) evaluated PRIMER-BLAST predictions for 788 nontarget templates and noted just 67% accuracy under the most stringent specificity settings. This degree of error is unsuitable for the majority of eDNA assays and highly concerning given the ubiquity of PRIMER-BLAST in most eDNA assay development pipelines.

Several studies have utilized machine learning to increase primer design capabilities. These include classifying primers as "good" or "bad" at amplifying the intended sequence (Ashlock et al., 2002), filtering out inefficient primers to increase the processing speed of thermodynamic models (Mann et al., 2009), designing primers that are specific according to user-defined mismatch tolerances (Chuang et al., 2012), and locating diagnostic SARS-CoV-2 sequences that contain effective primer binding sites (Lopez–Rincon et al., 2021). A few studies have used machine learning to predict amplification in particular (and specificity, by extension). Models by Kayama et al. (2021) and Cordaro et al. (2021) had 70% and 81% accuracy, respectively. Better performance was achieved by Döring et al. (2019), whose model produced an AUC of 0.953—slightly higher than our primer-only model, but lower than our full-assay model. To our knowledge, all uses of machine learning to predict amplification have relied on training data from conventional PCR and gel electrophoresis (or in one study, CRISPR-Cas13a; Metsky et al., 2021). Specificity in conventional PCR cannot be directly translated to specificity in qPCR because the latter is much more sensitive (sensu So et al., 2020). In other words, lack of a discernible band on a gel is very different than lack of a qPCR amplification curve; an assay specific in conventional PCR testing may not be specific in qPCR testing. These studies also do not account for hydrolysis probes—which are becoming ubiquitous in eDNA assays—and do not explicitly test DNA at the low concentrations typically found in environmental samples. Our training data set includes each of these factors (qPCR analysis, hydrolysis probes and relatively low DNA concentrations), giving our random forest classifiers clear advantages over existing methods of evaluating eDNA assay performance.

Although we believe our models have great utility, their performance inevitably reflects the particular assays we tested and the PCR conditions we used. For example, different reaction chemistries may produce different results; Stadhouders et al. (2010) amplified mismatched templates using three different master mixes and found considerable variation in mismatch effects. Another important factor is thermal cycling profile, particularly the annealing temperature and number of cycles. For example, we used 45-cycle reactions to produce our training data set. If we tested our models using 40-cycle reactions we may overestimate specificity, but with 50-cycle reactions we may underestimate specificity. However, to our knowledge, no models account for all chemical and thermal cycling parameters in estimates of specificity. We recommend that users of our tool and other predictive models either (i) use the PCR conditions under which the models were developed, or (ii) reassess or retrain the models using their PCR conditions of choice. Furthermore, our training assays had melting temperatures between ~58-60C for primers and ~68-70C for probes, and forward primers and probes annealed to the antisense strand and reverse primers to the sense strand. These characteristics should also be incorporated for optimal performance.

Machine learning algorithms are trained on the results of empirical tests and overfitting can be an issue. Compared to 99.6% accuracy of the primer-only model and 100% accuracy of the full-assay model in cross-validation, accuracy was 92.4% and 96.5%, respectively, when the models were tested against six independent assays. This suggests that our models were minimally overfit. The decrease in accuracy may have resulted in part from imperfect sequence information in tests of independent assays; we did not sequence the templates and instead relied on publicly available sequences, as we would in practice. If confirmed sequences were input into our models their accuracy may have been higher. Furthermore, the primer-only model was built using SYBR Green reaction chemistry. As SYBR Green dye binds indiscriminately to double-stranded DNA, it is prone to fluorescing nonspecific targets (e.g., primer dimers), which are sometimes difficult to differentiate from the intended amplicon. This uncertainty may have caused us to identify some templates as amplifying when they in fact did not.

