DETECTING NATIVE FRESHWATER MUSSELS IN PENNSYLVANIA WATERWAYS: COMPARISON & VALIDATION OF ENVIRONMENTAL DNA METHODS

Meredith Bennett

Duquesne University

Follow this and additional works at: https://dsc.duq.edu/etd

Part of the Biodiversity Commons, Other Animal Sciences Commons, and the Terrestrial and Aquatic Ecology Commons

Recommended Citation

This Immediate Access is brought to you for free and open access by Duquesne Scholarship Collection. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Duquesne Scholarship Collection. For more information, please contact beharyr@duq.edu.
DETECTING NATIVE FRESHWATER MUSSELS IN PENNSYLVANIA WATERWAYS: COMPARISON & VALIDATION OF ENVIRONMENTAL DNA METHODS

A Thesis
Submitted to the Bayer School of Natural and Environmental Sciences

Duquesne University

In partial fulfillment of the requirements for the degree of Master of Science

By
Meredith Bennett

May 2023
Copyright by

Meredith Bennett

2023
DETECTING NATIVE FRESHWATER MUSSELS IN PENNSYLVANIA
WATERWAYS: COMPARISON & VALIDATION OF ENVIRONMENTAL DNA

METHODS

By
Meredith Bennett
Approved November 29, 2022

Brady Porter
Associate Professor of Biology
(Committee Chair)

Jan Janecka
Associate Professor of Biology
(Committee Member)

John F. Stolz
Professor of Biology
(Committee Member)

Ellen Gawalt
Dean, Bayer School of Natural and
Environmental Sciences

John F. Stolz
Director, Center for Environmental
Research and Education
ABSTRACT

DETECTING NATIVE FRESHWATER MUSSELS IN PENNSYLVANIA WATERWAYS: COMPARISON & VALIDATION OF ENVIRONMENTAL DNA METHODS

By

Meredith Bennett

May 2023

Thesis supervised by Dr. Brady Porter

North America is home to approximately one third of the world's freshwater mussel species. They are highly imperiled organisms due to habitat destruction and invasive species. Traditional surveys rely on visual identification of mussels, but individuals tend to be rare and difficult to identify. An alternative method is to extract environmental DNA (eDNA) from water samples, which has advantages over traditional sampling, including less sampling effort and fewer hazards to researchers and organisms. We conducted a review of the two main eDNA approaches: single-species detection and metabarcoding. We also developed and validated metabarcoding primers for the detection of native mussels. Four primer pairs in the mitochondrial gene cytochrome oxidase 1 were validated using an equimolar mock mixture, revealing their amplification bias. The eDNA methods described in this project could make surveys faster, more affordable, and more accurate, leading to more effective conservation of mussels and the environment.
DEDICATION

I would like to dedicate this thesis to my family and to the friends who have offered so much support during my time at Duquesne.
ACKNOWLEDGEMENT

This project was made possible by funding from several sources, including a grant from 3 Rivers Quest (3RQ). In addition, the EQT Corporation provided funding in 2016 that enabled the development of our metabarcoding primers. Finally, Duquesne University provided financial support through the Bayer Fellowship.

I would like to thank Doug Locy from Aquatic Systems Inc. and Art Bogon for providing the original mussel specimens used in the mock mixture. Thank you also to Dr. Beth Dakin, who provided monthly eDNA samples through the 3RQ program. In addition, thank you to Zachary Lane, who conducted experiments with Unionids that informed this project.

Special thanks to Dr. Porter for his support throughout this project. His mentorship was invaluable, and I greatly appreciate the opportunities I have received in his lab. I would also like to acknowledge the contributions of Emily Bierer, who guided sequencing preparations and analysis. Thank you also to the entire Porter lab (past and present) for their support, including Kathleen Wilson, Riley Williams, Cassandra Ziegler, Brandon Hoenig, and Felicia Bedford.

Thank you to Dr. Brady Porter, Dr. Jan Janecka, and Dr. John Stolz for acting as my committee for this thesis. Finally, I would like to thank Duquesne’s Center for Environmental Research and Education (CERE) and the Department of Biological Sciences for support throughout this project.
ABSTRACT ............................................................................................................................. iv
DEDICATION .......................................................................................................................... v
ACKNOWLEDGEMENT .......................................................................................................... vi
LIST OF TABLES ................................................................................................................... x
LIST OF FIGURES ................................................................................................................. xii
CHAPTER 1: TWO APPROACHES TO AQUATIC ENVIRONMENTAL DNA SURVEYS: ADVANTAGES AND LIMITATIONS .......................................................... 1
  1.1 INTRODUCTION .............................................................................................................. 1
  1.2 SINGLE-SPECIES DETECTION VS. METABARCODING .............................................. 7
  1.3 ADVANTAGES AND DISADVANTAGES OF SINGLE-SPECIES DETECTION VS.
METABARCODING .............................................................................................................. 8
    1.3.1 General ..................................................................................................................... 8
    1.3.2 PCR Challenges ...................................................................................................... 12
    1.3.3 Replication Challenges and Information per Sample ............................................. 14
  1.4 LIMITATIONS OF eDNA STUDIES ............................................................................. 16
    1.4.1 Quantification ........................................................................................................ 16
    1.4.2 eDNA Persistence and Transport ...................................................................... 18
  1.5 TIME AND COST ANALYSIS FOR eDNA SURVEYS VS. TRADITIONAL SURVEYS ...... 21
  1.6 CONCLUSIONS ................................................................................................ .......... 26
Chapter 2: Designing and Validating Metabarcoding Primers for Detection of Freshwater Mussel (Unionid) eDNA ......................................................................................... 28
  2.1 Introduction .................................................................................................................. 28
2.1.1 Freshwater Mussels (Unionidae) ................................................................. 28
2.1.2 Sampling Methods ....................................................................................... 31
2.2 Methods ........................................................................................................... 34
2.2.1 Tissue Extraction from Specimens .............................................................. 34
2.2.2 Primer Design and Initial Amplification ...................................................... 35
2.2.3 Mock Mixture Creation and Internal Primer Validation ............................... 39
2.2.4 Next-Generation Sequencing ....................................................................... 40
2.2.5 Data Analysis ............................................................................................... 42
2.3 Results ............................................................................................................. 43
2.3.1 Overview of Results from 1st Sequencing Run (August) .............................. 43
2.3.2 Stochasticity Across PCR Replicates ............................................................ 43
2.3.3 Amplification Bias of Internal Primer Pairs .................................................. 48
2.3.4 Results from 2nd Sequencing Run (October) and Comparison with August ... 52
2.3.5 Comparing Results to a Previous Study of Amplification Bias .................... 55
2.4 Discussion ........................................................................................................ 59
2.4.1 Amplification Bias ....................................................................................... 60
2.4.2 Amplification Bias Across Separate Sequencing Runs ............................... 67
2.4.3 Comparison with Previous Mock Mixture Study ......................................... 67
2.5 Conclusion and Future Directions ................................................................... 69

Chapter 3: Using Metabarcoding Primers to Detect Unionid eDNA ..................... 71
3.1 Methods .......................................................................................................... 71

3.1.1 Site Selection and eDNA Collection ............................................................ 71
3.1.2 Filtration of Water Samples ....................................................................... 73
3.1.3 eDNA Extraction from Filters ................................................................. 73
3.1.4 Amplification of Extracted eDNA ............................................................. 75
3.1.5 Next-generation Sequencing ................................................................. 77
3.1.6 Data Analysis ....................................................................................... 78

3.2 Results ...................................................................................................... 79
3.2.1 Overview of eDNA Results ................................................................. 79
3.2.2 Unique Sequences in eDNA Samples ................................................. 82

3.3 Discussion ................................................................................................. 84
3.3.1 Unique Unionid Species ................................................................. 84
3.3.2 Contamination ..................................................................................... 85

3.4 Conclusion and Future Directions .......................................................... 86

Chapter 4: Conclusions ................................................................................. 88

References ...................................................................................................... 90

Appendix ......................................................................................................... 117

References for Cost Analysis of eDNA and Traditional Methods ............... 117

Backpack Electrofishing ................................................................................ 117

Environmental DNA (eDNA) Methods ......................................................... 117
Table 1. Advantages and disadvantages of the single-species detection and metabarcoding methods………………………………………………………………….11
Table 2. Cost analysis of backpack electrofishing and two eDNA methods…………23
Table 3. Labor costs for backpack electrofishing and two eDNA methods…………24
Table 4. Internal primer combinations tested on the mock mixture………………..37
Table 5. Universal primer combination used for initial amplification of mock mixture species…………………………………………………………………38
Table 6. Number of sequences identified to each mock mixture species for each PCR replicate……………………………………………………………………45
Table 7. Number of sequences identified to each mock mixture species for each primer pair……………………………………………………………………….49
Table 8. Number of sequences identified to each mock mixture species for each individual primer…………………………………………………………51
Table 9. Number of sequences identified to each mock mixture species for each primer pair in October sequencing run……………………………………..53
Table 10. Number of sequences identified to each mock mixture species for each primer pair in the two sequencing runs (August vs. October)……………………….54
Table 11. Number of paired-end sequences identified to each mock mixture species for two primer pairs in this study (Bennett) vs. Zachary Lane’s study in 2017………..57
Table 12. Number of base pair mismatches between the mock mixture species and the internal primers designed for this study……………………………..65
Table 13. Sequence counts for the most common Unionid species identified by primer pair 1F-1R………………………………………………………………………………………….80

Table 14. Sequence counts for the most common Unionid species identified by primer pair 2F-2R………………………………………………………………………………………….81

Table 15. Sequence counts for the 3 unique (non-mock mixture) Unionid species identified in this study………………………………………………………………………………………….83
LIST OF FIGURES

Figure 1. General workflow of eDNA studies .................................................................2
Figure 2. Single-species detection vs. metabarcoding ..................................................3
Figure 3. Common applications of eDNA .................................................................6
Figure 4. Determining a source of eDNA .................................................................20
Figure 5. Recommendations for selecting the optimal eDNA method .......................25
Figure 6. Spatial map of universal and internal primers across COI and COII .............37
Figure 7. Number of sequences identified to each mock mixture species for three PCR replicates with primer pair 1F-1R (A) .................................................................46
Figure 8. Number of sequences identified to each mock mixture species for three PCR replicates with primer pair 2F-2R (B) .................................................................46
Figure 9. Number of sequences identified to each mock mixture species for three PCR replicates with primer pair 3F-3R (C) .................................................................47
Figure 10. Number of sequences identified to each mock mixture species for three PCR replicates with primer pair 3F-4R (D) .................................................................47
Figure 11. Proportion of sequences identified to each mock mixture species for each primer pair and for an even distribution (Expected) .............................................50
Figure 12. Proportion of sequences identified to each mock mixture species for each individual primer .................................................................52
Figure 13. Proportion of sequences identified to each mock mixture species for each primer pair in two different sequencing runs (August vs. October) ..................55
Figure 14. Proportion of sequences identified to each mock mixture species for primer pair 1F-1R in this study (Bennett) vs. Lane’s .......................................................58
Figure 15. Proportion of sequences identified to each mock mixture species for primer pair 3F-4R in this study (Bennett) vs. Lane’s .................................................................58
Figure 16. Phylogeny of Unionidae ...........................................................................62
Figure 17. Alignment of forward primer 1 (1F) with reference sequences of the 6 mock mixture species ........................................................................................................66
Figure 18. Alignment of reverse primer 3 (3R) with reference sequences of the 6 mock mixture species ........................................................................................................66
Figure 19. Locations of monthly 3RQ eDNA collection sites ....................................71
Figure 20. Locations of eDNA collection sites in the Buffalo Creek and Crooked Creek watersheds ..................................................................................................................72
Figure 21. Layout of eDNA sample sequencing for each primer pair ....................77
CHAPTER 1: TWO APPROACHES TO AQUATIC ENVIRONMENTAL DNA SURVEYS: ADVANTAGES AND LIMITATIONS

1.1 INTRODUCTION

Environmental DNA (eDNA) consists of DNA molecules found in the environment. It can occur in many forms, including tissue, feces, gametes, and dead cells.\textsuperscript{1, 2} Environmental DNA can be found in many substrates, including soil, permafrost, and even air.\textsuperscript{1, 3, 4} The vast majority of eDNA studies, however, are conducted in aquatic ecosystems.\textsuperscript{5-8} Unlike studies of microbes where entire genomes are detected by directly sampling living organisms,\textsuperscript{9, 10} eDNA studies target DNA molecules.\textsuperscript{1} It is important to note that eDNA is not a scientific method, in itself.\textsuperscript{1} Instead, environmental DNA is a useful resource that can be extracted from the environment and analyzed with scientific methods, such as polymerase chain reactions (PCRs) and metabarcoding.\textsuperscript{11} These methods are made possible by rapidly developing technology, such as next generation sequencing.\textsuperscript{11} Once it is shed from aquatic organisms, environmental DNA becomes available for scientists to collect. DNA is “captured” from water samples using various methods, including precipitation, centrifugation, and filtration.\textsuperscript{12} Filtration has been found to be the most successful of these capture methods for aquatic samples.\textsuperscript{12} The eDNA is then extracted from the filters for analysis (Figure 1).\textsuperscript{12-14}
There are two main approaches for analyzing eDNA: single-species detection and metabarcoding. The single species approach requires the development of highly specific probes that only amplify a particular species of interest (Figure 2). These probes are used in tandem with primers to perform quantitative PCR (qPCR) on eDNA, in which many copies are made of the target region of DNA. The probes contain reporter molecules that emit fluorescence upon each successful replication. Once the number of copies/fluorescence reaches a certain threshold, detection of the target species is confirmed through successful amplification. In metabarcoding studies, universal primers are designed to amplify a broad taxonomic group of organisms (Figure 2). Next generation sequencing is then performed on the mixture of PCR products, simultaneously detecting numerous species by comparison with a reference sequence database. In most eDNA studies, primers are designed to target mitochondrial DNA. Mitochondrial genes are optimal for eDNA because their high copy number per cell helps ensure survival through environmental degradation. Mitochondrial genes also have an extensive reference database and contain higher species-specific variation.
Environmental DNA studies have been applied to a wide range of topics (Figure 3). Two of the more common applications of eDNA are the study of rare and elusive species \(^3,4,23-25\) and the monitoring of invasive species. \(^8,17,26-29\) The Hula painted frog is a rare species that was recently rediscovered after it was believed to be extinct. \(^4\) Environmental DNA was used to locate additional populations of the frog and determine the features that classify suitable habitat for the organism. \(^4\) A successful species-specific primer has also been designed for the highly endangered Dwarf Wedgemussel, native to North America. \(^30\) Studies like these help researchers target important areas for conservation. In the realm of invasive species, invertebrate stowaways on transatlantic vessels have been identified in the Laurentian Great Lakes using two new eDNA metabarcoding assays. \(^17\) The assays detected invasive bivalve and gastropod species while contributing to the library of assays applicable to studies of mollusks. \(^17\) Similarly, researchers in
New Zealand were able to successfully detect the invasive mosquito fish, *Gambusia affinis* using environmental DNA. Like studies of rare species, eDNA-based monitoring can help inform efforts to manage/eradicate invasive species. Environmental DNA can also be used to measure biodiversity and genetic variation. For example, metabarcoding has been used to measure the diversity of Guianese fish species, providing a detailed inventory of spatial data for 132 total species. Similarly, in a study of Mediterranean fish, eDNA metabarcoding was shown to detect more functional diversity within the communities than visual methods. The genetic variation of perch-like fishes (Perciformes) in Russia has also been analyzed using eDNA metabarcoding. Researchers identified 42 official species along with many distinct cryptic species. These studies illustrate that eDNA can be used to study intraspecific genetic variation in addition to simply detecting species in the environment.

