Received: 23 February 2023 | Revised: 21 December 2023 | Accepted: 15 January 2024

DOI: 10.1111/1755-0998.13932

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

The unknown unknown: A framework for assessing environmental DNA assay specificity against unsampled taxa

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

USDA Forest Service, Rocky Mountain Research Station, National Genomics Center for Wildlife and Fish Conservation, Forestry Sciences Laboratory, Missoula, Montana, USA

Correspondence
Taylor M. Wilcox, USDA Forest Service, Rocky Mountain Research Station, National Genomics Center for Wildlife and Fish Conservation, Forestry Sciences Laboratory, 800 E Beckwith, Missoula, Montana, USA.
Email: taylor.wilcox@usda.gov

Funding information
Environmental Security Technology Certification Program, Grant/Award Number: RC21-5121

Handling Editor: Paula Arribas

Abstract

Taxon-specific quantitative PCR (qPCR) assays are commonly used for environmental DNA sampling-based inference of animal presence. These assays require thorough validation to ensure that amplification truly indicates detection of the target taxon, but a thorough validation is difficult when there are potentially many non-target taxa, some of which may have incomplete taxonomies. Here, we use a previously published, quantitative model of cross-amplification risk to describe a framework for assessing qPCR assay specificity when there is missing information and it is not possible to assess assay specificity for each individual non-target confamilial. In this framework, we predict assay specificity against unsampled taxa (non-target taxa without sequence data available) using the sequence information that is available for other confamilials. We demonstrate this framework using four case study assays for: (1) An endemic, freshwater arthropod (meltwater stonefly; *Lednia tumana*), (2) a globally distributed, marine ascidian (*Didemnum perlucidum*), (3) a continentally distributed freshwater crustacean (virile crayfish; *Faxonius virilis*, *deanae* and *nais* species complex) and (4) a globally distributed freshwater teleost (common carp; *Cyprinus carpio* and its close relative *C. rubrofuscus*). We tested the robustness of our approach to missing information by simulating application of our framework for all possible subsamples of 20—all non-target taxa. Our results suggest that the modelling framework results in estimates which are largely concordant with observed levels of cross-amplification risk using all available sequence data, even when there are high levels of data missingness. We explore potential limitations and extensions of this approach for assessing assay specificity and provide users with an R Markdown template for generating reproducible reports to support their own assay validation efforts.

KEYWORDS
eDNA, environmental DNA, qPCR

1 | INTRODUCTION

1.1 | Problem overview

> As we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns — the ones we don't know we don't know.
>
> Donald Rumsfeld, United States Secretary of Defense (2002)

Environmental DNA (eDNA) sampling—the inference of organism presence from genetic material in the environment—has been found to be particularly effective for the detection of rare aquatic animals. The most commonly used tool for the application of eDNA sampling is taxon-specific quantitative PCR (qPCR; Tsuji et al., 2019), which can identify DNA from a single taxon with high sensitivity and specificity (e.g. Bylemans et al., 2019; Harper et al., 2018; Wilcox et al., 2013).

Single-taxon eDNA assays require extensive testing to ensure that cross-amplification of non-target, co-occurring taxa is unlikely. We can determine how unlikely cross-amplification is using in silico (based on a comparison of sequences; e.g. Kronenberger et al., 2022) or in vitro (based on laboratory testing with tissue-based DNA extracts or synthetic gene fragments) approaches to specificity testing. There is, however, little guidance on the optimal taxonomic or geographic extent and intensity of specificity testing, beyond that 'multiple individuals' of the target and non-target species be tested from throughout the area of assay application (e.g. Langlois et al., 2021; Thalinger et al., 2021). Truly comprehensive specificity testing may be infeasible, particularly when the geographic scope is large and the target taxon's genus or family is speciose and incompletely represented in public sequence databases. Without relevant sequence data, specificity testing must proceed in vitro, posing serious bottlenecks for assay development due to the difficulty of obtaining DNA from rare and protected species. Testing against 'multiple individuals' of all unsequenced, non-target but closely related taxa to an extent that truly accounts for intraspecific haplotype diversity could require tissues from hundreds or even thousands of specimens—an order of magnitude more than is typical of in vitro eDNA assay validations in the literature (primary literature reviewed in Thalinger et al., 2021). These unsequenced, non-target but closely related taxa represent known unknowns in that their existence is known, but their DNA sequences are not. They can be tested against, albeit with difficulty. Efforts at comprehensive assay testing are further thwarted by the potential presence of undiscovered and undescribed taxa—unknown unknowns—which may be relatively common in groups with incomplete taxonomies. These taxa cannot be tested against at all.

We can (and do), however, make assumptions about assay specificity against unsequenced and even undiscovered non-target taxa. The most common of these assumptions is the implicit belief that if an assay is fully specific against all known closely related taxa, then it will likely be specific against more distantly related taxa (Figure 1), leading researchers to only test assay specificity against close relatives (e.g. Langlois et al., 2021; Thalinger et al., 2021). Although it is best practice to also conduct a broader in silico test of assay specificity against distantly related taxa (e.g. Primer-BLAST of the NCBI nucleotide database; Ye et al., 2012), this is again only a sample of the larger population of all biological sequences because of database incompleteness. The approach focusing on closely related taxa is largely effective for producing specific eDNA assays because (1) distantly related taxa generally have greater genetic differentiation from the target than those that are more closely related and (2) even in the absence of sequence data, taxonomy generally correlates with phylogenetic relatedness (but see Discussion Section 4.3: Taxonomic–phylogenetic discordance and uncertainty).

Here, we seek to better characterize assay specificity against unsampled taxa by stating this implicit model more explicitly: when taxonomic designations are largely concordant with phylogenetic relationships, we can predict the likelihood of assay specificity against unsampled taxa based on specificity against sampled taxa. Then, we demonstrate how to use this framework with examples involving qPCR assays from four varied freshwater and marine taxa. Finally, we provide a user-friendly R Markdown template for other researchers using this framework. What is different about this framework from previous tools is that we explicitly recognize that the suite of described and sequenced non-target taxa is just a sub-sample of the potentially problematic taxa on the landscape. Whereas previous reference database construction and in silico testing tools have focused on refining our assessment of known, sequenced non-target taxa, the framework here provides a way to make an inference about the larger, unknowable suite of potentially problematic taxa.