To account for model uncertainty while also designing reliably specific assays, we propose lowering the probability threshold used to assign class labels based on the relative costs of false negative (FN) and false positive (FP) errors. False negative errors (when a template is predicted to not amplify but does) are typically more costly for eDNA practitioners. Cross-validation of our full-assay model shows the optimal threshold to fall rapidly as the cost of false negatives (the FN:FP ratio) increases, levelling out around 0.3 (Figure 4). When independent assays were tested, the false negative error rate fell to zero by 0.35 (Figure 3b). A threshold around 0.3 may therefore be a good starting place for many assay developers. In this case, only templates with assignment probabilities ≥0.3 would need in vitro testing. It is important to note that assignment probabilities do not represent the likelihood of amplifying, but the likelihood of being assigned to a particular class. In other words, an assignment probability of 0.3 does not mean that the template will amplify in 30% of reactions; our results indicate that such a template would be very unlikely to amplify no matter how many times it was tested.

Optimal thresholds may change depending on the application. An assay targeting a cryptic species that is protected and consequential for management may have a high FN:FP cost ratio (e.g., 100:1), and therefore a low threshold (e.g., 0.2). Conversely, an assay targeting a conspicuous species that is unprotected and less consequential for management may have a low FN:FP ratio (e.g., 1:1), and therefore a high threshold (e.g., 0.5). Moving the threshold can refine in vitro testing needs. For example, we applied our full-assay model to a new candidate *Faxonius virilis* assay (FVR3; unpublished) and predicted the amplification of all 279 confamilials in the continental United States with publicly available sequences. An assignment threshold of 0.3 resulted in 45 species needing in vitro testing, whereas a threshold of 0.4 resulted in only 19 species needing in

vitro testing. However, given the underrepresentation of many taxa in reference databases, the actual testing demands are likely to be higher. For the FVR3 assay, an additional 120 nontarget confamilials do not have sequences. These would either need to be tested in vitro or the taxonomic or geographical scope of the assay would need to be revised, although it may be reasonable in some cases to forego testing of confamilial genera with fully resolved phylogenies and universally low assignment probabilities of sequenced members.

We conducted this study to inform assay specificity, but our models are equally informative for assay generality. Most eDNA assays fall within mitochondrial genes with relatively high mutation rates, and polymorphisms within target taxa can be common. The presence of target polymorphisms may disqualify assay regions that are otherwise highly diagnostic, causing developers to settle for less-specific assays. Some polymorphisms may be tolerated if they allow for increased specificity. But how many mismatches can be tolerated before an assay no longer amplifies its target? Users of our models can predict target amplification to increase confidence in assay generality and maximize design options. In tests of the full-assay model, all templates with assignment probabilities >0.55 amplified (Figure 3b), and thus an assignment probability around 0.7 may be a reasonable upper threshold to indicate assay generality. Other considerations include: (i) care should be taken to assess as many target sequences as possible from throughout the geographical range of assay application, and (ii) a high assignment probability does not necessarily mean strong amplification (i.e., a low C_t value); extensive in vitro testing is still recommended.

Both models may be accessed via our GitHub page (https://github.com/NationalGenomicsCenter/eDNAssay). Given the exceptional accuracy of our full-assay model and the ubiquity of TaqMan probe-based qPCR assays in eDNA monitoring, we have also made it available as eDNAssay, a user-friendly and highly accurate tool for in silico prediction of eDNA assay specificity (https://NationalGenomicsCenter.shinyapps.io/eDNAssay). Users must simply input aligned sequences, a brief metadata file and melting temperatures to output mismatch information and class assignment probabilities. In addition, eDNAssay includes features to help users optimize the assignment threshold—lowering the class assignment probability threshold from 0.5 to better hedge against false negative errors.

This study demonstrates the impressive ability of machine learning algorithms to predict qPCR cross-amplification. Future studies are needed to refine existing models and build new models for other applications. For example, models can be retrained to increase accuracy as additional specificity data become available. Likewise, data can be generated to train models for different PCR conditions (e.g., probe moieties, polymerases and thermal cycling profiles), templates (e.g., with base-pair insertions and deletions) and oligonucleotides (e.g., with degenerate bases or locked nucleic acids). The use of qPCR-based eDNA analysis has grown tremendously in recent years and several authors have called for universal and stringent standards for assay validation (e.g., Goldberg et al., 2016; Langlois et al., 2021; Thalinger et al., 2021). Such standards are necessary for the ongoing utility of eDNA assays and the respect eDNA analysis has garnered as a wildlife monitoring tool. We believe and our predictive models have immense utility for helping to ensure these standards are met.