Ecological studies are often concerned with the distribution and behavior of aquatic species, and eDNA has been applied to these questions as well. Movements of the invasive Bigheaded Carp have been described across time and location using eDNA monitoring. A relationship was also found between the presence of carp eDNA and hydrographic conditions. DNA metabarcoding has even been applied to the reproductive behavior of aquatic species. By sampling ichthyoplankton (fish eggs and larvae) in the Amazon River and extracting their DNA, researchers were able to determine the reproductive timing of various fish species. This differs from most eDNA studies, in that complete organisms (ichthyoplankton) were collected, not just water samples. Based on the views of the scientific community, studies like this one in the Amazon qualify as eDNA studies. As researchers have found more applications for eDNA, the definition of eDNA has become quite broad, including microbes, and meiofauna. One such application is the use of diatom eDNA for ecological assessments of rivers and streams. Because
diatoms are highly sensitive to environmental changes, their community structures can indicate the health of aquatic ecosystems. \(^{47}\) Recent studies have used eDNA metabarcoding to describe diatom communities and inform ecological assessments. \(^{48,49}\) These studies are especially useful, because they often do not require taxonomic information for diatoms, which in many cases has not been resolved to the species level. \(^{48,49}\) In addition to diatoms, eDNA studies have used bacteria and phytoplankton to glean information about aquatic ecosystems, especially in response to nutrient pollution. These studies can illustrate how various fertilizers and pesticides affect the biological communities in rivers and streams. \(^{50,51}\)

Recent studies have applied eDNA to substrates other than water. For example, eDNA collected from streams can be used to detect terrestrial mammals in the surrounding land area, often more effectively than traditional camera traps. \(^{52-54}\) In addition, it is possible to collect eDNA from the air to detect species within a confined space. \(^{55}\) Researchers have even shown that spider webs can be reliable sources for eDNA. \(^{56}\) Webs act as filters of the air column, meaning they accumulate DNA from diverse groups of organisms and can be used to detect species in the area surrounding the web. \(^{56}\) Another study found that aquatic biofilms can serve a similar function for fish DNA. \(^{57}\) The biofilms naturally collect eDNA, which can be extracted and used to characterize aquatic communities. \(^{57}\) Additional applications of eDNA will likely appear with the development of more advanced sequencing technology and the standardization of eDNA collection and extraction methods.
Environmental DNA surveys generally have equal or higher detection probabilities than traditional methods. They also provide accurate taxonomic identification and may detect cryptic species/life stages that traditional surveys miss. Surveys using eDNA are less invasive and require reduced sampling effort. They also reduce the risk of translocating invasive species or diseases. The manual identification of species required in traditional surveys can be difficult, especially when individuals are damaged or morphologically cryptic. Surveys using eDNA avoid this challenge, along with some logistical constraints including sampling locations with low accessibility. These characteristics of eDNA have led scientists to supplement traditional methods with eDNA surveys. When combined, the two methods (eDNA and morphological surveys) provide more
Despite the many advantages and opportunities associated with eDNA studies, challenges remain. Natural resource managers will likely have concerns about eDNA persistence in ecosystems and the distance eDNA can travel once it is shed from target organisms. In addition, there is ongoing debate over whether eDNA can be quantified. In this review, we will discuss these challenges and their implications for the future of eDNA studies.

1.2 Single-species Detection vs. Metabarcoding

There are important differences to consider between the two methods of eDNA analysis, single-species detection and metabarcoding. In single-species detection, quantitative PCR (qPCR) is used for its advantages over standard PCR. By using fluorescent labelling probes, qPCR is more sensitive than general PCR and can provide some information about the quantity of the DNA signal in samples. The probes required for qPCR must be highly specific to the target species. Ideally, the species-specific probe and primer combination will amplify the target DNA sequence by making millions of copies during PCR. If PCR works and the DNA sequence is amplified, the target species has been detected. For example, species-specific probes were developed to detect the Macquarie perch (*Macquaria australasica*), a fish endemic to Australia. The primers were designed to detect the 12S region of the mitochondrial DNA along with a specific region of nuclear DNA for the Macquarie perch. Using qPCR, the primers successfully detected Macquarie perch DNA and were used to track perch spawning activity based on abundance data.

Unlike in single-species detection, metabarcoding requires the development of a more general primer used to detect large taxonomic groups. The primer must be general enough
to amplify all desired species, but not too general that it amplifies the DNA from non-target species, such as humans. The general workflow of a metabarcoding study consists of collection, extraction, amplification, and sequencing. Instead of detection through PCR amplification, metabarcoding studies rely on detection of species through data generated by next-generation sequencing. Sequences are clustered into Operational Taxonomic Units (OTUs) and identified to species by matching to a reference database. A metabarcoding study was used to evaluate the diversity of Guianese freshwater fishes. Around 8 million sequencing reads were generated and 132 Guianese fish species were identified using the metabarcoding method. Because of the differences between single-species detection and metabarcoding, the two methods are useful under different circumstances.

1.3 Advantages and Disadvantages of Single-Species Detection Vs. Metabarcoding

1.3.1 General

Single-species detection and metabarcoding have advantages and disadvantages based on the circumstances of a given study (Table 1). Single-species studies are convenient, because PCR amplification and detection occur simultaneously, eliminating the need for additional steps. Single-species studies are also generally less expensive than metabarcoding because they lack the extra expense of next-generation sequencing. In addition, single-species analyses can be performed as samples trickle in without affecting the cost. Metabarcoding becomes cost efficient only after collecting a large batch of samples, delaying analysis until samples are collected. As sequencing costs continue to decrease, however, metabarcoding studies will likely become more cost efficient.

There are many instances when metabarcoding is more suitable than single-species
detection. Conveniently, metabarcoding detection does not require any prior knowledge of the species that might be detected. 58 This allows scientists to detect species they were not necessarily looking for, greatly expanding the implications of surveys. Metabarcoding also requires less development and validation of primers, because a general primer will work for a broad taxonomic group. Single-species detection, on the other hand, requires the development and validation of a new probe for each species. 5, 58, 77 If single-species primers or probes are not designed correctly, they can generate false positives by cross-amplifying non-target species. 15, 58, 77 False negatives can also be produced in single-species detection by failing to amplify members of the target species with geographic or intraspecific variation. 15, 58, 77 False positives and false negatives are also known as Type I and Type II logistical errors, respectively. Metabarcoding also produces more reliable identifications than single-species detection. In metabarcoding studies, the identification of species is based on the analysis of multiple diagnostic sequence sites compared with a reference database, while single-species detection is based only on whether PCR works or not. 5, 78 Furthermore, additional validation of single-species detections might be needed to distinguish between closely related taxa, such as the various Asian Carp species. Finally, metabarcoding is generally more useful than single-species detection, because metabarcoding surveys contribute to future reference sequence databases, 15, 78 increasing knowledge of regional variation and potentially revealing cryptic species. In single-species studies, your results are limited to presence/absence data for the single species you were trying to detect. 70-73 Because metabarcoding is generally more resistant to logistical errors and produces more reliable species detection and identification, we argue that this method is superior to single-species detection in most situations. Metabarcoding also yields considerably more information, because it can be used to characterize an entire community of organisms. There are still
instances, however, when single-species detection might be advantageous over metabarcoding.
<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-species</td>
<td>• Less expensive than metabarcoding</td>
<td>• Requires a validated qPCR probe for each species</td>
</tr>
<tr>
<td>Detection</td>
<td>• Amplification and detection occur simultaneously</td>
<td>• Non-specific probes will cross-amplify with non-target species (false positives)</td>
</tr>
<tr>
<td></td>
<td>• qPCR can provide estimates of relative quantity of target eDNA in sample</td>
<td>• Overly specific probes fail to amplify target species with intraspecific variation (false negatives)</td>
</tr>
<tr>
<td></td>
<td>• Results can be obtained as sampling occurs over time without impacting cost</td>
<td>• Identification is less reliable (detection is based only on PCR amplification which can be impacted by poor primer design, PCR inhibitors or other assay issues)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Results show only presence/absence of one species</td>
</tr>
<tr>
<td>Metabarcoding</td>
<td>• Prior knowledge of the detectable species is not required when using generalized primers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Less primer development and validation; general primers are used</td>
<td>• More expensive than single-species approach</td>
</tr>
<tr>
<td></td>
<td>• Identifications are more reliable (based on multiple diagnostic sequence sites matched to reference databases.</td>
<td>• Analysis delayed until all samples are collected</td>
</tr>
<tr>
<td></td>
<td>• Surveys contribute to reference sequence databases and knowledge of regional variation</td>
<td>• Hindered by inadequate reference databases.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Amplification bias (primers detect certain species more effectively)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Stochastic bias (inconsistency in PCR due to uneven nature of eDNA)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Common taxa may outcompete rare taxa in sequencing data</td>
</tr>
</tbody>
</table>
1.3.2 PCR Challenges

PCR amplification is an important step in both single-species detection and metabarcoding. Therefore, challenges encountered during PCR can have detrimental implications for the detection of species in both methods. For both single-species detection and metabarcoding, successful amplification is limited by DNA concentration. Only a small portion of the DNA in an environmental sample will be from the target organism(s), reducing the likelihood of detecting those taxa. In single-species detection, PCR amplification itself is how species are detected in samples.¹⁸,²⁰,⁵⁸,⁷⁹ If PCR amplification does not occur, there is no detection of the target species.¹⁸,⁴⁰,⁷³ In other words, detection is dependent on whether PCR works or not. Unfortunately, PCR is not perfect, and many factors can cause amplification to fail even if the target DNA is present in the sample. Certain substances co-extracted with eDNA, such as humic and fulvic acids or organic matter are called PCR inhibitors, because they can cause PCR to fail by interfering with the extracted eDNA.¹⁵,¹⁸,²⁰,⁵⁸,⁸⁰ PCR might also fail if the samples have inadequate concentrations of eDNA or if a researcher inadvertently omits a reagent when creating the master mix for PCR. Dilution is often used to decrease the effect of PCR inhibitors.⁴⁰,⁵⁸,⁸⁰ In addition, using qPCR instead of normal PCR can reduce errors, because qPCR utilizes an internal positive control in which known numbers of a certain gene are added to samples.⁴¹,⁸⁰,⁸¹ If the internal positive control does not amplify, the researchers can assume there is a problem with PCR and try again. Even when using qPCR, however, errors can still cause amplification to fail, leading to incorrect conclusions about the detection of the target species.

While single-species detection is limited in reliability because detection is completely
reliant on the success of PCR amplification, metabarcoding encounters unique PCR challenges. In metabarcoding, if PCR amplification fails initially, researchers usually try again until something amplifies, after which they move on to sequencing. One of the major challenges associated with PCR in metabarcoding is amplification bias. Because of imperfect primer design and incomplete genetic databases, primers often match certain DNA sequences better than others, leading to biased amplification of some species over others. A gene often used for the creation of universal primers is the cytochrome oxidase subunit I (COI) mitochondrial gene primarily because it has the most extensive sequence libraries in genetic databases, such as GenBank. The COI gene has been criticized recently, however, because it has considerable species primer bias among other issues. Researchers have argued that COI has insufficient variable regions to discriminate between species and consequently underestimates biodiversity. Based on these revelations about the specificity of the COI gene, universal primer design should clearly not be a one-size-fits-all process. Techniques used to reduce primer bias include finding alternative genes to COI such as 16S rRNA genes or using statistical packages to evaluate potential primers. Primer bias can also be realized by constructing mock mixtures with known copy numbers of DNA from multiple species or using highly degenerate primers (primers with positions that contain more than one possible base). Despite these potential solutions, primer bias is likely to be a persistent problem in metabarcoding studies.

Stochasticity is another challenge to unbiased PCR amplification in metabarcoding studies. PCR amplification is an irregular process in which each molecule in a sample is not replicated during every PCR cycle as we would expect if PCR were perfect. Issues associated with stochasticity are reduced when samples possess an equal concentration of each
target sequence, but stochasticity may severely distort detection results if the samples did not have sufficient concentrations of eDNA. Due to stochasticity, sequences at higher concentrations may be differentially amplified over rarer sequences. This competition can lead to rare species being “masked” or even undetected. Furthermore, stochasticity is enhanced by the inherent amplification bias of the primers used. Although next-generation sequencing is generally a more reliable method of identification than qPCR, metabarcoding studies encounter unique challenges associated with PCR amplification, namely primer bias and stochasticity.

1.3.3 Replication Challenges and Information per Sample

As discussed in the previous section, PCR amplification is susceptible to errors. A particular PCR reaction might not occur for many reasons, including insufficient mixing, omission of a reagent, insufficient template concentration, etc. This can lead one to believe the target species is absent when that might not be true. Replicating the eDNA assay decreases the detrimental effects of these random errors in which PCR amplification fails. In single-species detection, PCR is replicated for each sample to gauge the consistency of amplification for the given sample. The number of replicates done for each sample varies with each study from 2 to 12. The most common number of replicates done for each sample, however, is 3 (triplicate). The requirements for detection also vary widely across single-species studies. Sometimes a species is considered present if at least one replicate amplifies successfully. In other studies, more than one replicate must amplify positively for the target species to be considered present in the sample, such as 2 positive replicates out of 6 total replicates. In qPCR, detection is confirmed by comparing the strength of fluorescence to a standard curve with a predetermined threshold for detection. In this case, special actions
are taken to contend with replicates that do not amplify, including assigning failed replicates with a value that will decrease the overall PCR efficiency measurement in a curve. 18, 40

A significant challenge with single-species detection is the limited amount of eDNA that can be extracted from each environmental sample. When using the single-species detection method to study one individual species, researchers can do as many replicates during PCR as possible with the amount of eDNA extracted from the sample. 4, 23, 73 On the other hand, single-species detection can be used to test for multiple species in samples by designing different species-specific primers for separate PCR reactions. 41, 71 Using the single-species approach to detect more than one species is not only tedious, but limits the number of replicates that can be done. 90 When the limited eDNA in samples must be used to detect many different species, less of that eDNA is available to do replicates for each species-specific DNA assay. 90 Using the single-species method to look for multiple species is generally less efficient than metabarcoding, because it provides much less information per sample due to fewer possible replications for each sample.