1.2 | Framework description

Imagine that there is some group (i.e. statistical population) of non-target confamilial taxa. This group includes all taxa (both described and undescribed) within a family that might occur within the geographic scope of assay application. This group of taxa varies in cross-amplification risk because some sequences are more similar to the target sequence (greater cross-amplification risk) and some sequences are less similar (lower cross-amplification risk). This variation in cross-amplification risk across the group has some distribution (Figure 1). We do not have access to sequence data for every member of this group, but we do have a sample of sequences in existing genetic repositories. As long as this sample is representative of the entire group (see Discussion for exploration of potential errors and solutions when this is not the case), we can estimate the probability density function that describes the distribution of some measure of cross-amplification risk across the entire group of non-target confamilials. If we can accurately

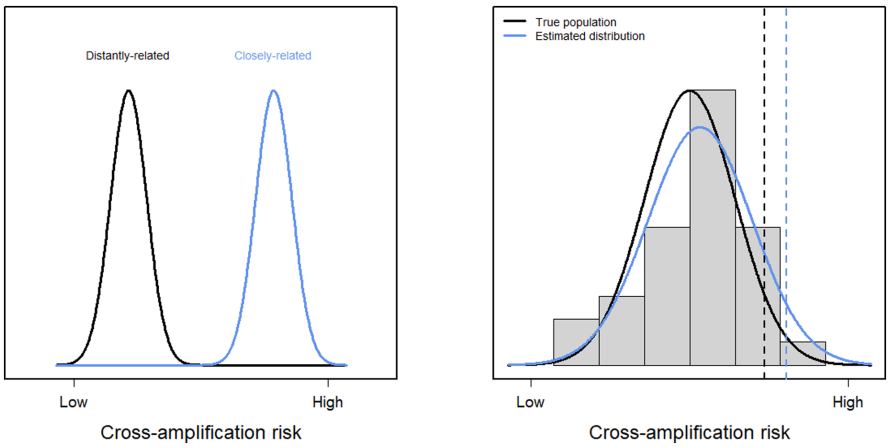

FIGURE 1 Conceptual figures for framework description. Left: There is an existing, implicit model where we understand that there tends to be low cross-amplification risk for distantly related non-target taxa and a higher risk for closely related non-target taxa (e.g. congenerics or confamilials to the target taxon). Thus, an assay which is specific when tested against closely related non-targets tends to be specific against distantly related non-target taxa as well. Right: The population of non-target taxa (e.g. potentially cooccurring confamilials) vary in cross-amplification risk (black curve). We do not have access to sequence data for the entire population of relevant non-target taxa, but we do have sequence data for a *sample* of this population (grey histogram). From this sample, we can estimate the distribution of cross-amplification risks across the larger population (blue curve). From the estimated probability density function, we can calculate a 95th percentile for cross-amplification risk (dotted blue line) which is a reasonable approximation of the true population (dotted black line).

estimate this probability distribution function, we can estimate group percentiles—that is, how likely an unsampled taxon—either unsequenced or entirely undescribed—is to exceed some threshold of cross-amplification risk (Figure 1).

1.3 | Analytical approach

Having laid out a conceptual groundwork, there are three elements required to put this approach into practice: (1) A measure of cross-amplification risk, (2) an appropriate probability density function to describe cross-amplification risk across non-target taxa group and (3) a method to assess the performance of this approach.

First, we need a way to quantify cross-amplification risk based on available DNA sequences. To do this, we need to use a model that relates assay-template sequence similarity to cross-amplification risk. Several models have been used for this purpose, including Primer-BLAST (Ye et al., 2012), PrimerMiner (Elbrecht & Leese, 2017) and rules of thumb such as 'having fewer than three primer base-pair mismatches'. Here, we use the eDNAssay model (Kronenberger et al., 2022)—a machine-learning classifier that has been trained on empirical data to predict qPCR cross-amplification. This tool assigns each template to the 'amplify' or 'non-amplify' class largely based on the type, number and position of base-pair mismatches with assay oligonucleotides. Any class assignment probability <0.5 is predicted to *not* cross-amplify. Assignment probabilities close to zero are almost certain not to cross-amplify. Classification of assay-template combinations with values close to 0.5 has greater uncertainty, although the model is highly accurate overall (96% across 610 independent specificity tests; see nationalgenomicscenter.shinyapps.io/eDNAssay/ for current documentation). Note that class assignment probability is *not* equivalent to cross-amplification probability (see Materials and Methods Section 2.4: Model Interpretation).

Second, we need a sample of sequenced taxa from a given taxonomic group (i.e. genus or family) to estimate a probability density function that represents the entire population of potential sequences. Multiple probability density functions could be potentially applied, but one natural fit is the beta distribution—a continuous probability distribution for values between zero and one, matching the bounded range of assignment probabilities from the eDNAssay model.

Finally, we need a method to assess the accuracy of this approach. The purpose of this framework is to make an inference about non-target taxa which are unsampled (i.e. are undescribed or described but unsequenced). Although, by definition, we cannot test our framework for these unsampled taxa, we can simulate data missingness by subsampling the data and comparing inferences with those using all available data.

1.4 | Specific case studies

We illustrate the proposed approach using four case study assays covering a wide range of animal taxonomies: (1) An endemic, freshwater arthropod (meltwater stonefly; *Lednia tumana*), (2) a globally distributed, marine ascidian (*Didemnum perlucidum*), (3) a continentally distributed freshwater crustacean (virile crayfish; *Faxonius virilis*, *deanae* and *nais* species complex) and (4) a globally distributed freshwater teleost (common carp; *Cyprinus carpio* and its close relative *C.rubrofuscus*). These taxa represent a range of potential assay validation situations with 0–24 congenerics and 427–3114 non-target confamilials described, of which 33.3%–90.7% are missing relevant sequence data.