AUTHOR CONTRIBUTIONS

All authors contributed to study design, results interpretation and writing the manuscript. John A. Kronenberger gathered empirical data and performed statistical analyses.

ACKNOWLEDGEMENTS

We thank Caleb Dysthe (formerly of the National Genomics Center for Wildlife and Fish Conservation) for the CAVI and RGSU assays, Amanda Mast (National Genomics Center for Fish and Wildlife Conservation) for help with specificity testing, and Chris Rees (U.S. Fish and Wildlife Service, Northeast Fishery Center) for insightful feedback. This work was funded by the Department of Defence through their Environmental Security Technology Certification Program (ESTCP) under Project RC21-5121.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

Raw data, script for estimating SYBR Green and TaqMan probe-based assay specificity, and the script behind eDNAssay are available on GitHub (https://github.com/NationalGenomicsCenter/eDNAssay) and Dryad (doi: https://doi.org/10.5061/dryad.cnp5hqc74).

ORCID

John A. Kronenberger https://orcid.org/0000-0003-3588-7572
Taylor M. Wilcox https://orcid.org/0000-0003-3341-7374
Daniel H. Mason http://orcid.org/0000-0002-2976-072X
Thomas W. Franklin https://orcid.org/0000-0003-1658-4664
Kevin S. McKelvey https://orcid.org/0000-0002-1399-3166
Michael K. Young https://orcid.org/0000-0002-0191-6112
Michael K. Schwartz https://orcid.org/0000-0003-3521-3367

REFERENCES

Ashlock, D., Wittrock, A., & Wen, T. J. (2002). Training finite state machines to improve PCR primer design. In *Proceedings of the 2002 congress on evolutionary computation. CEC'02 (cat. No. 02TH8600)* (Vol. *1*, pp. 13–18).

Bru, D., Martin-Laurent, F., & Philippot, L. (2008). Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Applied and Environmental Microbiology*, *74*(5), 1660–1663. https://doi.org/10.1128/AEM.02403-07

Bylemans, J., Gleeson, D. M., Duncan, R. P., Hardy, C. M., & Furlan, E. M. (2019). A performance evaluation of targeted eDNA and eDNA metabarcoding analyses for freshwater fishes. *Environmental DNA*, *1*(4), 402–414. https://doi.org/10.1002/edn3.41

Carim, K. J., Christianson, K. R., McKelvey, K. M., Pate, W. M., Silver, D. B., Johnson, B. M., Galloway, B. T., Young, M. K., & Schwartz, M. K. (2016). Environmental DNA marker development with sparse biological information: A case study on opossum shrimp (*Mysis diluviana*). *PLoS One*, *11*(8), e0161664. https://doi.org/10.1371/journal.pone.0161664

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chuang, L. Y., Cheng, Y. H., & Yang, C. H. (2012). URPD: A specific product primer design tool. *BMC Research Notes*, *5*(1), 1–9. https://doi.org/10.1186/1756-0500-5-306

Cordaro, N. J., Kavran, A. J., Smallegan, M., Palacio, M., Lammer, N., Brant, T. S., DuMont, V., Garcia, N. D., Miller, S., Jourabchi, T., & Sawyer, S. L. (2021). Optimizing polymerase chain reaction (PCR) using machine learning. *bioRxiv*. https://doi.org/10.1101/2021.08.12.455589

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, *18*(6), 1381–1391. https://doi.org/10.1111/1755-0998.12926

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Döring, M., Kreer, C., Lehnen, N., Klein, F., & Pfeifer, N. (2019). Modeling the amplification of immunoglobulins through machine learning on sequence–specific features. *Scientific Reports*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-47173-w

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*(4), 423–425. https://doi.org/10.1098/rsbl.2008.0118

Fountain-Jones, N. M., Smith, M. L., & Austerlitz, F. (2021). Machine learning in molecular ecology [Special issue]. *Molecular Ecology*, *21*(8), 2589–2597.