Metabarcoding is much more effective than single-species detection in studies with many target species. The general taxonomic primers used in metabarcoding are optimized for multiple species, and therefore do not limit the number of possible replicates for each sample. 26, 32, 82 Because there is only one primer set for all target species, researchers can do as many replicates as possible, limited only by the amount of extracted DNA. As with single-species detection, the number of replicates performed per sample varies across studies, 32, 89 but the most common number of replicates is 3 (triplicate). 59, 82 Most commonly, the PCR products from all replicates are pooled and sequenced together using next-generation sequencing. 26, 32, 59 As discussed earlier, the main challenges associated with PCR in the metabarcoding method are primer bias
and stochastic errors. While sequencing all PCR products together decreases the effects of primer bias and stochasticity, there still may be differences in replicates from the same sample. In many studies, sequences are filtered using software that removes sequences that do not meet certain quality criteria. In other studies, a species may be considered present in a given sample only if the number of sequence reads surpasses a predetermined threshold. Despite the attempt to decrease errors in sequence data, sequences of replicates are likely to remain inconsistent because of the inherent stochasticity in the PCR process and the uneven composition of eDNA samples. Rarer species often fail to appear in sequence data, for example, because their sequences are lost in the abundant sequences of more common species. Despite the complications of primer bias and stochasticity in metabarcoding, it is much more valuable than single-species detection when looking for many species. While the additional steps in metabarcoding can become tedious, this method provides researchers with more information per sample and requires less primer validation than single-species detection.

1.4 LIMITATIONS OF eDNA STUDIES

1.4.1 Quantification

In addition to presence/absence data, biologists are often interested in quantitative information, such as the number of individuals or biomass for various species. Quantifying organisms is crucial for biologists attempting to manage the populations of species in aquatic ecosystems. In addition, quantitative survey data informs conservation efforts more effectively.

However, there is substantial controversy over whether eDNA data can or should be quantified. Although scientific reviews of eDNA methodology are generally positive, very few claim that eDNA can provide reliable quantitative data. True quantification of eDNA is
likely only possible with single-celled prokaryotes, because they possess just one DNA copy per individual. Because single-celled eukaryotes and multicellular organisms possess many copies of organelle DNA, eDNA can, at best, only provide relative abundance or biomass for those species. In addition, there are many factors that complicate eDNA production and capture. For example, the methods used to capture and extract DNA from environmental samples are not equally efficient.\textsuperscript{12, 93, 94} The size and type of filter used to capture eDNA is an important factor affecting quantification, with glass fiber filters generally performing better than polycarbonate filters.\textsuperscript{12} It is also important to note that eDNA studies extract more than just free DNA. In fact, free DNA likely makes up a small portion of extractions compared to intact cells and mitochondria. This raises the issue of different particle sizes. Researchers may capture a greater diversity of genetic material by using filter papers with assorted pore sizes. More research is needed to validate whether filter types truly differ in this way.

The materials used in PCR amplification of samples also affect eDNA results. Commercial polymerases used in PCR are often biased towards particular species,\textsuperscript{89} making it difficult to determine the actual abundance of different species in a sample. The same problem is caused by primer bias.\textsuperscript{95} A considerable effort must be made to standardize the methods used in eDNA studies if they are to provide quantitative information.

Environmental and organismal factors also affect the production and subsequent quantification of eDNA. For example, water temperature seems to affect the relationship between species abundance and the concentration of eDNA, with higher temperatures showing stronger correlations.\textsuperscript{20, 96} Other studies have noted that individuals of the same species shed DNA at different rates depending on their life stage, with individuals generally producing less DNA as adults.\textsuperscript{97-99} Therefore, a comparatively large amount of eDNA for a given species does
not necessarily mean it is more abundant. Scientists using eDNA must account for these variables in their studies if they are trying to estimate species abundance.

Most studies find only a small or no correlation between eDNA concentration and species abundance. An extensive review of 22 metabarcoding studies found that collectively, biomass and number of sequences are only weakly correlated (slope = 0.52). This suggests that scientists should be hesitant to draw quantitative conclusions from eDNA data. In addition, although qPCR can offer some quantification of the DNA in samples, this does not necessarily translate to the environment itself. That being said, there are instances when scientists have found significant relationships between species abundance and sequencing data, especially when environmental factors are accounted for. In addition, new technologies show promise for improving the quantification of species from eDNA. For example, digital droplet PCR (ddPCR) reduces the effects of PCR inhibitors and improves the reliability of PCR amplification by partitioning a given sample into thousands of individual droplets. Using internal standards of known quantity and modeling the rates of eDNA shedding in particular species can also make estimates of abundance more accurate. Despite these innovations, eDNA-based methods have a long way to go before they can provide reliable quantitative information for individual species under a variety of environmental conditions.

1.4.2 eDNA Persistence and Transport

Another source of uncertainty in eDNA studies is the ecology of eDNA itself. Reviews of eDNA-based methods are often concerned with how long DNA persists in the environment and how it travels within stream systems. In traditional surveys, capture of a specimen assures it is currently living at that site. In eDNA studies, questions arise concerning residual eDNA.
Generally, eDNA becomes undetectable at a source site after about two days, but some studies report faster or slower degradation. Environmental factors affect the persistence of eDNA. Some increase degradation rates, including acidity and microbial activity. In addition, eDNA generally degrades faster in marine environments and at higher temperatures. On the other hand, eDNA can persist longer when it forms aggregates with sediment and biofilms. Finally, concentrations of eDNA for a given aquatic species will increase when that species is spawning. These environmental and life history variables illustrate the importance of characterizing potential sites for eDNA studies before collecting samples, noting hydrology, chemistry, and spawning season. Finally, eDNA persistence is dependent on the kind of genetic material being studied. Environmental RNA has been shown to degrade faster than DNA. Similarly, nuclear DNA appears to degrade faster than mitochondrial DNA. It is important for scientists to acknowledge these differences when interpreting the data from eDNA studies.

Another major uncertainty is how far eDNA can travel before being detected. Due to environmental factors, it is difficult to determine exactly where eDNA originates from. Stream discharge determines how eDNA is transported within an aquatic system. In a flowing system eDNA is known to have come from upstream. In one study of a Common Carp farm, eDNA was detected as far as 3 kilometers away from the discharge point. The intensity of an eDNA signal is strongest at the source of DNA and decreases with increasing distance from the source. Therefore, sampling upstream of an initial detection site could allow scientists to determine the source of eDNA. Like the process of detecting chemical pollutants, eDNA can be detected (or not) at upstream sites until a potential source is found (Figure 4). The accuracy of future eDNA studies can be improved by embracing more frequent and widespread sampling.
Scientists should also account for the environmental variables that affect the degradation and transport of eDNA.

**Figure 4:** Determining a source of eDNA. By sampling upstream of an initial detection site, the origin of eDNA may be determined.
Finally, a major concern in eDNA studies is contamination. Water samples may be contaminated by the collector, by residual DNA in lab settings, or through cross contamination with each other. \textsuperscript{81,120} If researchers are unaware of contamination in their samples, they may report false positives. In many eDNA studies, researchers take “blanks” to identify whether contamination occurs during sampling. \textsuperscript{58,121} Blanks often consist of a bottle filled with sterilized water and stored alongside actual samples. \textsuperscript{58,81,121} If the blanks are found to contain target DNA, then one can assume contamination has occurred. Researchers may omit any contaminated samples from their analysis, while others have tried to correct the number of sequence reads in samples by subtracting those found in corresponding blanks. \textsuperscript{121} Most studies implement practices to avoid contamination, \textsuperscript{4,79} but these sterilization techniques are often minimal and vague, such as conducting analyses in a “clean lab”. \textsuperscript{4} Some biologists have discussed the need to address contamination in eDNA studies, \textsuperscript{120,122} but a surprising number of studies make no mention of contamination or hygiene practices. If eDNA is to be a reliable tool in ecological studies, methods must be standardized to address the pervasive problem of contamination.

1.5 Time and Cost Analysis for eDNA Surveys Vs. Traditional Surveys

Cost is an important factor to consider when conducting biological surveys. Many resource managers will be concerned that eDNA methods are still too expensive to be feasible. Traditional morphological methods and eDNA methods differ in the specific expenses making up the total cost. However, as molecular technologies become less expensive, the total costs of eDNA studies are closer to the costs of traditional surveys. \textsuperscript{123,124} In addition, traditional methods tend to have much higher labor costs, because they require more time and person power. \textsuperscript{123,124} Beyond these generalities, relatively few studies have offered complete cost analyses comparing
eDNA methods to traditional surveys. For this paper, I conducted a detailed cost analysis of the equipment and labor costs for backpack electrofishing (traditional method), metabarcoding, and single-species eDNA detection. Tables 2 and 3 show the equipment and labor costs of each method. The price ranges shown in Table 2 are approximate but based on multiple sources (Appendix). The wages shown in Table 3 are based on a typical invoice for electrofishing field work, along with my own experience as a graduate researcher. The costs of sequencing and qPCR were determined under the assumption that samples would be sent out to a company for analysis. The actual cost of analyzing eDNA in the lab could be lower if the project is in collaboration with an academic or research institution.

This cost analysis was based on 100-meter single-pass electrofishing at 10 sites. This equated to 35 water samples (3 per site plus 5 field blanks) for eDNA analysis. Our protocol uses 2 paper filters per sample, yielding a final total of 70 eDNA samples. In general, the total costs of the three methods (equipment and labor) were relatively similar, with single-species detection being the least expensive (Table 2). As found in previous studies, the labor cost was much higher for electrofishing than for either eDNA method (Table 3). The equipment costs were also relatively similar across the three methods (Table 2). However, not all equipment can be reused in future surveys. In Table 2, equipment that must be purchased at the start of each new survey is highlighted in blue. The eDNA methods have more of these recurring expenses than traditional surveys. Resource managers will need to consider these costs if they are considering long-term eDNA monitoring efforts.
Table 2: Cost analysis of backpack electrofishing and two eDNA methods. Costs are based on 100-meter single-pass electrofishing at 10 sites and 35 water samples for eDNA. Highlighted cells indicate recurring expenses.

<table>
<thead>
<tr>
<th>Backpack Electrofishing</th>
<th>Metabarcoding</th>
<th>Single-species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment</td>
<td>Price (dollars)</td>
<td>Equipment</td>
</tr>
<tr>
<td>Electrofisher (batteries, charger, etc.)</td>
<td>5,000 – 7,000</td>
<td>1-liter collection bottles</td>
</tr>
<tr>
<td>Seine and dip nets</td>
<td>342 – 592</td>
<td>Vacuum holder assembly (funnel and base; metal clamp; glass frit filter support; 1-L flask)</td>
</tr>
<tr>
<td>Bubblers</td>
<td>44 – 52</td>
<td>Vacuum filtration pump</td>
</tr>
<tr>
<td>Tape measure</td>
<td>35 – 37</td>
<td>Filter papers</td>
</tr>
<tr>
<td>Collection and processing buckets</td>
<td>90 – 120</td>
<td>Lab consumables (tubes, pipet tips, bleach, etc.)</td>
</tr>
<tr>
<td>Processing nets</td>
<td>15 – 21</td>
<td>Extraction kit</td>
</tr>
<tr>
<td>Fishing permits</td>
<td>137 – 206</td>
<td>Lab usage and rental</td>
</tr>
<tr>
<td></td>
<td>Primmers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequencing (library prep, QC, and sequencing)</td>
<td>3710 – 5135</td>
</tr>
<tr>
<td>Total</td>
<td>5663 – 8028</td>
<td>Total</td>
</tr>
</tbody>
</table>
**Table 3:** Labor costs for backpack electrofishing and two eDNA methods. Costs are based on 100-meter single-pass electrofishing at 10 sites and 35 water samples for eDNA.

<table>
<thead>
<tr>
<th>Backpack Electrofishing</th>
<th>Number</th>
<th>Hours</th>
<th>Wage</th>
<th>Total Cost</th>
<th>eDNA methods</th>
<th>Number</th>
<th>Hours</th>
<th>Wage</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Scientist</td>
<td>1</td>
<td>48</td>
<td>75</td>
<td>3600</td>
<td>Principal Scientist</td>
<td>1</td>
<td>46 - 49</td>
<td>75</td>
<td>3450 – 3675</td>
</tr>
<tr>
<td>Fish Scientist/Processor</td>
<td>1</td>
<td>48</td>
<td>75</td>
<td>3600</td>
<td>Scientific technician</td>
<td>2 - 3</td>
<td>46 - 49</td>
<td>12</td>
<td>1104 - 1764</td>
</tr>
<tr>
<td>Scientific technician</td>
<td>4 - 7</td>
<td>48</td>
<td>12</td>
<td>2304 - 4032</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td>9504 - 11232</td>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td>4554 - 5439</td>
</tr>
</tbody>
</table>
Figure 5: Recommendations for selecting the optimal eDNA method.
In general, the single-species eDNA approach is slightly less expensive than metabarcoding, an observation validated by this cost analysis. This is largely because of the next-generation sequencing required by metabarcoding. We argue, however, that metabarcoding is often more cost effective because of the extensive information it provides about biological communities. In Figure 5, we provide guidelines to help resource managers choose the most cost effective eDNA approach for their needs. When monitoring one or two species, the single-species approach with qPCR can be highly effective. Primers and probes designed for single-species detection may even be able to amplify several closely related species, such as the invasive carp species. For studies of an entire taxonomic group, however, metabarcoding is more effective. Metabarcoding studies produce more reliable results and can offer insight into an entire biological community. Finally, one should consider the timeline of sample collection when deciding on the best method. Metabarcoding is more cost effective when used to analyze samples in bulk (around 96 at once), while single-species detection is beneficial if samples will trickle in over time.

1.6 CONCLUSIONS

Environmental DNA (eDNA) methods are quickly reshaping how we approach biological surveys. They have many advantages over traditional surveys. They are less invasive, require less sampling effort, and provide unique taxonomic information. They are especially useful for sites with low accessibility or when species are cryptic/difficult to identify. The two common eDNA approaches (metabarcoding and single-species detection) have advantages and disadvantages. They are suited to different applications. The single-species approach requires the design of specific qPCR probes and primers that detect one or two species. Metabarcoding uses
universal primers to detect many species within a taxonomic group. We argue that the metabarcoding approach is generally the more effective way to analyze eDNA because it provides extensive information on the biological community, is more robust to logistical errors, and provides higher confidence in species diagnosis and detection.

Other concerns addressed in this review were the quantification and ecology of eDNA. Reliable quantification of species is not yet possible using eDNA alone. Some studies have attempted to equate eDNA concentration with relative abundance, but even those estimations are affected by many confounding variables. There is still considerable uncertainty around the ecology (persistence and transport) of eDNA. Factors like stream flow, temperature, nutrient load, and biological oxygen demand (BOD) likely affect how fast DNA degrades and how it moves within an aquatic system.