2 | MATERIALS AND METHODS

2.1 | Molecular assays

The assay for the freshwater arthropod *Lednia tumana* was based on an alignment of 1509 Nemouridae *COI* sequences from 71 taxa (328 sequences from taxa which were assigned only to genus; Hutchins et al., 2023). This assay was designed to detect a narrowly endemic stonefly species in high-elevation streams in the North American Rocky Mountains. This assay was not designed with an MGB-binding moiety on the hydrolysis probe, but we assumed inclusion of this moiety for the purposes of assessment with eDNAssay. The assay for the ascidian *Didemnum perlucidum* (Didemnidae) was based on 33 *COI* sequences from non-target *Didemnum* taxa (Simpson et al., 2017) and designed to detect a widely distributed invasive species at global scales. The assay for the crustacean virile crayfish (*Faxonius virilis*, *deanae* and *nais* species complex; Cambaridae) was based on an alignment of 343 Cambaridae *COI* sequences from 287 taxa known to occur within the United States (Appendix S1). The purpose of this original assay design was to detect an invasive species within freshwaters of North America. The virile crayfish exists within species complex encompassing *F. virilis*, *F. deanae* and *F. nais* (Mathews et al., 2008; Filipová et al., 2010), so we considered each of these species as targets for this assessment. We did not consider sequences assigned to *F. punctimanus* because this taxon is difficult to identify morphologically and can hybridize with virile crayfish. Thus, some of the *F. punctimanus* mitochondrial sequences on GenBank with high eDNAssay assignment probabilities may have come from misidentified or hybridized individuals (Rozansky et al., 2021). The assay for the teleost *Cyprinus carpio* was based on an alignment of 1001 Cyprinidae and Leuciscidae *cytb* sequences from 268 taxa known to occur within the United States (Appendix S1). While common carp is placed in Cyprinidae, assay design is also considered Leuciscidae because it represents North American minnows which were classified under Cyprinidae until recently (Schönhuth et al., 2018). This assay was also designed to detect an invasive species within freshwater habitats of North America. It also targets *Cyprinus rubrofuscus*, a species that was previously considered a subspecies of common carp (*C. carpio haematopterus*; see Fricke et al., 2019 for revision) and is very similar at *COI* and *cytb* mitochondrial genes.

2.2 | Specificity data

For each assay, we built a list of all extant confamilial taxa based on a search of the Catalogue of Life (searched September 2023; Banki, 2022). Once species lists were assembled, we used a custom R (R Core Team, 2020) script based on the package *rentrez* (Winter, 2017) to pull all available sequence data from GenBank (Benson et al., 2015) for these species lists (September 2023). We aligned non-target sequences with our assay oligonucleotide sequences in MEGA7 (Kumar et al., 2016) or Clustal Omega (Sievers et al., 2011) using the ClustalW algorithm (Thompson et al., 1994) with a visual check of alignment quality. For Cambaridae and Cyprinidae, we restricted our alignments to one representative sequence per taxon due to the large number of sequences available (5925 for Cambaridae and 98,120 for Cyprinidae). We then analysed these alignments with eDNAssay, using the original TaqMan probe-based model version and methods described in Kronenberger et al. (2022). To be conservative, if multiple haplotypes were found for a single taxon (Nemouridae and Didemnidae only), we used the sequence with the highest eDNAssay assignment probability (i.e. the haplotype with greatest probability of cross-amplification). The final result for each assay was a list of eDNAssay assignment probabilities for all extant confamilials with publicly available sequences on GenBank.

2.3 | Data analysis

The objective of our data analysis was to predict how likely the eDNAssay assignment probability of an unsampled taxon is to exceed given thresholds based on the distribution of eDNAssay assignment probabilities of sampled taxa. Further, we explored the robustness of this approach to data missingness.

To estimate a beta distribution for non-target assignment probabilities, we bootstrap re-sampled the eDNAssay assignment probabilities of individual taxa ($n = 500$). We were initially concerned that variation in the number of sequenced species per genus could cause family-wide predictions of assay specificity to be biased, but sampling genera weighted by described species number (vs. sequencing effort) yielded results similar to an unweighted approach (Appendix S2). We used bootstrap re-sampled (unweighted) sets of non-target taxa to estimate parameters for a beta probability density function using the MASS package (Venables & Ripley, 2002) and to estimate the probabilities of exceeding eDNAssay assignment probability thresholds of 0.5 and 0.3 (see Section 2.4: Model interpretation for justification of these thresholds). We tested for goodness of fit of the beta model to data for each assay using an Anderson–Darling test with the package DescTools (Signorell, 2017) in R.

We then explored the robustness of this approach to data missingness for each assay. We randomly sampled a subset of sequenced non-target taxa to simulate the effect of incomplete data. For each assay, we simulated all possible subsamples of 20—all non-target taxa. We replicated each subsample level 100 times and each time extracted the probability of exceeding the 0.3 and 0.5 eDNAssay assignment probability thresholds. This allowed us to plot the average and bootstrapped 95% confidence intervals of these parameter estimates across possible sample sizes for each assay.

2.4 | Model interpretation

There are multiple ways in which we might use eDNAssay assignment probabilities in assay development and validation. The most meaningful thresholds, in part, depend on project-specific

tolerance of false positive detection errors (see discussion in Kronenberger et al., 2022). For these case studies, we use 0.5 and 0.3 assignment probability thresholds that have been used by the National Genomics Center for Wildlife and Fish Conservation in the process of assay validation to identify when in vitro testing of assay specificity is necessary. Templates with assignment probabilities <0.5 are predicted to not cross-amplify, with an accuracy of 98% across 587 tests of unique assay/template combinations that we have conducted since publication of the eDNAssay model (nationalgenomicscenter.shinyapps.io/eDNAssay/). The lower threshold of 0.3 is meant to reduce the likelihood of false-positive detection errors when analysing environmental samples (i.e. a template is predicted to not cross-amplify but does; a model false-positive error). Below this threshold, cross-amplification probability is or approaches zero and in vitro validation of specificity is usually unnecessary.