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. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307. https://doi.org/10.1111/2041-210X.12595

Harper, L. R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H. C., Gough, K. C., Lewis, E., Adams, I. P., Brotherton, P., Phillips, S., & Hänfling, B. (2018). Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (*Triturus cristatus*). *Ecology and Evolution*, *8*(12), 6330–6341. https://doi.org/10.1002/ece3.4013

Kayama, K., Kanno, M., Chisaki, N., Tanaka, M., Yao, R., Hanazono, K., Camer, G. A., & Endoh, D. (2021). Prediction of PCR amplification from primer and template sequences using recurrent neural network. *Scientific Reports*, *11*(1), 1–24. https://doi.org/10.1038/s41598-021-86357-1

Kuhn, M. (2008). Building predictive models in *R* using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Kutyavin, I. V., Afonina, I. A., Mills, A., Gorn, V. V., Lukhtanov, E. A., Belousov, E. S., Singer, M. J., Walburger, D. K., Lokhov, S. G., Gall, A. A., & Dempcy, R. (2000). 3′-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Research*, *28*(2), 655–661. https://doi.org/10.1093/nar/28.2.655

Kwok, S., Kellogg, D. E., McKinney, N., Spasic, D., Goda, L., Levenson, C., & Sninsky, J. J. (1990). Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nucleic Acids Research*, *18*(4), 999–1005. https://doi.org/10.1093/nar/18.4.999

Langlois, V. S., Allison, M. J., Bergman, L. C., To, T. A., & Helbing, C. C. (2021). The need for robust qPCR-based eDNA detection assays in environmental monitoring and species inventories. *Environmental DNA*, *3*(3), 519–527. https://doi.org/10.1002/edn3.164

Lefever, S., Pattyn, F., Hellemans, J., & Vandesompele, J. (2013). Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays. *Clinical Chemistry*, *59*(10), 1470–1480. https://doi.org/10.1373/clinchem.2013.203653

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Lopez–Rincon, A., Tonda, A., Mendoza–Maldonado, L., Mulders, D. G., Molenkamp, R., Perez–Romero, C. A., Claassen, E., Garssen, J., & Kraneveld, A. D. (2021). Classification and specific primer design for accurate detection of SARS–CoV–2 using deep learning. *Scientific Reports*, *11*(1), 1–11. https://doi.org/10.1038/s41598-020-80363-5

Mann, T., Humbert, R., Dorschner, M., Stamatoyannopoulos, J., & Noble, W. S. (2009). A thermodynamic approach to PCR primer design. *Nucleic Acids Research*, *37*(13), e95. https://doi.org/10.1093/nar/gkp443

McDowell, D. G., Burns, N. A., & Parkes, H. C. (1998). Localised sequence regions possessing high melting temperatures prevent the amplification of a DNA mimic in competitive PCR. *Nucleic Acids Research*, *26*(14), 3340–3347. https://doi.org/10.1093/nar/26.14.3340

McKelvey, K. S., Young, M. K., Knotek, W. L., Carim, K. J., Wilcox, T. M., Padgett-Stewart, T. M., & Schwartz, M. K. (2016). Sampling large geographic areas for rare species using environmental DNA: A study of bull trout *Salvelinus confluentus* occupancy in western Montana. *Journal of Fish Biology*, *88*(3), 1215–1222. https://doi.org/10.1111/jfb.12863

Metsky, H. C., Welch, N. L., Haradhvala, N. J., Rumker, L., Zhang, Y. B., Pillai, P. P., Yang, D. K., Ackerman, C. M., Weller, J., Blainey, P. C., & Myhrvold, C. (2021). Designing viral diagnostics with model–based optimization. *bioRxiv*. https://doi.org/10.1101/2020.11.28.401877

Nathan, L. M., Simmons, M., Wegleitner, B. J., Jerde, C. L., & Mahon, A. R. (2014). Quantifying environmental DNA signals for aquatic invasive species across multiple detection platforms. *Environmental Science & Technology*, *48*(21), 12800–12806. https://doi.org/10.1021/es5034052

Noguera, D. R., Wright, E. S., Camejo, P., & Yilmaz, L. S. (2014). Mathematical tools to optimize the design of oligonucleotide probes and primers. *Applied Microbiology and Biotechnology*, *98*(23), 9595–9608. https://doi.org/10.1007/s00253-014-6165-x