Finally, financial concerns were addressed with an in-depth cost analysis of three survey methods: backpack electrofishing, metabarcoding, and single-species detection. We found that the total costs did not differ greatly, indicating that eDNA methods are becoming more affordable. Based on these findings, we conclude that eDNA is a valuable tool that can be used to supplement traditional survey methods. Furthermore, the circumstances and aims of a given project will determine the optimal application of eDNA methods. As molecular technology becomes more accessible, eDNA will vastly improve our understanding of biological communities, allowing greater protection of the environment.
2.1 Introduction

2.1.1 Freshwater Mussels (Unionidae)

Freshwater mussels are part of the phylum Mollusca, which also includes snails, slugs, squid, and octopuses. They are classified as bivalves because their bodies are contained within 2 shells that are hinged. Though they occur elsewhere, North America is unique in its high density of freshwater mussel species. With over 300 species, the United States is estimated to contain around one third of the entire world's freshwater mussel diversity. Freshwater mussels are divided into several taxonomic families. The class Bivalvia is split into two orders, Venerida and Unionida, each containing two main families. Within Venerida are Sphaeriidae (fingernail clams) and Cyrenidae, the family that includes the invasive Asian clam (Corbicula fluminea). In the order Unionida are Hyriidae and Unionidae. Of these four main families, Unionidae contains the greatest number of species, and most North American species are in this family. Therefore, this study will focus on Unionidae (Unionids).

Freshwater mussels (Unionidae) are fascinating animals with unique life histories. As adults, they are filter feeders, meaning they filter food particles from the water column, including phytoplankton, detritus, and diatoms. Adult mussels are primarily sedentary, but they do possess a muscular foot that allows them to move horizontally and vertically within the substrates of lakes, rivers, and streams. Freshwater mussels are able to colonize new areas by using a unique strategy of dispersal. Male mussels release sperm into the water column, which is then taken up by female mussels by the same mechanism used to filter feed. Fertilization, taking place inside the female, yields larvae called glochidia. To continue development, the glochidia must then become attached to the gills of a suitable host fish, where they will remain
while they metamorphosize into juvenile mussels. The juvenile mussels then drop off the host fish and may continue developing if they land in a suitable habitat. Many mussel species have only one suitable host fish species, meaning the persistence of freshwater mussels is often directly linked to the presence of certain fish species. Female mussels often have fascinating adaptations that increase the chance of their glochidia attaching to host fish, such as elaborately shaped mantles that resemble minnows. Once the host fish is lured close enough, the mussel ejects the glochidia at the fish, sometimes even temporarily trapping the fish's head in its shell cavity. Once fully mature, adult mussels can have very long lifespans, with some living more than 100 years.

Freshwater mussels are a highly imperiled group of organisms. Of the freshwater mussel species once found in North America, around 70% are at risk of extinction or are already extinct. In comparison, 39% of freshwater fish species, 16.5% of mammal species, and 14.6% of bird species in North America are extinct or imperiled. In Pennsylvania alone, there are 67 known species of freshwater mussels and more than 20 are endangered or possibly extirpated. Freshwater mussels are confronted with many challenges, including habitat loss/degradation, invasive species, and pollution. Mussel species prefer to anchor themselves in sandy gravel substrates along the shallow edges of medium to large streams and rivers. Mussel habitat can be degraded by actions that change the natural flow of waterways, including damming and channelization. Dams often transform flowing, oxygenated bodies of water into stagnant areas that are inhospitable to mussels. Dam releases also replace natural conditions with alternating periods of very high and very low water levels. By restricting the movement of host fish, dams can also prevent mussels from colonizing new habitats. Although dam removals may have beneficial long-term effects, mussel communities may take many years to recover from
these projects. More research is needed to fully understand the effects of human-made barriers on mussels and other wildlife.

Agriculture and urbanization can also degrade mussel habitat by increasing sedimentation and pollution. Impervious surfaces and reduced riparian cover cause soil to erode into rivers and streams, which may smother mussels or scour them from the substrate. Sedimentation also hinders mussels’ ability to filter feed by clouding the water column. Erosion from agricultural lands is especially harmful because it introduces pollutants like pesticides and excess nutrients. Studies of freshwater mussels indicate that species are negatively affected by agricultural factors, including increased nitrate levels. In addition, species richness was found to decrease with increasing total dissolved solids (TDS). Several fungicides, including chlorothalonil and pyraclostrobin, are extremely toxic to the glochidia and juveniles of at least one species (Lampsilis siliquoidea). The same species was also negatively affected by chronic exposure to the herbicide atrazine. Industrial waste, deicing salts, and brines from mining operations all contribute to runoff, increasing the salinity and overall ionic load of streams and rivers. This increased salinity can negatively affect freshwater mussels by decreasing glochidia viability and preventing attachment to host fish. Our understanding of how mussels are affected by pollutants is greatly limited, and more research is needed in this area.

Native freshwater mussels are also threatened by invasive mussel species, introduced to North America in ship ballast water and on the hulls of watercraft. One such invasive is the Zebra mussel (Dreissena polymorpha), which harms native mussel species by competing for space and food. Zebra mussels may also attach themselves to the shells of native mussels, thus directly inhibiting respiration and feeding. Another invasive species affecting native mussels is the Asian clam (Corbicula fluminea). The trophic niche (diet) of C. fluminea overlaps
with many native freshwater mussel species. This means native mussels must compete with the invasive species for critical food resources. In the presence of *C. fluminea* native mussels show lower growth and poorer physical condition. The glochidia of native mussels also show higher mortality in the presence of *C. fluminea*. Finally, the Round Goby (*Neogobius melanostomus*) is an invasive fish species that threatens native fauna in North America, especially the Great Lakes. The Round Goby poses a considerable threat to freshwater mussels, with reports of the fish consuming many native species. The spread of these invasive species is likely to increase the challenges faced by native mussels.

There are many reasons to care about native freshwater mussels. In addition to the intrinsic value possessed by all living things, mussels provide ecosystem services that benefit humans. As filter feeders, mussels naturally improve water quality by filtering out suspended solids and pollutants, such as nitrogen and heavy metals. This natural filtration system reduces the treatment costs needed to make water drinkable. According to one study, the mussel community in a 3.2-kilometer stretch filtered a volume of water equal to the total stream discharge at least two times every day. Because they are especially sensitive to toxic contamination, native mussels are effective indicators of water quality, with the absence of mussels suggesting chronic water pollution. Native mussels are also ecologically important as a source of food for raccoons, otters, muskrats, waterfowl, and some fish. Healthy mussel beds also serve as excellent habitat for fish and aquatic macroinvertebrates. Finally, North America is unique in being home to so many freshwater mussel species. Thus, prioritizing native mussel conservation is crucial for increasing biodiversity in general.

2.1.2 Sampling Methods
Effective protection of native freshwater mussels requires knowledge of where they naturally occur and where they are struggling to persist. This requires reliable monitoring to track changes in species abundance and range. Traditional mussel surveys rely on the visual recognition of specimens in the substrate of a waterbody. Surveyors may simply walk through the waterbody or use snorkel masks/SCUBA gear to dive for specimens. These visual surveys may consist of sampling in randomly selected quadrats or transects. In some cases, surveys are based on an established search time, during which surveyors simply move through the site and identify any mussels they find. These surveys provide crucial information about native mussels and help inform conservation decisions, but there are drawbacks to relying on visual and tactile methods. These methods usually require at least 3 or 4 team members for a site to be adequately sampled. Even with enough people, some areas often go unsampled if they are not easily accessible or the water is too deep/turbid to see the substrate. Though useful, visual-based sampling methods yield imperfect species detection and often require multiple time-consuming surveys to detect the majority of species at a site. In addition, mussels are sensitive to human handling. Handling mussels for visual surveys has been shown to cause reduced growth and may decrease individuals' long-term fitness. Finally, distinguishing between native mussel species can be quite difficult, making misidentification a major problem for visual surveys.

The difficulties associated with traditional survey methods make Unionids a great candidate for the development of novel environmental DNA (eDNA) methods. Surveys using eDNA have several advantages over traditional visual-based methods. They generally have equal or higher detection probabilities compared to traditional methods and provide accurate taxonomic identification. eDNA surveys are also less invasive to wildlife, require less
sampling effort, and enable sampling in areas with low physical accessibility. As stated earlier, the manual identification required in traditional surveys can be difficult, especially when individuals are damaged or morphologically cryptic. With eDNA, researchers can correctly identify species without needing in-depth knowledge of their morphological traits.

Some researchers have developed eDNA-based methods for the detection of Unionids, but these studies are still relatively rare. Most have taken the species-specific approach described earlier, in which primers detect one species or a few species within a given genera. In general, these species-specific primers have successfully amplified the target taxa but need to be optimized further. Fewer researchers have taken the metabarcoding approach for the study of mussels, in which universal primers are used to detect many species within a taxonomic group, such as *Unionidae*. Metabarcoding studies of Unionids have shown varying levels of success. In one such study, metabarcoding primers were able to successfully identify over 80% of the species previously found during visual surveys. Primers from another study were less effective, only detecting around half of the species previously identified with traditional surveys. The authors explained that some species were too taxonomically similar to be distinguished by the primers. In most studies that compare eDNA and traditional survey methods, both methods detect species that the other method does not, emphasizing the fact that eDNA will not likely replace traditional surveying, but augment it.

As explained in the previous section, metabarcoding has several inherent limitations, namely primer bias. Primer bias describes the tendency of universal primers to amplify some species more successfully than others. For many metabarcoding studies, scientists use/modify previously designed universal primers published by other researchers. This makes it crucial for the original creators of the primers to properly test them and describe the bias.
associated with them. Unfortunately, relatively few researchers complete this validation step. The most common way to characterize primer bias is by creating a mock mixture. A mock mixture consists of tissue-extracted DNA from morphologically identified species within the taxonomic group targeted by the primers. The mock mixture is amplified using the newly designed primers, and the products are sequenced. Most of the time, the tissue DNA from each target species is equalized to a standard concentration and added to the mixture in equal volume. This reduces confounding variables and allows researchers to test amplification bias alone. Without knowing the amplification bias associated with universal primers, it is impossible to accurately interpret data generated by metabarcoding studies of eDNA.

In this study, we introduce four internal primer pairs designed to amplify sections of the Cytochrome c oxidase subunit 1 (COI) gene in Unionids. The primer pairs were tested using an equimolar mock mixture of six target Unionid species. An external primer pair was designed and used to generate amplicons of a standard length for each mock mixture species. By using these amplicons to create the mock mixture, we hoped to standardize copy number, further reducing the confounding variables associated with amplification bias. The results of the primer validation will allow us to determine the least biased primer pair(s) and select them for future use in eDNA studies of Unionids. The findings will also enable more accurate interpretation of the data generated in those studies.

2.2 Methods

2.2.1 Tissue Extraction from Specimens

Six freshwater mussel (Unionid) species were used for the creation of an equimolar mock mixture: Lasmigona complanata (White heelsplitter), Lasmigona costata (Fluted shell),
*Obliquaria reflexa* (Three-horned wartyback), *Quadrula quadrula* (Mapleleaf), *Potamilus alatus* (Pink heelsplitter), and *Ligumia recta* (Black sandshell). All the species except for *L. costata* were collected from Dashields Pool in the Ohio River by Doug Locy and Arthur Bogon on August 26, 2014. DNA extractions for those species took place in June 2017. The *L. costata* specimen was collected from Pine Creek near North Park in Allegheny County by Brady Porter and Zachary Lane on October 26, 2017. DNA was extracted from that specimen in October 2017. All specimens were preserved in absolute ethanol after collection. For DNA extractions, a portion of tissue about 25 millimeters in size was cut from each individual’s foot. DNA was then extracted from the tissue using the procedure outlined for “Purification of Total DNA from Animal Tissues (Spin-Column Protocol)” in the Qiagen DNeasy Blood and Tissue Handbook. 165

2.2.2 Primer Design and Initial Amplification

To validate their ability to amplify species in the family Unionidae, we tested universal metabarcoding primers on a mock mixture of the 6 mussel species mentioned above. In 2017, Dr. Brady Porter designed 4 internal primer pairs for the Cytochrome c oxidase subunit 1 (COI) gene in the mitochondria of Unionids (Table 4). The PCR products (amplicons) produced by these primer pairs are around 300 base pairs long. Porter also designed a reverse primer (6R) downstream of the 4 internal pairs near the end of the COI gene (Table 5). We set out to design a forward primer that could be paired with 6R to amplify a section of DNA that contained all 4 primer pairs. Molecular Evolutionary Genetics Analysis (MEGA) software was used to design the primers used in this study. 166 Eighty-seven unique sequences for over 30 native Unionid species were retrieved from the NIH genetic sequence database, GenBank and aligned to each other in MEGA. Using the alignment, highly conserved regions were identified as potential
locations for the forward primer. A conserved region was identified in the Cytochrome c oxidase
subunit 2 (COII) gene around 240 base pairs upstream of the start of COI. A primer called
UCO2F (23 base pairs) was designed and aligned to this region using IUPAC nucleotide codes.
No degenerate bases were used. When paired, UCO2F and 6R amplified a section of DNA 866
base pairs long (Figure 6), which encompasses all 4 internal primer pairs.
Table 4: Internal primer combinations tested on the mock mixture. All were designed by Dr. Brady Porter, 2017. Underlined and italicized portion delineates the adaptor sequence used for Illumina MiSeq indexing.

<table>
<thead>
<tr>
<th>Combination</th>
<th>Sequence</th>
<th>Amplicon Length (base pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1F-1R</td>
<td>1F: TCGTCCGAGCGTCAGATGTGTTAAGAGACAGKTTCTYACWAAAYCATAARGATATTGG 1R: GTCTCGTGGGGCTCGGAGATGTGTTAAGAGACAGAHTYAVATTRTTYARYCGAGG</td>
<td>282</td>
</tr>
<tr>
<td>2F-2R</td>
<td>2F: TCGTCCGAGCGTCAGATGTGTTAAGAGACAGCCNGATATRKCTTTCYCTCG 2R: GTCTCGTGGGGCTCGGAGATGTGTTAAGAGACAGCAHACRAAYAHSKHATHCGYTC</td>
<td>292</td>
</tr>
<tr>
<td>3F-3R</td>
<td>3F: TCGTCCGAGCGTCAGATGTGTTAAGAGACAGGAYCARYTRTATAAYKRYATTGT 3R: GTCTCGTGGGGCTCGGAGATGTGTTAAGAGACAGCCMGCHARRRTGYAAGAAAAAAT</td>
<td>310</td>
</tr>
<tr>
<td>3F-4R</td>
<td>3F: TCGTCCGAGCGTCAGATGTGTTAAGAGACAGGAYCARYTRTATAAYKRYATTGT 4F: GTCTCGTGGGGCTCGGAGATGTGTTAAGAGACAGAHATCHACHSARRCHCCAGAATG</td>
<td>283</td>
</tr>
</tbody>
</table>

Figure 6: Cartoon of universal and internal primers across COI and COII. Each tick mark denotes 100 base pairs. All primers are in COI except for UCO2F, which lies near the end of COII.
PCR reactions with this external primer pair were carried out on the 6 mock mixture species: *L. complanata, L. costata, O. reflexa, L. recta, P. alatus, and Q. quadrula*. A negative control of PCR-grade water was also included. The total volume of the reactions was 50 microliters (μL). Each reaction included 2 μL of tissue-extracted template DNA at 50 nanograms per microliter, 2.5 μL of the forward and reverse primers (UCO2F and 6R) at 10 μM, 5X Phusion HF Buffer, 10 mM dNTPs, PCR-grade water, and Phusion DNA polymerase (Thermo Fisher Scientific). The reactions were then run on a thermocycler with the following conditions: 30 second hot start at 98 °C followed by 46 cycles of denaturation (98 °C for 5 seconds), annealing (49 °C for 20 seconds), and extension (72 °C for 15 seconds). Cycling was followed by a 5-minute final extension at 72 °C. Amplification was confirmed via gel electrophoresis using 2% agarose gel in SB buffer and visualized by ethidium bromide staining and UV illumination.