As noted above, eDNAssay assignment probabilities should not be interpreted as equivalent to probabilities of cross-amplification. To help readers interpret our modelling results, we approximated probability of cross-amplification for each case study assay by leveraging in vitro tests of eDNAssay model predictions conducted previously (out-of-sample test dataset in Kronenberger et al., 2022). There were 61 tests of assay/template combinations with eDNAssay assignment probabilities ≥0.3 but <0.5, of which two were cross-amplified (cross-amplification rate = 3.3%). Of 59 tests of assay–template combinations with eDNAssay assignment probabilities <0.3, none cross-amplified. Based on this experience, a conservative estimate of cross-amplification risk would be the probability of exceeding the 0.3 threshold, multiplied by a cross-amplification probability of 0.033, plus the probability of exceeding the 0.5 threshold. This approach is conservative (i.e. likely over-estimates cross-amplification risk) because cross-amplification rates for assignment probabilities ≥0.5 are still <1.0 (22/24 tests cross-amplified), at least in part because the eDNAssay model was optimized for recall of cross-amplification (i.e. minimization of false-negative model errors; see Kronenberger et al., 2022 for details). We calculated this metric of risk for each assay because the units (i.e. probability of cross-amplification) are more intuitive than eDNAssay assignment probabilities.

3 | RESULTS

3.1 | Molecular assays

In this manuscript, we focus on the use of the case study assays to test our proposed framework for assessing specificity in silico, but details of assay sensitivity and specificity testing performed in vitro (known-composition tissue DNA extractions and synthetic gene fragments) and in situ (known presence/absence environmental samples) are summarized in the original publications for the arthropod and ascidian assays (Hutchins et al., 2023; Simpson et al., 2017) and in Appendix S1 (including Tables S1 and S2) for the crustacean and teleost assays. The virile crayfish and common carp assays in particular meet requirements for the current version of the eDNAssay model (qPCR assays with a MGB hydrolysis probe and 45 thermal cycles with 60°C annealing/extension, tested using the same PCR mix as in the original model training). See Appendix S1 for details on assay validation.

3.2 | Case study model performance

As of September 2023, the Catalogue of Life database identified 760 Nemouridae, 624 Didemnidae, 427 Cambaridae and 3114 Cyprinidae taxa (not including target and excluded taxa for each assay as listed above). Of these, we were able to identify sequences in GenBank covering the assay region for 78 Nemouridae (10.3%), 57 Didemnidae (9.3%), 285 Cambaridae (66.7%) and 1471 Cyprinidae taxa (47.2%). These resulted in eDNAssay assignment probabilities of 0.056–0.405 (mean = 0.123) for the Nemouridae assay, 0.119–0.839 (mean = 0.253) for the Didemnidae assay, 0.049–0.506 (mean = 0.165) for the Cambaridae assay and 0.077–0.788 (mean = 0.181) for the Cyprinidae assay (Figure 2).

For *Didemnum*, there was a single reference for the target species (*D. etiolum*) in GenBank (Accession KY741541; direct submission, unpublished) with a high eDNAssay assignment probability (0.839), which is much higher than the next highest observed non-target assignment probability (0.386). In a BLAST search of GenBank, the highest hit (95% identity) was for *D. perlucidum*. This GenBank entry is also the only entry in the Barcode of Life Database (BOLD) and Global Biodiversity Information Facility (GBIF). The collection information indicates an inland collection site from India (BOLD), which is surprising given the species description by Kott (2004) being for the Atlantic coast of North America. Because of this, we discarded this sequence from consideration.

Beta distributions for all available sequences appeared to have a reasonable fit for three of the four case study assays. The Anderson–Darling goodness-of-fit tests the correspondence between the observed data and the proposed beta distribution, the null hypothesis being that the data fit the distribution. A significant *p*-value, thus, indicates that the data do not follow the tested distribution. For three of the four case study assays, p-value > 0.05. There was evidence that the beta distribution may not be a good fit for all available Cyprinidae sequence data ($p < 0.01$), primarily because this dataset is heavily right skewed (Figure 2). However, predictive ability of the beta distribution for exceedance of 0.3 and 0.5 thresholds was good: 9.75% predicted versus 8.57% observed for the 0.3 threshold and 0.11% predicted versus 2.79% observed for the 0.5 threshold. Across assays, the model-predicted probability of threshold exceedance was similar to the observed rate (Table 1).

When we tested varying sample sizes for estimating probability of exceeding eDNAssay thresholds, we found that the reproducibility increased with sample size. For *Lednia tumana*, 95% of bootstraps with 20 taxa were within 2.12% of the 0.3 threshold exceedance probability predicted from all data ($n = 78$) and 95% of bootstraps