Ogram, A., Sayler, G. S., & Barkay, T. (1987). The extraction and purification of microbial DNA from sediments. *Journal of Microbiological Methods*, *7*(2–3), 57–66. https://doi.org/10.1016/0167-7012(87)90025-x

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy S. (2020). Biostrings: Efficient manipulation of biological strings. R package version 2.58.0. https://bioconductor.org/packages/Biostrings

Parsley, M. B., Torres, M. L., Banerjee, S., Tobias, Z., Goldberg, C. S., Murphy, M. A., & Mims, M. C. (2020). Local and landscape factors influence functional connectivity in an amphibian metapopulation: Evidence from multiple lines of genetic inquiry. *Landscape Ecology*, *35*, 319–335. https://doi.org/10.1007/s10980-019-00948-y

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rees, H. C., Maddison, B. C., Middleditch, D. J., Patmore, J. R., & Gough, K. C. (2014). The detection of aquatic animal species using environmental DNA—A review of eDNA as a survey tool in ecology. *Journal of Applied Ecology*, *51*(5), 1450–1459. https://doi.org/10.1111/1365-2664.12306

Rychlik, W. J. S. W., Spencer, W. J., & Rhoads, R. E. (1990). Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Research*, *18*(21), 6409–6412. https://doi.org/10.1093/nar/18.21.6409

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch–Mizrachi, I. (2019). GenBank. *Nucleic Acids Research*, *47*(D1), D94–D99. https://doi.org/10.1093/nar/gky989

Simmons, M., Tucker, A., Chadderton, W. L., Jerde, C. L., & Mahon, A. R. (2016). Active and passive environmental DNA surveillance of

aquatic invasive species. *Canadian Journal of Fisheries and Aquatic Sciences*, *73*(1), 76–83. https://doi.org/10.1139/cjfas-2015-0262

So, K. Y. K., Fong, J. J., Lam, I. P. Y., & Dudgeon, D. (2020). Pitfalls during in silico prediction of primer specificity for eDNA surveillance. *Ecosphere*, *11*(7), e03193. https://doi.org/10.1002/ecs2.3193

Somerville, C. C., Knight, I. T., Straube, W. L., & Colwell, R. R. (1989). Simple, rapid method for direct isolation of nucleic acids from aquatic environments. *Applied and Environmental Microbiology*, *55*(3), 548–554. https://doi.org/10.1128/aem.55.3.548-554.1989

Stadhouders, R., Pas, S. D., Anber, J., Voermans, J., Mes, T. H., & Schutten, M. (2010). The effect of primer–template mismatches on the detection and quantification of nucleic acids using the 5′ nuclease assay. *The Journal of Molecular Diagnostics*, *12*(1), 109–117. https://doi.org/10.2353/jmoldx.2010.090035

Süß, B., Flekna, G., Wagner, M., & Hein, I. (2009). Studying the effect of single mismatches in primer and probe binding regions on amplification curves and quantification in real-time PCR. *Journal of Microbiological Methods*, *76*(3), 316–319. https://doi.org/10.1016/j.mimet.2008.12.003

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. https://doi.org/10.1002/edn3.189

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Thomsen, P. F., Kielgast, J. O. S., Iversen, L. L., Carsten, W., Rasmussen, M., Gilbert, M. T. P., Orlando, L., & Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*(11), 2565–2573. https://doi.org/10.1111/j.1365-294X.2011.05418.x

Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), 1986–1994. https://doi.org/10.1093/bioinformatics/btr300

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—New capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115. https://doi.org/10.1093/nar/gks596

Wang, K., Li, H., Xu, Y., Shao, Q., Yi, J., Wang, R., Cai, W., Hang, X., Zhang, C., Cai, H., & Qu, W. (2019). MFEprimer-3.0: Quality control for PCR primers. *Nucleic Acids Research*, *47*(W1), W610–W613. https://doi.org/10.1093/nar/gkz351