The 6 mock mixture species and a negative control were also amplified in 20 μL reactions using a different commercial polymerase: KAPA HiFi HotStart ReadyMix. Each reaction included 2 μL of template DNA at 50 nanograms per microliter, 0.75 μL of the forward and reverse primers at 10 μM, PCR-grade water, and KAPA HiFi DNA polymerase. (Thermo

**Table 5:** Universal primer combination used for initial amplification of mock mixture species. UCO2F was designed by Meredith Bennett for this study. 6R was designed by Dr. Brady Porter, 2017.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Amplicon Length (base pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCO2F</td>
<td>TCTGCTGATGTTATTCATGCTTG</td>
<td>866</td>
</tr>
<tr>
<td>6R</td>
<td>ATHGTRATRGCACCRGCYAAAAC</td>
<td></td>
</tr>
</tbody>
</table>

PCR reactions with this external primer pair were carried out on the 6 mock mixture species: *L. complanata, L. costata, O. reflexa, L. recta, P. alatus, and Q. quadrula*. A negative control of PCR-grade water was also included. The total volume of the reactions was 50 microliters (μL). Each reaction included 2 μL of tissue-extracted template DNA at 50 nanograms per microliter, 2.5 μL of the forward and reverse primers (UCO2F and 6R) at 10 μM, 5X Phusion HF Buffer, 10 mM dNTPs, PCR-grade water, and Phusion DNA polymerase (Thermo Fisher Scientific). The reactions were then run on a thermocycler with the following conditions: 30 second hot start at 98 °C followed by 46 cycles of denaturation (98 °C for 5 seconds), annealing (49 °C for 20 seconds), and extension (72 °C for 15 seconds). Cycling was followed by a 5-minute final extension at 72 °C. Amplification was confirmed via gel electrophoresis using 2% agarose gel in SB buffer and visualized by ethidium bromide staining and UV illumination.

The 6 mock mixture species and a negative control were also amplified in 20 μL reactions using a different commercial polymerase: KAPA HiFi HotStart ReadyMix. Each reaction included 2 μL of template DNA at 50 nanograms per microliter, 0.75 μL of the forward and reverse primers at 10 μM, PCR-grade water, and KAPA HiFi DNA polymerase. (Thermo
Fisher Scientific). The thermocycler protocol used for these reactions was named “KAPAMUSSEL” and consisted of the following steps: 30 second hot start at 98 °C followed by 46 cycles of denaturation (98 °C for 5 seconds), annealing (63 °C for 20 seconds), and extension (72 °C for 15 seconds). Cycling was followed by a 5-minute final extension at 72 °C. Again, amplification was confirmed via gel electrophoresis. The PCR products from the KAPA reactions were combined with those from the reactions with Phusion, resulting in a total of 60 uL of amplified DNA for each of the 6 species.

From the pooled amplified DNA for each species, 40 uL was cleaned using Mag-Bind Total Pure NGS magnetic beads. The DNA was cleaned according to the protocol for Left Side Size Selection in the SPRIselect User Guide. This is done to remove unincorporated primers and primer dimers from the amplicons. Based on the length of the desired amplicon (866 bp), the products were cleaned using a 0.6 ratio of beads to sample. The cleaned products were eluted to 25 uL in PCR-grade water.

2.2.3 Mock Mixture Creation and Internal Primer Validation

The cleaned products from the initial amplification with UCO2F and 6R were analyzed using a Qubit Fluorometer to determine DNA concentration. Three readings were taken on the Qubit and averaged for each sample. The samples were then run through a 4150 Tape Station System to confirm that the amplicons were the correct size and free from other bands. According to the Qubit, L. costata had the lowest average concentration at 4.34 ng/uL. Using the Qubit readings, the other 5 mock mixture species were diluted to 4.34 ng/uL. An equimolar mock mixture was created by combining 20 uL of each species at this concentration.
The mock mixture was then used to validate the 4 internal primer pairs (1F-1R, 2F-2R, 3F-3R, 3F-4R) and evaluate their performance. Three separate PCR runs, each in triplicate, were done for each primer pair, resulting in a total of 9 reactions for each pair. This was done to evaluate the stochastic variability associated with PCR. The total volume of all reactions was 20 uL. They consisted of KAPA HiFi DNA polymerase, 0.75 uL of the forward and reverse primers at 10 uM, PCR-grade water, and the desired amount of mock mixture. The amount of mock mixture added to each reaction was determined for each primer pair separately based on preliminary PCR runs testing various volumes. The optimal volume was determined using gel electrophoresis, choosing one that produced a bright, clean band. The “KAPAMUSSEL” thermocycler protocol described earlier was used for these reactions. For each PCR run, the 3 replicates were pooled, resulting in 12 samples (3 for each primer pair). The samples were given unique labels to denote the primer pair and the replicate. Primer pair 1F-1R was designated as “A”, 2F-2R was “B”, 3F-3R was “C”, and 3F-4R was “D”. The replicates were labeled as “1”, “2”, or “3”. Therefore, the first replicate done with 1F-1R, for example, was labeled “A1”.

2.2.4 Next-Generation Sequencing

The 12 samples were cleaned using Mag-Bind Total Pure NGS magnetic beads. Cleaning followed the protocol for Left Side Size Selection in the SPRIselect User Guide. Based on the length of the desired amplicon (around 300 bp), the products were cleaned using a 1.2 ratio of beads to sample. The cleaned samples were eluted to 20 uL in PCR-grade water. Each sample was then labeled with a unique set of indexing primers, which attach to the adaptor sequences on the internal primers (Illumina, cat. #FC-131-1002). The indexed samples were cleaned using a 1.2 bead-to-sample ratio and eluted to 20 uL. The samples were then measured on the Qubit and
pooled together based on similar concentrations, resulting in 5 tubes. These samples were run on the Tape Station to confirm the amplicon size was correct. Concentrations were measured on the Qubit again, and all samples were diluted to the same concentration and pooled. The pooled sample was measured on the Qubit and Tape Station to ensure quality. Finally, the pooled sample was loaded onto the Illumina MiSeq in Dr. Jan Janecka’s lab and run according to the protocol associated with Illumina’s MiSeq Reagent Micro Kit v2 (300 cycles).

The same 12 samples were run again in October using Illumina’s MiSeq Reagent Nano Kit v2 (500 cycles). Beginning with the indexing step, preparation of the samples for sequencing followed the same protocol described above for the run in August. The pooled sample was run according to the protocol associated with the Nano Kit v2 (500 cycles). Due to an error in setup, however, the sequencer only ran for 300 cycles. The purpose of this experiment was to determine the variability associated with separate sequencing runs. We also wanted to test whether the type of kit/number of cycles (Micro v2 vs. Nano v2) affected the sequencing data.

The software, QIIME 2 \(^{168}\) was used to analyze the sequence reads produced by the MiSeq runs. The primers were removed from the sequences using the “cutadapt” package. The sequences were processed as single-end reads, and the “DADA2” package was used to remove errors from the data. Using DADA2, the sequences were denoised, removing replicates and chimeras. The sequences were then clustered into operational taxonomic units (OTUs) based on 97% similarity. Each of the unique OTUs were identified to species level by entering the sequence into the Basic Local Alignment Search Tool (BLAST) \(^{169}\). The species for each OTU was identified as the top hit from BLAST, and the e-value (an indication of the probability of finding the sequence by random chance) was recorded. In the case of 2 equally likely results (e-value and percent identity were the same), both species were recorded as possibilities.
2.2.5 Data Analysis

Once the sequences were processed in QIIME 2, they were ready to be analyzed. All figures were created in Microsoft Excel, and the statistical analyses were conducted in the programming language, R.  

For each PCR replicate, the total number of sequences for each mock mixture species was found by adding the sequences generated by the forward and reverse primer. Bar graphs were created to show how the number of sequences for each species varied across the 3 replicates for each primer pair. Analysis of variance (ANOVA) tests were conducted in R for each primer pair by comparing the average number of sequences produced by mock mixture species in each replicate.

The sequences from the 3 replicates were summed to yield the total number of sequences identified to each mock mixture species. This was done for each primer pair. The proportion of sequences represented by each species was calculated and graphed alongside a hypothetical even distribution for each primer pair. In addition, a Chi-square test was performed in R to determine whether the observed proportions significantly differed from an even distribution. Again, this was done separately for each primer pair. The data from the second sequencing run in October was prepared in the same way. The October species distributions were compared to the first sequencing run using bar graphs and Chi-square tests.

The number of sequences identified to each mock mixture species was determined for each individual forward and reverse primer. Using this information, the species proportions for each primer were calculated and graphed. Chi-square tests were performed in R to determine
whether the species proportions significantly differed between the forward and reverse primer for each primer pair.

2.3 Results

2.3.1 Overview of Results from 1st Sequencing Run (August)

Each of the four internal primer pairs were able to successfully detect at least five of the six mock mixture species. After trimming the primers, the generated sequences were between 126 – 128 base pairs for each forward and reverse primer. Nearly 100% of all the generated sequences were identified to mock mixture species (Table 7). For the identifications of mock mixture species, the E values were all below 1.0 x e^{-50}.

Initially, BLAST identified *L. complanata* sequences as *Arcidens wheeleri/Arcidens confragosus* for forward primer 1 (1F). For the purposes of data analysis, the *Arcidens* sequences produced by 1F were attributed to *L. complanata*. This is justified by the fact that the mock mixture composition was known and because the number of sequences identified as *Arcidens* roughly matched the number of sequences identified as *L. complanata* by 1R. As of December 19, 2022, however, BLAST correctly identified the forward sequences as *L. complanata*. In addition, BLAST could not distinguish between *Potamilus alatus* and *Potamilus purpuratus* for several of the primers. The percent identity and E values were the same for the two species. For data analysis, these sequences were attributed to *P. alatus*.

2.3.2 Stochasticity Across PCR Replicates

Stochastic variability (stochasticity) is the random variability associated with the removal of a mix-template aliquot from the mock mixture, and the PCR process itself. In any given PCR
reaction, certain species may be amplified more successfully than others. To test stochastic bias in this experiment, we compared the number of sequences identified to each mock mixture species across three PCR replicates of each primer pair. Table 6 shows the sequence distributions along with the results of ANOVA tests for each primer pair. The ANOVA tests indicated no statistically significant change in the number of sequences across the three PCR replicates. Figures 7 – 10 illustrate this relationship for each primer pair.
<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. complanata</td>
<td>5942</td>
<td>5202</td>
<td>4157</td>
<td>306</td>
<td>343</td>
<td>247</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>18</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>O. reflexa</td>
<td>4419</td>
<td>4209</td>
<td>3694</td>
<td>1494</td>
<td>2012</td>
<td>1834</td>
<td>148</td>
<td>89</td>
<td>148</td>
<td>139</td>
<td>56</td>
<td>75</td>
</tr>
<tr>
<td>L. recta</td>
<td>3828</td>
<td>4674</td>
<td>3265</td>
<td>582</td>
<td>560</td>
<td>503</td>
<td>4722</td>
<td>2497</td>
<td>4600</td>
<td>349</td>
<td>188</td>
<td>185</td>
</tr>
<tr>
<td>P. alatus</td>
<td>2146</td>
<td>2437</td>
<td>2103</td>
<td>749</td>
<td>850</td>
<td>714</td>
<td>1749</td>
<td>946</td>
<td>1724</td>
<td>2544</td>
<td>1176</td>
<td>1211</td>
</tr>
<tr>
<td>L. costata</td>
<td>128</td>
<td>149</td>
<td>133</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>12</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Q. quadrula</td>
<td>100</td>
<td>92</td>
<td>54</td>
<td>2599</td>
<td>3714</td>
<td>3443</td>
<td>8771</td>
<td>5113</td>
<td>8774</td>
<td>11818</td>
<td>5367</td>
<td>5729</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F-statistic</td>
<td>0.1262</td>
<td>0.1018</td>
<td>0.2789</td>
<td>0.3288</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>0.8823</td>
<td>0.904</td>
<td>0.7614</td>
<td>0.7248</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>2 and 15</td>
<td>2 and 12</td>
<td>2 and 12</td>
<td>2 and 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 6:** Number of sequences identified to each mock mixture species for each PCR replicate. Primer pair 1F-1R is denoted by A, 2F-2R is B, 3F-3R is C, and 3F-4R is D. F statistics indicate the results of ANOVA tests comparing the average number of sequences across 3 PCR replicates.
Figure 7: Number of sequences identified to each mock mixture species for three PCR replicates with primer pair 1F-1R (A).

Figure 8: Number of sequences identified to each mock mixture species for three PCR replicates with primer pair 2F-2R (B).
**Figure 9:** Number of sequences identified to each mock mixture species for three PCR replicates with primer pair 3F-3R (C). *L. costata* was amplified in 2 out of 3 replicates but was removed for easier viewing (see Table 6).

**Figure 10:** Number of sequences identified to each mock mixture species for three PCR replicates with primer pair 3F-4R (D). *L. complanata* was amplified in all 3 replicates and *L. costata* was amplified in 1 of 3 but were removed for easier viewing (see Table 6).
2.3.3 Amplification Bias of Internal Primer Pairs

The universal primers used for metabarcoding are subject to amplification bias because they are designed to be degenerate and capable of amplifying many species. We determined the bias associated with our four internal primer pairs by testing them on an equimolar mock mixture. Table 7 shows the sequence distributions along with the results of Chi-Squared tests for each primer pair. The Chi-Squared values represent the degree to which the sequence distribution differed from an even distribution (equal representation of each mock mixture species). The Chi-Squared tests indicated that the sequence distributions for each primer pair differed significantly from an even distribution, with primer pairs 3F-3R and 3F-4R differing the most. Figure 11 illustrates the sequence distributions for each primer pair compared to the even distribution.
Table 7: Number of sequences identified to each mock mixture species for each primer pair. Chi-Squared values indicate the degree to which the observed sequence distribution differed from an even distribution where each mock mixture species was equally represented.