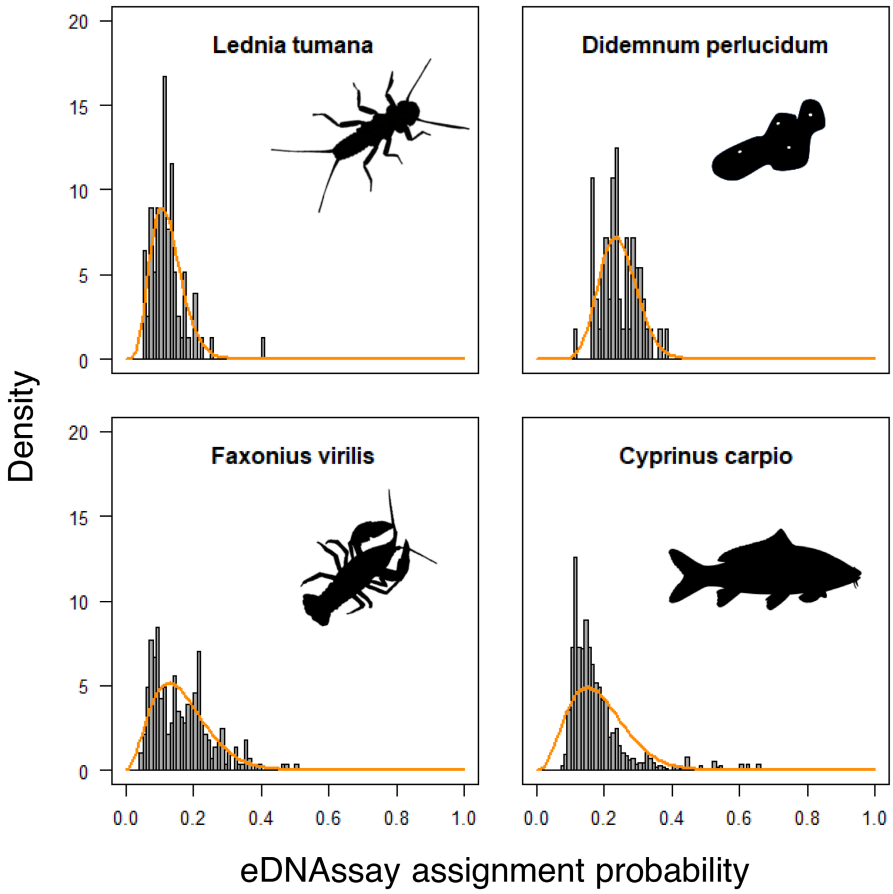

FIGURE 2 Histograms showing the distribution of eDNAssay assignment probabilities across all non-target taxa available for each assay (grey) and modelled beta probability density function (orange).

	0.3 threshold		0.5 threshold	
	Predicted	Observed	Predicted	Observed
Lednia tumana	0.12%	0%	<0.01%	0%
Didemnum perlucidum	15.07%	16.07%	<0.01%	0%
Faxonius virilis	6.69%	7.37%	0.06%	0.03%
Cyprinus carpio	9.75%	8.57%	0.11%	2.79%

TABLE 1 Modelled probability of exceeding the 0.3 and 0.5 eDNAssay thresholds versus observed in available sequence data for the four case study assays using 500 bootstrap re-samples per assay.

with 50 taxa were within 0.53%. For *Didemnum perlucidum*, 95% of bootstraps with 20 taxa were within 11.77% of the 0.3 threshold exceedance probability predicted from all data ($n = 56$) and 95% of bootstraps with 50 taxa were within 3.10%. For *Faxonius virilis*, 95% of bootstraps with 20 taxa were within 11.70% of the 0.3 threshold exceedance probability predicted from all data ($n = 285$) and 95% of bootstraps with 50 taxa were within 5.54%. For *Cyprinus carpio*, 95% of bootstraps with 20 taxa were within 13.31% of the 0.3 threshold exceedance probability predicted from all data ($n = 1471$) and 95% of bootstraps with 50 taxa were within 11.66% (Figure 3).

3.3 | Model interpretation

Using the approximation approach described in Section 2.3, we estimated the probability that an unsampled non-target taxon would cross-amplify to be <0.1% for *Lednia tumana*, 0.5% for *Didemnum perlucidum* and 0.3% for *Faxonius virilis* and *Cyprinus carpio*. Note that these represent estimated cross-amplification risk per taxon, not the estimated cross-amplification risk posed by any one taxon among the full set of unsampled taxa, which would be substantially higher.

4 | DISCUSSION

4.1 | Case study model performance

We found that our approach for assessing cross-amplification risk of non-target sequences for an eDNA qPCR assay generally had robust performance in simulations, even with high levels of taxonomic missingness, suggesting that this approach is useful for predicting specificity for unsampled taxa when comprehensive in silico or in vitro assessments of specificity are infeasible. In all four cases, estimated

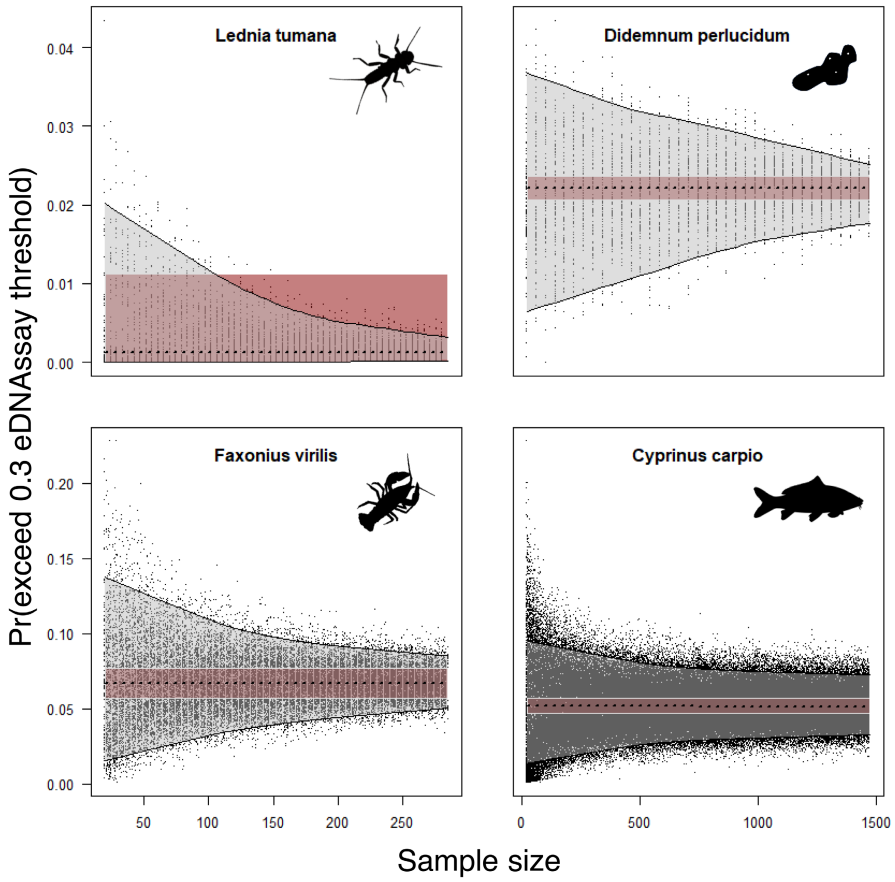

FIGURE 3 Modelled probability of an unsampled taxon exceeding the eDNAssay 0.3 assignment probability threshold based on a sample of 20 to all available non-target confamilial taxa sequences for each assay. Each level of subsampling is bootstrapped 100 times (black points). Grey region represents 95% of bootstrap replicates (quantiles with local regression smoother). The dotted black line indicates the point estimate using all available data. The red rectangle region represents ±1% of the point estimate using all available data for comparison across sub-plots.

cross-amplification risk was similar to observed rates of exceeding the eDNAssay assignment probability thresholds with all available data (Table 1).