Wilcox, T. M., Carim, K. J., McKelvey, K. S., Young, M. K., & Schwartz, M. K. (2015). The dual challenges of generality and specificity when developing environmental DNA markers for species and subspecies of *Oncorhynchus*. *PLoS One*, *10*(11), e0142008. https://doi.org/10.1371/journal.pone.0142008

Wilcox, T. M., McKelvey, K. S., Young, M. K., Engkjer, C., Lance, R. F., Lahr, A., Eby, L. A., & Schwartz, M. K. (2020). Parallel, targeted analysis of environmental samples via high-throughput quantitative PCR. *Environmental DNA*, *2*(4), 544–553. https://doi.org/10.1002/edn3.80

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. https://doi.org/10.1371/journal.pone.0059520

Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., Bunce, M., Wiuf, C., Gilichinsky, D. A., & Cooper, A. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, *300*(5620), 791–795. https://doi.org/10.1126/science.1084114

Williams, M. A., O'Grady, J., Ball, B., Carlsson, J., de Eyto, E., McGinnity, P., Jennings, E., Regan, F., & Parle-McDermott, A. (2019). The application of CRISPR-Cas for single species identification from environmental DNA. *Molecular Ecology Resources*, *19*(5), 1106–1114. https://doi.org/10.1111/1755-0998.13045

Williams, M. R., Stedtfeld, R. D., Engle, C., Salach, P., Fakher, U., Stedtfeld, T., Dreelin, E., Stevenson, R. J., Latimore, J., & Hashsham, S. A. (2017). Isothermal amplification of environmental DNA (eDNA) for direct field-based monitoring and laboratory confirmation of *Dreissena* sp. *PLoS One*, *12*(10), e0186462. https://doi.org/10.1371/journal.pone.0186462

Wilson, C., Wright, E., Bronnenhuber, J., MacDonald, F., Belore, M., & Locke, B. (2014). Tracking ghosts: Combined electrofishing and environmental DNA surveillance efforts for Asian carps in Ontario waters of Lake Erie. *Management of Biological Invasions*, *5*(3), 225–231. https://doi.org/10.3391/mbi.2014.5.3.05

Wood, S. A., Pochon, X., Laroche, O., von Ammon, U., Adamson, J., & Zaiko, A. (2019). A comparison of droplet digital polymerase chain reaction (PCR), quantitative PCR and metabarcoding for species-specific detection in environmental DNA. *Molecular Ecology Resources*, *19*(6), 1407–1419. https://doi.org/10.1111/1755-0998.13055

Wright, E. S., Yilmaz, L. S., & Noguera, D. R. (2012). DECIPHER, a search–based approach to chimera identification for 16S rRNA sequences. *Applied and Environmental Microbiology*, *78*(3), 717–725. https://doi.org/10.1128/AEM.06516-11

Wright, E. S., Yilmaz, L. S., Ram, S., Gasser, J. M., Harrington, G. W., & Noguera, D. R. (2014). Exploiting extension bias in polymerase chain reaction to improve primer specificity in ensembles of nearly identical DNA templates. *Environmental Microbiology*, *16*(5), 1354–1365. https://doi.org/10.1111/1462-2920.12259

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: A tool to design target–specific primers for polymerase chain reaction. *BMC Bioinformatics*, *13*(1), 1–11. https://doi.org/10.1186/1471-2105-13-134

Yilmaz, L. S., Loy, A., Wright, E. S., Wagner, M., & Noguera, D. R. (2012). Modeling formamide denaturation of probe-target hybrids for improved microarray probe design in microbial diagnostics. *PLoS One*, *7*(8), e43862. https://doi.org/10.1371/journal.pone.0043862

Young, M. K., Isaak, D. J., Schwartz, M. K., McKelvey, K. S., Nagel, D. E., Franklin, T. W., Greaves, S. E., Dysthe, J. C., Pilgrim, K. L., Chandler, G. L., Wollrab, S. P., Carim, K. J., Wilcox, T. M., Parkes-Payne, S. L., & Horan, D. L. (2018). *Species occurrence data from the aquatic eDNAtlas database*. Forest Service Research Data Archive. Updated 30 June 2021. https://doi.org/10.2737/RDS-2018-0010

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: 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*, *00*, 1–12. https://doi.org/10.1111/1755-0998.13681