<table>
<thead>
<tr>
<th>Species</th>
<th>1F-1R</th>
<th>2F-2R</th>
<th>3F-3R</th>
<th>3F-4R</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Lasmigona complanata</em></td>
<td>15301</td>
<td>896</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td><em>Obliquaria reflexa</em></td>
<td>12322</td>
<td>5340</td>
<td>385</td>
<td>270</td>
</tr>
<tr>
<td><em>Ligumia recta</em></td>
<td>11767</td>
<td>1645</td>
<td>11819</td>
<td>722</td>
</tr>
<tr>
<td><em>Potamilus alatus</em></td>
<td>6686</td>
<td>2313</td>
<td>4419</td>
<td>4931</td>
</tr>
<tr>
<td><em>Lasmigona costata</em></td>
<td>410</td>
<td>0</td>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td><em>Quadrula quadrula</em></td>
<td>246</td>
<td>9756</td>
<td>22658</td>
<td>22914</td>
</tr>
</tbody>
</table>

| Total Mock Mixture Sequences | 46732  | 19950  | 39297 | 28903 |
| Mock Mixture Sequences + Unidentified Sequences | 46732  | 20102  | 39656 | 28911 |
| Percent Mock Mixture Sequences | 100    | 99     | 99    | 100   |
| Chi-Squared ($X^2$) Value   | 38.686 | 46.601 | 75.078 | 99.029 |
| P-value                    | 2.7e-7 | 6.8e-9 | 9.0e-15 | 2.2e-16 |
| Degrees of Freedom         | 5      | 5      | 5     | 5     |
The sequence distributions above were determined for each primer pair by combining the sequences produced by the forward and reverse primer. In addition to this general bias, we wanted to determine if the individual forward and reverse primers produced significantly different results. Table 8 shows the sequence distributions along with the results of Chi-Squared tests for each primer. The Chi-Squared values represent the degree to which the sequence distributions differed between the forward and reverse primer for each pair. To improve the accuracy of the Chi-Squared tests, a species was removed from the test if the proportion of sequences identified to that species was 0. This can be seen in Table 8 by observing the degrees of freedom for each primer pair. The degrees of freedom were calculated as the number of species included in the test minus 1. The tests indicated that the sequence distributions differed significantly for primer pairs 3F-3R and 3F-4R but not for 1F-1R and 2F-2R. Figure 12 illustrates the sequence distributions for each primer.

**Figure 11:** Proportion of sequences identified to each mock mixture species for each primer pair and for an even distribution (Expected).
**Table 8:** Number of sequences identified to each mock mixture species for each individual primer. Chi-Squared values indicate the degree to which the observed sequence distributions differed between the forward and reverse primer for each pair.

<table>
<thead>
<tr>
<th>Species</th>
<th>1F</th>
<th>1R</th>
<th>2F</th>
<th>2R</th>
<th>3F</th>
<th>3R</th>
<th>3F</th>
<th>4R</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Lasmigona complanata</em></td>
<td>7667</td>
<td>7634</td>
<td>423</td>
<td>473</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td><em>Obliquaria reflexa</em></td>
<td>6310</td>
<td>6012</td>
<td>2919</td>
<td>2421</td>
<td>199</td>
<td>186</td>
<td>138</td>
<td>132</td>
</tr>
<tr>
<td><em>Ligumia recta</em></td>
<td>5850</td>
<td>5917</td>
<td>843</td>
<td>802</td>
<td>6058</td>
<td>5761</td>
<td>378</td>
<td>344</td>
</tr>
<tr>
<td><em>Potamilus alatus</em></td>
<td>3449</td>
<td>3237</td>
<td>1296</td>
<td>1017</td>
<td>4318</td>
<td>101</td>
<td>4796</td>
<td>135</td>
</tr>
<tr>
<td><em>Lasmigona costata</em></td>
<td>212</td>
<td>198</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>6</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td><em>Quadrula quadrula</em></td>
<td>118</td>
<td>128</td>
<td>5390</td>
<td>4366</td>
<td>10758</td>
<td>11900</td>
<td>11424</td>
<td>11490</td>
</tr>
</tbody>
</table>

**Chi-Squared (X^2) Value**  
0.088343  0.24926  20.747  30.861

**P-value**  
0.9999  0.9928  1.188e-4  9.98e-06

**Degrees of Freedom**  
5  4  3  5
As in August, each of the four internal primer pairs were able to successfully detect at least five of the six mock mixture species. After trimming the primers, the generated sequences were between 126 – 128 base pairs for each forward and reverse primer. Nearly 100% of all the generated sequences were identified to mock mixture species (Table 9). For the identifications of mock mixture species, the E values were all below $1.0 \times 10^{-50}$.

Table 9 shows the sequence distributions along with the results of Chi-Squared tests for each primer pair in October. The Chi-Squared values represent the degree to which the sequence distribution differed from the hypothetical even distribution. The Chi-Squared tests indicated that the sequence distributions for each primer pair differed significantly from an even distribution.

2.3.4 Results from 2nd Sequencing Run (October) and Comparison with August

As in August, each of the four internal primer pairs were able to successfully detect at least five of the six mock mixture species. After trimming the primers, the generated sequences were between 126 – 128 base pairs for each forward and reverse primer. Nearly 100% of all the generated sequences were identified to mock mixture species (Table 9). For the identifications of mock mixture species, the E values were all below $1.0 \times 10^{-50}$.

Table 9 shows the sequence distributions along with the results of Chi-Squared tests for each primer pair in October. The Chi-Squared values represent the degree to which the sequence distribution differed from the hypothetical even distribution. The Chi-Squared tests indicated that the sequence distributions for each primer pair differed significantly from an even distribution.

---

**Figure 12:** Proportion of sequences identified to each mock mixture species for each individual primer.
We wanted to determine if the two sequencing runs produced significantly different results. Table 10 shows the sequence distributions along with the results of Chi-Squared tests for each primer pair in August vs. October. The Chi-Squared values represent the degree to which the sequence distributions differed between the two sequencing runs for each pair. A species was removed from the Chi-Squared test if the proportion of sequences identified to that species was 0. The degrees of freedom were calculated as the number of species included in the test minus 1. The tests indicated that the sequence distributions did not differ significantly across the two sequencing runs for any of the primer pairs. Figure 13 illustrates the sequence distributions for each primer pair in August vs. October.

**Table 9:** Number of sequences identified to each mock mixture species for each primer pair. Chi-Squared values indicate the degree to which the observed sequence distribution differed from an even distribution where each mock mixture species was equally represented.

<table>
<thead>
<tr>
<th></th>
<th>1F-1R</th>
<th>2F-2R</th>
<th>3F-3R</th>
<th>3F-4R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lasmigona complanata</strong></td>
<td>9265</td>
<td>724</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td><strong>Obliquaria reflexa</strong></td>
<td>7514</td>
<td>4764</td>
<td>108</td>
<td>227</td>
</tr>
<tr>
<td><strong>Ligumia recta</strong></td>
<td>6835</td>
<td>1433</td>
<td>4179</td>
<td>596</td>
</tr>
<tr>
<td><strong>Potamilus alatus</strong></td>
<td>4185</td>
<td>1991</td>
<td>1652</td>
<td>4708</td>
</tr>
<tr>
<td><strong>Lasmigona costata</strong></td>
<td>210</td>
<td>0</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td><strong>Quadrula quadrula</strong></td>
<td>84</td>
<td>8703</td>
<td>9145</td>
<td>23025</td>
</tr>
</tbody>
</table>

**Table 10:** Number of sequences identified to each mock mixture species for each primer pair. Chi-Squared values indicate the degree to which the observed sequence distribution differed from an even distribution where each mock mixture species was equally represented.

<table>
<thead>
<tr>
<th></th>
<th>1F-1R</th>
<th>2F-2R</th>
<th>3F-3R</th>
<th>3F-4R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lasmigona complanata</strong></td>
<td>9265</td>
<td>724</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td><strong>Obliquaria reflexa</strong></td>
<td>7514</td>
<td>4764</td>
<td>108</td>
<td>227</td>
</tr>
<tr>
<td><strong>Ligumia recta</strong></td>
<td>6835</td>
<td>1433</td>
<td>4179</td>
<td>596</td>
</tr>
<tr>
<td><strong>Potamilus alatus</strong></td>
<td>4185</td>
<td>1991</td>
<td>1652</td>
<td>4708</td>
</tr>
<tr>
<td><strong>Lasmigona costata</strong></td>
<td>210</td>
<td>0</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td><strong>Quadrula quadrula</strong></td>
<td>84</td>
<td>8703</td>
<td>9145</td>
<td>23025</td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Mock Mixture Sequences</strong></td>
<td>28093</td>
<td>17615</td>
<td>15090</td>
<td>28593</td>
</tr>
<tr>
<td><strong>Total Sequences</strong></td>
<td>28093</td>
<td>17634</td>
<td>15090</td>
<td>28648</td>
</tr>
<tr>
<td><strong>Percent Mock Mixture Sequences</strong></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td><strong>Chi-Squared (X^2) Value</strong></td>
<td>57.175</td>
<td>102.88</td>
<td>174.96</td>
<td>305.18</td>
</tr>
<tr>
<td><strong>P-value</strong></td>
<td>4.6e-11</td>
<td>&lt;2.2e-16</td>
<td>&lt;2.2e-16</td>
<td>&lt;2.2e-16</td>
</tr>
<tr>
<td><strong>Degrees of Freedom</strong></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 10: Number of sequences identified to each mock mixture species for each primer pair in the two sequencing runs (August vs. October). Chi-Squared values indicate the degree to which the observed sequence distributions differed between the two runs.

<table>
<thead>
<tr>
<th>Species</th>
<th>1F-1R (Aug)</th>
<th>1F-1R (Oct)</th>
<th>2F-2R (Aug)</th>
<th>2F-2R (Oct)</th>
<th>3F-3R (Aug)</th>
<th>3F-3R (Oct)</th>
<th>3F-4R (Aug)</th>
<th>3F-4R (Oct)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Lasmigona complanata</em></td>
<td>15301</td>
<td>9265</td>
<td>896</td>
<td>724</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>24</td>
</tr>
<tr>
<td><em>Obliquaria reflexa</em></td>
<td>12322</td>
<td>7514</td>
<td>5340</td>
<td>4764</td>
<td>385</td>
<td>108</td>
<td>270</td>
<td>227</td>
</tr>
<tr>
<td><em>Ligumia recta</em></td>
<td>11767</td>
<td>6835</td>
<td>1645</td>
<td>1433</td>
<td>11819</td>
<td>4179</td>
<td>722</td>
<td>596</td>
</tr>
<tr>
<td><em>Potamilus alatus</em></td>
<td>6686</td>
<td>4185</td>
<td>2313</td>
<td>1991</td>
<td>4419</td>
<td>1652</td>
<td>4931</td>
<td>4708</td>
</tr>
<tr>
<td><em>Lasmigona costata</em></td>
<td>410</td>
<td>210</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>6</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td><em>Quadrula quadrula</em></td>
<td>246</td>
<td>84</td>
<td>9756</td>
<td>8703</td>
<td>22658</td>
<td>9145</td>
<td>22914</td>
<td>23025</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chi-Squared ($X^2$) Value</th>
<th>0.14695</th>
<th>0.038454</th>
<th>0.1951</th>
<th>0.14227</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-value</td>
<td>0.9996</td>
<td>0.9998</td>
<td>0.9784</td>
<td>0.9996</td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
2.3.5 Comparing Results to a Previous Study of Amplification Bias

A previous study conducted by Zachary Lane in 2017 investigated the amplification bias of the four internal primer pairs mentioned here. As in this study, an equimolar mock mixture of 6 Unionid species was created and used to test the primers. Lane’s mock mixture, however, consisted of genomic DNA from each species (original DNA extraction). We used an outside primer pair (UCO2F and 6R) to generate PCR products and created the mock mixture using those amplicons. We theorized that this independent replication step could help standardize the number of copies of the target gene for each species. To test this theory, we compared our sequence distributions with Lane’s results to determine whether they differed significantly.

Lane used paired-end analysis to explain the sequencing data generated in his study. In this study, we used single-end analysis. For single-end analysis, QIIME was used to extract the
sequences generated by the forward and reverse primers separately. To generate paired-end reads, the sequences generated by the forward and reverse primers must be joined together to produce a contig for each primer pair. There must be a sufficient region of overlap between the separate forward and reverse primer sequences if they are to be joined together reliably. According to a QIIME error message, the region of overlap was only sufficient for two of our internal primer pairs: 1F-1R and 3F-4R. The raw sequences for these two primer pairs were analyzed as paired-end reads and compared to Lane’s results for the same pairs. After trimming the primers, the paired-end sequences in our study were between 235 – 238, and 100% of all the generated sequences were identified to mock mixture species. Finally, E values for the identifications of mock mixture species were much lower than those generated by single-end analysis. All values were below 1.0 x e^{-112}.

Table 11 compares Lane’s sequence distributions with our own for the two primer pairs submitted to paired-end analysis (1F-1R and 3R-4R). The Chi-Squared values in light blue represent the degree to which the sequence distribution for each primer pair differed from a hypothetical even distribution. In both studies, the Chi-Squared tests indicated that the sequence distributions for both primer pairs differed significantly from an even distribution. The Chi-Squared values in dark blue represent the degree to which the sequence distributions for the two primer pairs differed between our study and Lane’s. The Chi-Squared tests indicated that the distribution differed significantly for 1F-1R but not for 3F-4R. Figure 14 illustrates the sequence distributions for primer pair 1F-1R in our study vs. Lane’s, and figure 15 shows the distributions for 3F-4R in both studies.
Five of the mock mixture species were common to Lane’s study and this one (Table 11). Instead of *Lasmigona costata*, the sixth species in Lane’s mixture was *Fusconaia flava*. Therefore, “Other” was used to represent the sequences belonging to *L. costata* or *F. flava*.

Table 11: Number of paired-end sequences identified to each mock mixture species for two primer pairs in this study (Bennett) vs. Zachary Lane’s study in 2017. Chi-Squared values in light blue indicate the degree to which the observed sequence distributions differed from an even distribution. Chi-Squared values in dark blue indicate the degree to which the observed sequence distributions differed between the two studies.