Our analyses show that modelled, unsampled taxa cross-amplification risk is sensitive to reference database sample size, but it remains unclear if there is a generalizable 'minimum' sample size required for robust assessment using this method. A reasonable minimum sample size likely depends on both the level of genetic variability among non-target taxa and the level of accuracy and precision that is required. For *L. tumana*, a sample size of less than 40 confamilials consistently resulted in estimated probability of eDNAssay threshold exceedance within 1% of the rate determined using all available data. This may be because, unlike the other three case studies, *Lednia* is a monotypic genus (i.e. there are no congeneric non-target taxa). Conversely, even with >1000 confamilials, probability of eDNAssay threshold exceedance varied between bootstrap replicates by up to 10% for *C. carpio*. Future studies may benefit from conducting a sample size sensitivity analysis via bootstrapping as done here to understand potentially taxon-specific characteristics.

4.2 | Model interpretation

Having estimated the probabilities of exceeding eDNAssay assignment probability thresholds, what should one do with this information? As described by Kronenberger et al. (2022), different assignment probability thresholds might be used depending on project context. In some cases, unexpected cross-amplification may be tolerable (e.g. the potential non-target is highly endemic and inhabits a basin which is not critical to data interpretation). In other cases, an unexpected cross-amplification could incur substantial social, economic and ecological costs (e.g. detection triggers management actions, including cost-intensive follow-up sampling with other methodologies). Here, we assessed thresholds of 0.3 and 0.5 because they may be broadly relevant. Current test data suggest that cross-amplification rates at eDNAssay assignment probabilities <0.3 are close to zero. The 0.5 assignment probability threshold is relevant because above this level the eDNAssay model is predicting cross-amplification (although the observed rate is <1.0). Our proposed framework is valuable because it enables practitioners to use these thresholds to estimate probability of cross-amplifying an unsampled non-target taxon ('model interpretation' in Methods), which for our case study assays ranged from <0.01 to 0.3%. Note that this probability is per taxon, so the cumulative risk across many non-target taxa could be substantially higher (i.e. approximately $1 - (1–0.003)^n$ for virile crayfish, where n is the number of unsampled non-target taxa). When taxonomies are incomplete, n may be unknown or estimated with uncertainty. Thus, this approximation is simply another metric which requires application-specific context.

One important piece of application-specific context is the intended geographic scope of application. For example, here we

assessed the *Cyprinus carpio* assay at a global scale and found thousands of potential non-target Cyprinidae, including some with high eDNAssay assignment probabilities. However, the original intended scope of this assay was to screen for invasive populations in North America. No taxa in Cyprinidae (Schönhuth et al., 2018) are native to North America and <10% of described Cyprinidae are thought to have been introduced. Thus, we might apply this assay with substantially reduced risk in North America as compared to regions with high Cyprinidae diversity (e.g. Asia).

4.3 | Taxonomic–phylogenetic discordance and uncertainty

It is important to note that the effectiveness of the proposed framework is contingent on the assumption that taxonomy is a good proxy for phylogenetic relatedness. This is not always the case. In some taxonomic groups, there are instances of discord between taxonomic designations and phylogenetic relationships and in others, substantial uncertainty due to data incompleteness. Some of these issues arise from incompatibilities between taxonomies relying largely or solely on morphological differences and those informed by molecular data (e.g. phylogenetic species concept; Mayden, 1997). Only species consistent with the latter can be specifically and accurately detected with eDNA sampling. Thus, although the ideal described taxon would be supported by concordant molecular and morphological data (concordant taxa; Figure 4), some data may be missing or discordant, divorcing established taxonomy from the molecular features that eDNA assays can distinguish. Phylogenetically supported taxa with morphological data that are missing (dark taxa; Paige, 2016) or ambiguous (cryptic taxa) often lack taxonomic designation (Figure 4), even though they may have molecular features recognizable with an eDNA assay. Conversely, taxonomic designations based on morphology may lack molecular data (morphotaxa) or even be knowingly discordant with molecular features (pseudotaxa; Figure 4), making taxon-specific eDNA assay design uncertain or impossible. Simultaneously, particularly for diverse systems within poorly sampled regions or ecosystems, substantial numbers of undiscovered taxa lack any molecular or morphological data, representing an 'unknown unknown' risk that is difficult to address in assay design (Figure 4).

The potential for taxonomic–phylogenetic complexity is illustrated by our case studies. For example, within one family we assessed (Cambaridae), the single, currently recognized taxon *Faxonius neglectus* probably encompasses at least three lineages with levels of genetic differentiation similar to the differences among other recognized crayfish species (Dillman et al., 2007). Similar incongruities between current taxonomy and phylogenetic relatedness are relatively common in fishes, amphibians, molluscs and other aquatic taxa that are the frequent focus of eDNA sampling efforts. Because our proposed framework focuses on