<table>
<thead>
<tr>
<th>Species</th>
<th>1F-1R (Bennett)</th>
<th>1F-1R (Lane)</th>
<th>3F-4R (Bennett)</th>
<th>3F-4R (Lane)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Lasmigona complanata</em></td>
<td>2421</td>
<td>1074</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td><em>Obliquaria reflexa</em></td>
<td>1932</td>
<td>331</td>
<td>42</td>
<td>92</td>
</tr>
<tr>
<td><em>Ligumia recta</em></td>
<td>1854</td>
<td>463</td>
<td>110</td>
<td>36</td>
</tr>
<tr>
<td><em>Potamilus alatus</em></td>
<td>965</td>
<td>23</td>
<td>43</td>
<td>6</td>
</tr>
<tr>
<td><em>Quadrula quadrula</em></td>
<td>36</td>
<td>1293</td>
<td>3780</td>
<td>6469</td>
</tr>
<tr>
<td>Other (Unassigned reads)</td>
<td>57</td>
<td>57</td>
<td>4</td>
<td>754</td>
</tr>
<tr>
<td>Total Sequences</td>
<td>7265</td>
<td>3241</td>
<td>3985</td>
<td>7366</td>
</tr>
<tr>
<td>Chi-Squared ($X^2$) Value</td>
<td>39.573</td>
<td>43.081</td>
<td>124.97</td>
<td>112.14</td>
</tr>
<tr>
<td>P-value</td>
<td>1.8e-7</td>
<td>3.6e-8</td>
<td>&lt;2.2e-16</td>
<td>&lt;2.2e-16</td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Chi-Squared ($X^2$) Value  
43.081  
P-value  
3.6e-8  
Degrees of Freedom  
5
**Figure 14:** Proportion of sequences identified to each mock mixture species for primer pair 1F-1R in this study (Bennett) vs. Lane’s. For reference, “Bennett Sep.” represents the proportions generated by single-end analysis.

**Figure 15:** Proportion of sequences identified to each mock mixture species for primer pair 3F-4R in this study (Bennett) vs. Lane’s. For reference, “Bennett Sep.” represents the proportions generated by single-end analysis.
2.4 Discussion

In general, the sequences generated from both MiSeq runs were of high quality. For all 4 primer pairs, nearly 100% of the sequences were identified to mock mixture species. In addition, the e-values were appropriately low, indicating strong identifications. The sequences were generally the correct size according to our predicted amplicon length.

Despite the high quality of our sequences, there were some uncertainties around the identifications provided by GenBank’s BLAST tool. Originally, BLAST identified *Lasmigona complanata* as *Arcidens confragosus* or *Arcidens wheelerii*. The genus *Arcidens* is within the same subfamily as *Lasmigona*, which may explain the misidentification. When these sequences were rerun in December 2022, BLAST correctly identified them as *L. complanata*. This is because new sequences of *L. complanata* were added to the database in November (OM736871.1) and December (NC_068679.1) of 2022. In addition, BLAST could not distinguish between *Potamilus alatus* and a related species, *Potamilus purpuratus*. Our mixture used *P. alatus*, but BLAST identified *P. purpuratus* as an equally likely match. These misidentifications were likely due to an insufficient reference database for Unionid mussels. GenBank lacks reference sequences for many Unionid species, and the species with more entries are more likely to be successfully identified. This uncertainty was not a major problem for this study because we already knew the 6 species in our mock mixture. For studies of unknown sequences, however, as in the use of eDNA, the incompleteness of GenBank will make reliable identifications difficult. Therefore, more researchers should contribute mussel sequences to public databases like GenBank.

The first aim of our study was to determine the stochastic bias associated with PCR amplification of the mock mixture. Stochastic bias or stochasticity is the random variability
associated with PCR amplification and subsampling from a complex mixture of templates. Certain species may be amplified more successfully than others in a given PCR reaction due to random circumstances. For each primer pair, three PCR replicates were compared. Those replicates were themselves composed of three separate rounds of PCR pooled together. We compared the number of sequences identified to each mock mixture species across the three PCR replicates. The number of sequences identified to each species did not differ significantly between the replicates for any of the primer pairs. This indicates that stochastic bias may be decreased substantially by conducting all PCR amplifications in triplicate and pooling before sequencing. Stochastic bias is likely to decrease steadily as more and more replicates are performed and pooled together.

2.4.1 Amplification Bias

The second aim of this study was to determine the amplification bias associated with the 4 internal primer pairs. If the primers were completely unbiased, we would expect equal representation of each mock mixture species in the sequence results. We expected our primers to have some inherent bias, and therefore produce uneven sequence distributions. Our results did indicate substantial amplification bias for all 4 primer pairs. The sequence distributions were significantly different than a hypothetical even distribution of the 6 species, with primer pairs 3 and 4 (3F-3R and 3F-4R) having the most bias. Only primer pair 1 successfully amplified all 6 species across all 3 PCR replicates. In addition, primer pair 1 was the most successful at amplifying the *Lasmigona* species: *L. complanata* and *L. costata*. Both species were poorly represented by the other 3 primer pairs, especially *L. costata*. Primer pairs 2, 3, and 4 were all
highly biased towards *Quadrula quadrula*. In fact, *Q. quadrula* made up nearly half or more than half of the sequences produced by those primer pairs.

The bias against *Lasmigona spp.* may be explained by the fact that all other genera represented in the mock mixture belong to the subfamily Ambleminae. *Lasmigona* alone belongs to Anodontinae, which is cladistically distinct from Ambleminae (Figure 16). Therefore, it makes sense that the two *Lasmigona* species were poorly represented by primers based largely on Ambleminae species. In future studies, it may be wise to design more specific primers for each Unionid subfamily and use them in tandem.
Anodontinae (includes genus Lasmigona)

Unioninae

Rectidentinae

Gonideinae

Ambleminae (includes genera Ligumia, Potamilus, Quadrula, & Obliquaria)

Parreysiinae

Figure 16: Phylogeny of Unionidae showing the 6 major subfamilies. The two subfamilies represented in this study are highlighted. Adapted from Phylogeny of the most species-rich freshwater bivalve family (Bivalvia: Unionida: Unionidae): Defining modern subfamilies and tribes by Lopes-Lima et al. 2017
We also compared the sequence distributions for the individual forward and reverse primers for each pair. In other words, did the forward primers have similar amplification bias to their corresponding reverse primers? No significant difference was found between the forward and reverse primers for pairs 1 and 2. For pairs 3 and 4, however, the sequence distributions were significantly different for the forward and reverse primers. This difference was largely due to a large dropout of *Potamilus alatus* sequences in the reverse primers for each pair (3R and 4R).

To explain amplification bias, we looked at the number and nature of base pair mismatches between each primer and the 6 mock mixture species (Table 12). Due to insufficient reference sequences for Unionids in GenBank, it was impossible to determine the number of mismatches between forward primer 1 and some species. We were unable to find any reference sequences of *Ligumia recta* or *Lasmigona costata* that overlapped with the region containing 1F. The general sparseness of reference sequences for this region may explain why 1F identified *Lasmigona complanata* sequences as *Arcidens spp* (Figure 17).

The number of mismatches often differed between haplotypes of the same species. In these cases, a range was recorded of the highest and lowest number of mismatches found among the haplotypes. In addition, base pair mismatches in the 3’ end of a primer are notable, because they are more likely to cause PCR to fail than mismatches in the middle or 5’ end of the primer. These mismatches are especially inhibitory to a primer’s ability to amplify the species’ DNA. In Table 12, parentheses indicate the number of mismatches within 5 base pairs of the 3’ end of the primer.

The major patterns of amplification bias can be explained by these base pair mismatches. For example, there are two mismatches between *Quadrula quadrula* and primer pair 1,
explaining why 1F-1R generated so few sequences for *Q. quadrula* compared to the other three pairs. The mismatches can also explain the differences observed between the forward and reverse primers for pairs 3 and 4. Both reverse primers (3R and 4R) have mismatches near their 3’ end for several species (Figure 18), including *P. alatus*, which may contribute to the dropout of sequences for that species.

When sequences are analyzed as paired-end reads, the forward and reverse sequences are combined. We were unable to process our sequences as paired-end reads, because there was insufficient overlap for QIIME software to combine the forward and reverse sequences. Instead, we performed single-end analysis, in which the forward and reverse sequences are processed separately. This unforeseen roadblock allowed us to understand the unique biases of the forward and reverse primers in each pair, which turned out to be important information. Although paired-end analysis provides more reliable identifications (lower e-values), it provides less overall information, because sequences are limited by the primer (forward or reverse) that amplifies the species less effectively. In our case, paired-end analysis would have shown very few sequences of *P. alatus* for primer pairs 3 and 4 solely because of the reverse primers. We would not have known that the forward primer amplified a much larger number of *P. alatus* sequences. This information is crucial for the accurate interpretation of sequence results and could help guide improvements to the study. Namely, more effective reverse primers could be designed and paired with the current forward primer, yielding a more accurate picture of the abundance of *P. alatus*. 
**Table 12**: Number of base pair mismatches between the mock mixture species and the internal primers designed for this study. Numbers are based on multiple entries from GenBank for each species. Question marks indicate no reference sequences containing the primer region. Ranges indicate the highest and lowest number of mismatches for haplotypes of the species. Parentheses indicate the number of mismatches that occur within 5 base pairs of the 3’ end of the primer.

<table>
<thead>
<tr>
<th>Species</th>
<th>1F</th>
<th>1R</th>
<th>2F</th>
<th>2R</th>
<th>3F</th>
<th>3R</th>
<th>4R</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Lasmigona complanata</em></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0-1</td>
<td>0</td>
</tr>
<tr>
<td><em>Obliquaria reflexa</em></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>Ligumia recta</em></td>
<td>0?</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0-1(1)</td>
<td>1 (1)</td>
</tr>
<tr>
<td><em>Potamilus alatus</em></td>
<td>0</td>
<td>0</td>
<td>0-1</td>
<td>0-1</td>
<td>0-1</td>
<td>0-1(1)</td>
<td>0-1</td>
</tr>
<tr>
<td><em>Lasmigona costata</em></td>
<td>0?</td>
<td>0-1</td>
<td>0-2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>Quadrula quadrula</em></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0-1(1)</td>
<td>0</td>
</tr>
</tbody>
</table>
**Figure 17:** Alignment of forward primer 1 (1F) with reference sequences of the 6 mock mixture species. Reference sequences were taken from GenBank. Black circles indicate meaningful base pair mismatches.

**Figure 18:** Alignment of reverse primer 3 (3R) with reference sequences of the 6 mock mixture species. Reference sequences were taken from GenBank. Black circles indicate meaningful base pair mismatches.
2.4.2 Amplification Bias Across Separate Sequencing Runs

Another aim of this study was to determine whether sequence distributions varied across separate sequencing runs. The same mock mixture samples (3 replicates per primer pair) were sequenced again in October 2022. Instead of Illumina’s MiSeq Reagent Micro Kit v2, which consists of 300 cycles, we ran the samples with the Nano Kit v2, which runs 500 cycles. We expected the Nano kit to produce longer forward and reverse sequences, providing the necessary overlap to analyze them as paired-end reads. Due to an oversight during setup, however, the sequencer only ran for 300 cycles, meaning we could not evaluate the effect of the Nano kit on sequencing results.

In general, the second sequencing run produced fewer overall sequences, which can be explained by the use of a smaller kit (Nano kit v2). The Nano kit produces about ¼ the reads of the Micro kit. While the number of sequences decreased, the patterns of bias did not differ between the two sequencing runs. However, the Chi-square values suggest the biases were more pronounced for the second sequencing run. This is likely due to the overall decrease in the number of sequences generated by the primer pairs. The Nano kit produces longer but fewer sequence reads. Therefore, researchers should consider their priorities when choosing the type of sequencing kit for their project.

2.4.3 Comparison with Previous Mock Mixture Study

Mock mixture experiments are relatively common in studies of eDNA metabarcoding. Usually, these mixtures consist of equal concentrations of genomic DNA from each species. However, equal concentration does not necessarily ensure equal copy numbers of the target gene.
In this study, we attempted to standardize copy numbers for COI by introducing an independent replication step. Genomic DNA from each species was amplified using the primers UCO2F and 6R, and the mock mixture was created from normalizing the concentration of the resulting amplicons. Zachary Lane’s mock mixture study, conducted in 2017, did not include this step. His mock mixture consisted of unamplified genomic DNA from the species at equal concentration. We compared the results of Lane’s study with our own to determine whether the independent replication step indeed affected the observed amplification bias.

As stated earlier, sequences were analyzed as single-end reads in this study. Lane, however, used paired-end analysis. As stated earlier, paired-end analysis requires a certain amount of overlap between the forward and reverse sequences. Only primer pairs 1 (1F-1R) and 4 (3F-4R) generated sequences with enough overlap to be analyzed as paired-end reads. Thus, only pairs 1 and 4 were compared with Lane’s results. The two primer pairs showed substantial amplification bias in both studies, differing significantly from an even distribution. The sequence distribution for primer pair 4 did not differ significantly between the two studies, with Lane’s results showing a similar bias towards *Quadrula quadrula*. The distribution of sequences for primer pair 1, however, did differ significantly between the studies. In our study, primer pair 1 was biased against *Quadrula quadrula*. In Lane’s study, all 4 primer pairs were heavily biased towards *Quadrula quadrula*, including primer pair 1. The results of our study align more closely with the base pair mismatches between the primers and mock mixture species. There are no consistent mismatches between *Q. quadrula* and primer pairs 2, 3, and 4, but there are two mismatches between the species and primer pair 1. Therefore, we would expect primer pair 1 to amplify *Q. quadrula* less effectively than the other pairs, a hypothesis validated by our sequencing results but not Lane’s. This implies that the independent replication step in our study
may have led to a more accurate assessment of primer bias by controlling equal copy numbers for each template in the mock mixture.

It is important to note that our study is not completely comparable to Lane’s mock mixture experiment. One of the species differed between the two studies. Instead of *Lasmigona costata*, Lane used *Fusconaia flava*. The primers likely showed different levels of bias towards these species, affecting the amplification of the other five. However, the proportion of sequences attributed to *L. costata* in this study was nearly equal to the proportion for *F. flava* in Lane’s study.

2.5 Conclusion and Future Directions

Metabarcoding is a rapidly evolving technique allowing scientists to monitor biological communities using eDNA. There are some significant challenges associated with this method, however, including stochastic and amplification bias. This study introduces novel metabarcoding primers for the detection of Unionid mussel species. The stochastic and amplification bias associated with these primers was evaluated using an equimolar mock mixture of 6 common Unionids.

Variability across separate PCR runs was not statistically significant, suggesting that stochastic bias may be substantially reduced by performing multiple PCR replicates. There was statistically significant amplification bias for all 4 primer pairs. Specifically, primer pairs 2, 3, and 4 were highly effective at amplifying *Quadrula quadrula* but produced very few sequences for the two *Lasmigona* species. These patterns can be explained by the observed base pair mismatches between the primers and mock mixture species. Overall, primer pairs 1 and 2 were the least biased and will be used for future studies of Unionid eDNA. Additionally, these two
primer pairs minimally overlap and together sequence over 500 base pairs of COI, thereby maximizing the sequence coverage across this variable gene (Figure 6).