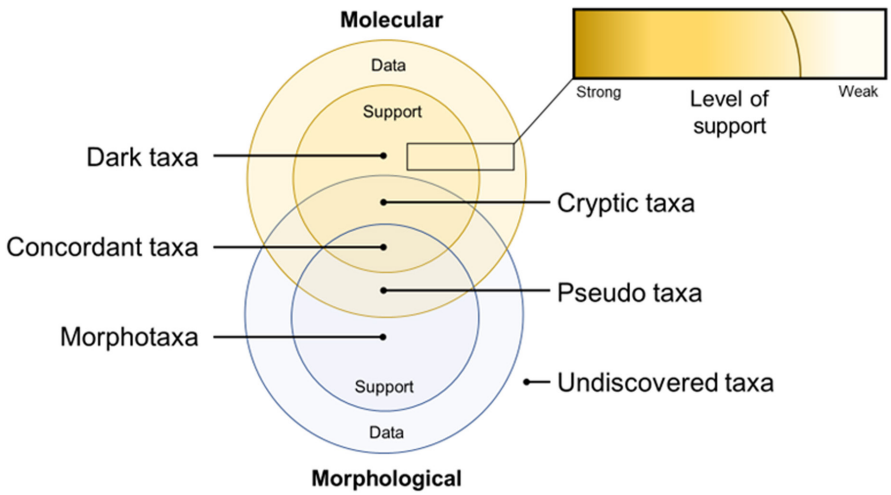

FIGURE 4 Venn diagram indicating the varying levels of support available for different types of taxa. Concordant taxa are supported by both molecular (top) and morphological (bottom) data. Cryptic taxa have both molecular and morphological data but are only supported molecularly. These cryptic, phylogenetic taxa often lack taxonomic recognition because they are morphologically indistinguishable. Dark taxa are supported by molecular data, but no morphological data are available and typically lack taxonomic recognition. Future morphological assessments may either be concordant with molecular data (concordant taxa) or discordant (cryptic taxa). Pseudotaxa have both molecular and morphological data, but their taxonomic designation is only supported morphologically. It is impossible to design eDNA assays which distinguish pseudotaxa. Morphotaxa are supported by morphological data, but no molecular data are available. Future molecular assessments may either be concordant with morphological data (concordant taxa) or discordant (pseudotaxa). Undiscovered taxa are biologically real entities that lack current molecular or morphological data. Note that although this figure implies that support is discrete (a taxonomic designation is either supported or not supported by data), levels of both molecular and morphological support lie on a continuum from weak to strong (top-right box), which may change over the course of subsequent research and taxa may move across types with new information.

family-scale inferences, relatively large phylogenetic–taxonomic discordance would be necessary to cause issues (i.e. a taxon is misplaced at the family level). However, this degree of discordance is possible, particularly in groups where morphological identification is very difficult (i.e. dark and cryptic taxa are common; Figure 4) or where phylogenetic assessments are rare (i.e. morpho- and pseudotaxa are common; Figure 4). Thus, researchers might benefit from testing the assumption of taxonomic/phylogenetic concordance at the family level before applying this framework. Conversely, our proposed framework can be extended to incorporate more fine-scale taxonomic information and make genus-level inferences about specificity. However, this approach would be even more sensitive to incomplete and discordant taxonomies (see Appendix S3; including Table S3 and Figures S1 and S2).

The primary takeaway that we hope to communicate is simply to remind readers that although we attempt to address model uncertainty in our analytical approach (Section 1.4) and model interpretation (Section 2.4), the greatest sources of uncertainty may occur outside the model. Designing assays for taxonomic groups with high levels of diversity, incomplete taxonomies and limited molecular exploration will always come with greater levels of uncertainty and the proposed framework should be applied thoughtfully. Further, our ability to predict amplification risk accurately is contingent on the accuracy of our in silico amplification model (eDNAssay), which may require reparameterization and testing in order to expand application to different assay chemistries.

4.4 | Potential framework extensions

Future studies may expand this framework to new chemistries and applications. Here, we infer taxon-specific assay specificity for unsampled taxa from a given sub-sample. Conversely, an analogous approach might be used to infer assay generality for unsampled haplotypes or taxa. The eDNAssay model can be used to assess qPCR assay generality (see discussion in Kronenberger et al., 2022), so it may be possible to model probability of amplification for unsampled haplotypes. Beyond qPCR, a similar concept might be applied to models for amplicon sequencing (e.g. Wright et al., 2014) to assess metabarcoding assay generality for unsampled taxa.

4.5 | Conclusion

Single-taxon qPCR assays are well known for being highly specific, but they are not intrinsically so. Rather, taxon-specific assay design is only feasible when (1) there are sufficiently distinct molecular features that align with the taxonomic designation and (2) assays are carefully designed and validated to achieve specificity with a good understanding of potential cross-amplification risks across the geographic scope of application. Although many papers have discussed qPCR assay design considerations (e.g. Goldberg et al., 2016; Klymus et al., 2020; Langlois et al., 2021; Thalinger et al., 2021; Wilcox et al., 2013, 2015), there remains little consensus on the level of non-target specificity testing required prior to assay application and, even for assays that do report specificity testing, the extent and intensity varies wildly across peer-reviewed studies (e.g. Thalinger et al., 2021). Assay specificity can only be guaranteed if there is comprehensive specificity testing. However, truly comprehensive specificity testing of qPCR assays which are planned for application over large areas, in biodiverse regions, and among confamilials with incomplete taxonomies is not always possible. Thus, some method for assessing cross-amplification risk across potentially unsampled non-targets is desirable. Here, we demonstrate a framework for using available non-target sequence data to make inferences about these unsampled non-target taxa. This framework was robust to data missingness in our four case study examples. To help other researchers conduct this analysis on their own assays, we have developed a publicly available R Markdown (rmarkdown.rstudio.com) template that automatically generates a fully reproducible report from a single flat file input. The template and instructions can be found at https://github.com/NationalGenomicsCenter/AssayStatisticalFramework.

We hope that this framework enables a more formal consideration of assay specificity against unsampled non-target taxa in future eDNA-based studies.

AUTHOR CONTRIBUTIONS

TMW conceived the study, which was further refined by JAK, MKY, DHM, TWF and MKS. JAK, DHM and TWF contributed molecular protocols and generated data. Analyses were conducted by TMW and JAK. All co-authors contributed to writing the manuscript.

ACKNOWLEDGEMENTS

This work was funded by the Department of Defence Environmental Security Technology Certification Program (ESTCP RC21-5121). Tissue collections for testing assay specificity were provided by Eric Larson, Yvette Paroz, Mike Ruggles, North Carolina Museum of Natural Sciences (Hendrickson et al., 2022; catalogue number 36331), Academy of Natural Sciences (tissue numbers 13181–13183, 17946–50), Florida Museum of Natural History (catalogue numbers 239421, 179954; FLMNH tissue #2010-0256 190216, 166419) and Texas Natural History Collections (catalogue numbers 43513, 55148, 63205, 65050, 69530, 69734). The framework approach was shaped, in part, by conversations with Amanda Mast. The manuscript was improved by helpful comments from Laura Hauck as well as three anonymous reviewers.

DATA AVAILABILITY STATEMENT

Data and scripts underlying this manuscript are available at: https://github.com/NationalGenomicsCenter/AssayStatisticalFramework.

ORCID

Taylor M. Wilcox https://orcid.org/0000-0003-3341-7374
John A. Kronenberger https://orcid.org/0000-0003-3588-7572
Michael K. Young https://orcid.org/0000-0002-0191-6112

REFERENCES

Banki, O. (2022). Catalog of life: From a list to a service. *Biodiversity Information Science and Standards*, *6*(s6), e94040. https://doi.org/10.3897/biss.6.94040

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015). GenBank. *Nucleic Acids Research*, *43*(Database issue), D30–D35.

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.

Dillman, C. B., Wagner, B. K., & Wood, R. M. (2007). Systematics of Species of Special Concern in the Genus *Orconectes* from the Ozark Highlands. Case Studies from Species Pairs: *Orconectes neglectus neglectus*, *Orconectes neglectus chaenodactylus* and *Orconectes macrus*, *Orconectes nana*. State Wildlife Grant Project Final Report (Arkansas Game and Fish Commission). Project T20-2. P 68.

Elbrecht, V., & Leese, F. (2017). PrimerMiner: An R package for development and *in silico* validation of DNA metabarcoding primers. *Methods in Ecology and Evolution*, *8*(5), 622–626.

Filipová, L., Holdich, D. M., Lesobre, J., Grandjean, F., & Petrusek, A. (2010). Cryptic diversity within the invasive virile crayfish *Orconectes virilis* (Hagen, 1870) species complex: New lineages recorded in both native and introduced ranges. *Biological Invasions*, *12*, 983–989.

Fricke, R., Eschmeyer, W. N., & Van der Laan, R. (2019). Eschmeyer's catalog of fishes: Genera, species, references. http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp. Accessed on 06/26/2019.

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

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.

Hendrickson, D. A., Cohen, A. E., & Casarez, M. J. (2022). University of Texas, Biodiversity Center, Ichthyology Collection (TNHCi). Version 5.210. University of Texas at Austin, Biodiversity Collections. Occurrence dataset https://doi.org/10.15468/h8gxdr accessed via GBIF.org on 2022-08-31. https://www.gbif.org/occurrence/2248102392

Hutchins, P. R., Giersch, J. J., Sepulveda, A. J., & Muhlfeld, C. C. (2023). Validation of a species-specific probe-based qPCR assay for the threatened meltwater stonefly, *Lednia tumana*, in environmental samples. *Conservation Genetics Resources*, *16*, 105–107.

Klymus, K. E., Merkes, C. M., Allison, M. J., Goldberg, C. S., Helbing, C. C., Hunter, M. E., Jackson, C. A., Lance, R. F., Mangan, A. M., Monroe, E. M., Piaggio, A. J., Stokdyk, J. P., Wilson, C. C., & Richter, C. A. (2020). Reporting the limits of detection and quantification for environmental DNA assays. *Environmental DNA*, *2*(3), 271–282.

Kott, P. (2004). A new species of *Didemnum* (Ascidiacea, Tunicata) from the Atlantic coast of North America. *Zootaxa*, *732*(1), 1–18.

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*, *22*, 2994–3005.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874.

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*, 519–527.

Mathews, L. M., Adams, L., Anderson, E., Basile, M., Gottardi, E., & Buckholt, M. A. (2008). Genetic and morphological evidence for substantial hidden biodiversity in a freshwater crayfish species complex. *Molecular Phylogenetics and Evolution*, *48*, 126–135.

Mayden, R. L. (1997). A hierarchy of species concepts: The denouement in the saga of the species problem. In M. F. Claridge, H. A. Dawah, & M. R. Wilson (Eds.), *Species: The units of biodiversity* (pp. 381–424). Springer.

Paige, R. D. (2016). DNA barcoding and taxonomy: Dark taxa and dark texts. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *371*(1702), 20150334.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rozansky, Z., Larson, E. R., & Taylor, C. A. (2021). Invasive virile crayfish (*Faxonius virilis* Hagen, 1870) hybridizes with native spothanded crayfish (*Faxonius punctimanus* creaser, 1933) in the Current River watershed of Missouri, U.S. *Aquatic Invasions*, *16*, 690–709.

Schönhuth, S., Vukić, J., Šanda, R., Yang, L., & Mayden, R. L. (2018). Phylogenetic relationships and classification of the Holarctic family Leuciscidae (Cypriniformes: Cyprinoidei). *Molecular Phylogenetics and Evolution*, *127*, 781–799.

Signorell, A. (2017). DescTools: Tools for descriptive statistics. R Package Version 0.99.23. https://cran.r-project.org/package=DescTools

Sievers, F., Wilm, A., Dineen, D., Givson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. *Molecular Systems Biology*, *7*, 539.

Simpson, T. J. S., Dias, P. J., Snow, M., Muñoz, J., & Berry, T. (2017). Real-time PCR detection of *Didemnum perlucidum* (Monniot, 1983) and *Didemnum vexillum* (Kott, 2002) in an applied routine marine biosecurity context. *Molecular Ecology Resources*, *17*, 443–453.

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.

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.

Tsuji, S., Takahara, T., Doi, H., Shibata, N., & Yamanaka, H. (2019). The detection of aquatic macroorganisms using environmental DNA analysis—A review of methods for collection, extraction, and detection. *Environmental DNA*, *1*(2), 99–108.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

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.

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.

Winter, D. J. (2017). Rentrez: An R package for the NCBI eUtils API. *The R Journal*, *9*, 520–526.

Wright, E. S., Yilmaz, L. S., Ram, S., Gasser, J. M., Harrington, G. W., & Noguera, D. R. (2014). Exploiting extension bias in PCR to improve primer specificity in ensembles of nearly identical DNA templates.

Environmental Microbiology, *16*, 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.

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: Wilcox, T. M., Kronenberger, J. A., Young, M. K., Mason, D. H., Franklin, T. W., & Schwartz, M. K. (2024). The unknown unknown: A framework for assessing environmental DNA assay specificity against unsampled taxa. *Molecular Ecology Resources*, *00*, e13932. https://doi.org/10.1111/1755-0998.13932