The consistently low representation of *Lasmigona* species suggests that the 4 primer pairs are biased towards the subfamily Ambleminae. *Lasmigona* belongs to the subfamily Anodontinae. Thus, future studies should introduce primer pairs designed specifically for Anodontinae and the other subfamilies within Unionidae. In addition, this study indicates that reference databases like GenBank currently lack the information needed to produce species-level identifications of Unionids with COI primers. Researchers need to contribute more sequences to these databases if metabarcoding studies are to maximize utility. Finally, it is important to understand the inherent challenges of eDNA metabarcoding, including amplification bias. Mock mixture experiments can help scientists interpret sequencing data more accurately and should be used more widely in eDNA studies.
CHAPTER 3: USING METABARCODING PRIMERS TO DETECT UNIONID eDNA

3.1 Methods

3.1.1 Site Selection and eDNA Collection

The 3 Rivers Quest (3RQ) monitoring program is an initiative to monitor water quality in 3 northeastern states, including Pennsylvania. Duquesne University is one of the participating organizations in the program. Three sites associated with the 3RQ program were chosen for monthly eDNA monitoring: Allegheny Lock & Dam 5 (40.688471, -79.663267), Allegheny-Parker (41.106499, -79.680766), and Kiski-Leechburg (40.635405, -79.612801) (Figure 19). Dr. Elizabeth Dakin collected a single 1-Liter water sample from each of these sites once a month in tandem with the usual sampling activities required by 3 Rivers Quest. 3RQ samples analyzed for this study were collected between March 2021 and August 2022.

Figure 19: Locations of monthly 3RQ eDNA collection sites. Map was created using Google Earth.
In addition to these monthly samples, sites from two watersheds were chosen for eDNA collection: 3 sites in Buffalo Creek and 1 site in Crooked Creek (Figure 20). The Buffalo Creek sampling was conducted by Meredith Bennett and Kathleen Wilson in September 2022. At each site, water samples were collected at 3 points along a transect perpendicular to the stream flow (left descending bank, center, and right descending bank). The Crooked Creek sampling was conducted by Katie Stupar in July 2020. Because Crooked Creek is a wider stream, Stupar collected samples at 5 points along the perpendicular transect.

All water samples for eDNA analysis were collected in sterile 1-L plastic bottles and placed on ice for transport to the laboratory. The samples were stored in a refrigerator at 4 °C until filtered. Field blanks were taken for the Buffalo Creek sampling trip and for half of the 3RQ...
sampling dates. Field blanks consisted of a sterile 1-L plastic bottle filled with deionized water from the laboratory prior to sampling. In the field, blanks were stored on ice alongside the eDNA samples.

3.1.2 Filtration of Water Samples

All water samples and field blanks were processed within 24 hours of collection. The eDNA was captured via vacuum filtration. Filtration took place in a tissue culture hood sterilized by UV radiation. The filtering apparatus consisted of a 1-L Sidearm flask, a glass filter support funnel, a glass funnel barrel, and a spring clamp. Prior to filtration, the filter support funnel, the funnel barrel, and a pair of forceps were sterilized by the following protocol: 30-minute soak in 25% bleach, tap water rinse, and 15-minute soak in deionized water. The forceps were used to place a Grade 1 qualitative filter paper (Whatman, cat. # 1001-045) on the filter support funnel and a 0.7 um glass fiber filter (Whatman, cat. # 1825-047) on top of the Grade 1 filter. The funnel barrel was then clamped to the filter support with the spring clamp, and the sidearm flask was connected to the vacuum pump with a rubber hose. After filtering the 1-L sample, the filtrate was discarded and the two filter papers were stored in zip-lock bags at -80 °C. A newly sterilized set of equipment was used for every sample.

3.1.3 eDNA Extraction from Filters

This protocol for DNA extraction from filters is based on an eDNA monitoring project conducted by the Midwest region of the U.S. Fish and Wildlife Service for the detection of invasive carp. Each filter was extracted as a separate sample and labeled to indicate the site, filter type, and transect (if applicable).
A 1.7-mL microcentrifuge tube was prepared for each sample, containing 370 uL of ASL or ATL buffer and 30 uL of proteinase K. The filters were retrieved and allowed to thaw before folding them to fit in the prepared tubes. The tubes with filters were briefly vortexed and allowed to incubate at 55 °C for at least 1 hour. During incubation, the tubes were vortexed periodically. A modified spin column was prepared for each sample by removing the plunger from a 5-mL syringe and placing the barrel in a sterile, labeled 15-mL conical tube. After incubation, the filters and buffer solution were transferred to the syringe barrels using a sterile pipette. The modified spin columns were spun in a Beckman GS-15R centrifuge at 1,000 rpm for 15-20 minutes. This step separated the buffer solution from the filters. The filters and syringe barrels were discarded. The filtrate in each conical tube was transferred to a clean, labeled 1.7-mL microcentrifuge tube.

To each of these tubes was added 400 uL of buffer AL and 400 uL of ethanol (100%). The tubes were briefly vortexed after each addition. Roughly half of the mixture in each tube was transferred to a DNeasy spin column nestled in a 2-mL collection tube. The spin columns were centrifuged at 6,000 x g for one minute. The filtrate was discarded, and the column was placed back in the collection tube. The remaining mixture was transferred to the spin column and centrifuged at the same settings. The filtrate and collection tubes were discarded.

The spin columns were transferred to new collection tubes, and 500 uL of buffer AW1 (wash buffer 1) was added to the center of each column. The columns were centrifuged at 6,000 x g for 1 minute, after which the filtrate was discarded. The columns were transferred to new collection tubes, and 500 uL of buffer AW2 (wash buffer 2) was added to the center of each column. The columns were centrifuged at 18,000 x g for 3 minutes, and the filtrate was discarded. The spin columns were transferred to clean, labeled 1.7-mL microcentrifuge tubes for
the final elution steps.

To the center of each spin column was added 100 μL of preheated (70 °C) buffer AE. The columns were incubated at room temperature for 1 minute, then centrifuged at 6,000 x g for one minute. Another 100 μL of the warm buffer AE was added to the columns, which incubated at room temperature for 1 minute. The samples were centrifuged at 6,000 x g for one minute. The spin columns were discarded, and the 200 μL of pooled eluate were stored at -20 °C.

3.1.4 Amplification of Extracted eDNA

According to the previous mock mixture experiment, the two primer pairs with the least amplification bias were 1F-1R and 2F-2R. Therefore, the eDNA samples and field blanks were amplified using these primers. Preliminary amplifications were conducted for a subset of samples to determine the optimal volume of extracted eDNA to add to PCR reactions. The reactions were tested with 1 μL, 3 μL, and 5 μL of eluate. Amplification was confirmed via gel electrophoresis using 2% agarose gel in SB buffer, stained with ethidium bromide, and visualized under UV light. According to the resulting gels, the 0.7 um filters were optimized at 3 μL of eluate, while the Grade 1 filters were optimized at 1 μL of eluate.

All eDNA extractions were amplified in 20 μL triplicate reactions using KAPA HiFi HotStart ReadyMix. Each reaction included 0.75 μL of the forward (1F or 2F) and reverse (1R or 2R) primers at 10 μM, PCR-grade water, and KAPA HiFi DNA polymerase. For the reactions of 0.7 um filters, 3 μL of eDNA eluate was added. For the reactions of Grade 1 filters, 1 μL of eluate was added. The thermocycler protocol (KAPAMUSSEL) consisted of the following steps: 30 second hot start at 98 °C followed by 46 cycles of denaturation (98 °C for 5 seconds), annealing (63 °C for 20 seconds), and extension (72 °C for 15 seconds). Cycling was followed
by a 5-minute final extension at 72 °C. For each PCR run, a positive and negative control were included. The positive control reaction contained 1 uL of tissue DNA from *Obliquaria reflexa* at 50 ng/uL. The negative control reaction consisted of PCR-grade water. After each PCR run, the positive and negative controls were visualized via gel electrophoresis to confirm amplification success and the absence of contamination.

The three replicates for each sample were pooled together, producing a total 60 uL of PCR product for each sample. At this point, we had separate samples for each extracted filter for both primer pairs, yielding 236 total samples. To conserve sequencing resources, we reduced this number by combining the two filter types for the Buffalo Creek and Crooked Creek samples and for the 3RQ samples amplified with primer set 1 (1F-1R). We kept the filter types separate for the 3RQ samples amplified with primer set 2 (2F-2R), because we were interested in whether the filters produce significantly different results. After combining the filter types, we had a total of 165 samples (Figure 21).
3.1.5 Next-generation Sequencing

The samples were cleaned using Mag-Bind Total Pure NGS magnetic beads. Cleaning followed the protocol for Left Side Size Selection in the SPRIselect User Guide. Based on the length of the desired amplicon (around 300 bp), the products were cleaned using a 1.2 ratio of beads to sample. The cleaned samples were eluted to 20 uL in PCR-grade water. Each sample was then labeled with a unique set of indexing primers, which attach to the adaptor sequences on the internal primers (Illumina, cat. #FC-131-1002). The indexed samples were cleaned using a 1.2 bead-to-sample ratio and eluted to 20 uL. The samples were then measured on the Qubit and pooled together based on similar concentrations, resulting in 7 tubes. These samples were run on

\textbf{Figure 21:} Layout of eDNA sample sequencing for each primer pair.
the Tape Station to confirm the amplicon size was correct. Concentrations were measured on the Qubit again, and all samples were diluted to the same concentration and pooled. The pooled sample was measured on the Qubit and Tape Station to ensure quality. Finally, the pooled sample was loaded onto the Illumina MiSeq in Dr. Jan Janecka’s lab and run according to the protocol associated with Illumina’s MiSeq Reagent Nano Kit v2 (500 cycles). Due to an error in setup, however, the sequencer only ran for 300 cycles.

The software, QIIME 2 was used to analyze the sequence reads produced by the MiSeq run. The primers were removed from the sequences using the “cutadapt” package. The sequences were processed as single-end reads, and the “DADA2” package was used to remove errors from the data. Using DADA2, the sequences were denoised, removing replicates and chimeras. The sequences were then clustered into operational taxonomic units (OTUs) based on 97% similarity. Each of the unique OTUs were identified to species level by entering the sequence into the Basic Local Alignment Search Tool (BLAST). The species for each OTU was identified as the top hit from BLAST, and the e-value (an indication of the probability of finding the sequence by random chance) was recorded. In the case of 2 equally likely results (e-value and percent identity were the same), both species were recorded as possibilities.

3.1.6 Data Analysis

Molecular Evolutionary Genetics Analysis (MEGA) software was used to align the sequences generated by the MiSeq. The evolutionary distances within and between certain Unionid groups were computed using MEGA. The p-distance method was used to calculate the proportion of nucleotides that were different between the specified species.
3.2 Results

3.2.1 Overview of eDNA Results

Both primer pairs (1F-1R and 2F-2R) amplified Unionid DNA from many of the eDNA samples. Over 99% of the sequences generated by each individual primer were identified to a Unionid species. After trimming the primers, the sequences identified as Unionids were all between 125 – 128 base pairs for primer pair 1 (1F-1R). The Unionid sequences for primer pair 2 (2F-2R) were nearly all 128 base pairs long. As in the mock mixture experiment, BLAST could not distinguish between *Potamilus alatus* and *Potamilus purpuratus*, reporting equal E values and percent similarities for the two species. For data analysis, those sequences were attributed to *P. alatus*, because it is more likely to occur in the sampled areas.

Tables 13 and 14 show the most common sequences generated by the two primer pairs for a subset of samples. The E values for these sequences were all below $1 \times e^{-54}$. Overall, primer pair 1F-1R produced more sequences than 2F-2R. The most common sequences were identified to Unionid species used in the previous mock mixture experiment. In addition, the field blanks were heavily contaminated with DNA from mock mixture species. The field blanks not displayed in tables 13 and 14 were similarly contaminated. Due to the abundant contamination, we are unable to draw reliable conclusions about the sequences identified as mock mixture species. Recall that the two filter types (0.7 um and Grade 1) were sequenced separately for primer pair 2F-2R. Table 14 shows the separate sequence counts for each filter type. In tables 13 and 14, the first four columns represent the Buffalo Creek samples plus the associated field blank. The next four columns are the 3RQ samples for August 2022, along with the field blank from the same date. Finally, “PCR Neg-3RQ” refers to the negative PCR control associated with the 3RQ samples from August 2022.
Table 13: Sequence counts for the most common Unionid species identified by primer pair 1F-1R. Columns 1-3 show the center transect samples for Buffalo Creek, followed by the associated field blank. Columns 5-7 are 3 Rivers Quest samples from August 2022, followed by the associated field blank. Column 9 is the PCR negative control associated with the 3RQ samples.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Lasmigona complanata</em></td>
<td>0</td>
<td>0</td>
<td>612</td>
<td>2539</td>
<td>986</td>
<td>2202</td>
<td>1409</td>
<td>537</td>
<td>0</td>
</tr>
<tr>
<td><em>Obliquaria reflexa</em></td>
<td>311</td>
<td>32</td>
<td>3</td>
<td>844</td>
<td>703</td>
<td>2570</td>
<td>1887</td>
<td>338</td>
<td>96</td>
</tr>
<tr>
<td><em>Ligumia recta</em></td>
<td>73</td>
<td>37</td>
<td>269</td>
<td>1269</td>
<td>327</td>
<td>807</td>
<td>376</td>
<td>578</td>
<td>0</td>
</tr>
<tr>
<td><em>Potamilus alatus</em></td>
<td>72</td>
<td>22</td>
<td>11</td>
<td>1615</td>
<td>84</td>
<td>876</td>
<td>469</td>
<td>148</td>
<td>0</td>
</tr>
</tbody>
</table>
**Table 14:** Sequence counts for the most common Unionid species identified by primer pair 2F-2R. Columns 1-3 show the center transect samples for Buffalo Creek, followed by the associated field blank. Columns 5-7 are 3 Rivers Quest samples from August 2022, followed by the associated field blank. Column 9 is the PCR negative control associated with the 3RQ samples. Columns 5-8 are separated to show the sequence counts for the two filter types: 0.7 um glass fiber and Grade 1 qualitative.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quadrula quadrula</strong></td>
<td>91</td>
<td>48</td>
<td>62</td>
<td>164</td>
<td>0</td>
<td>194</td>
<td>29</td>
<td>647</td>
<td>13</td>
</tr>
<tr>
<td><strong>Obliquaria reflexa</strong></td>
<td>18</td>
<td>4</td>
<td>39</td>
<td>63</td>
<td>5</td>
<td>69</td>
<td>0</td>
<td>238</td>
<td>9</td>
</tr>
<tr>
<td><strong>Potamilus alatus</strong></td>
<td>0</td>
<td>21</td>
<td>5</td>
<td>69</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>117</td>
<td>0</td>
</tr>
<tr>
<td><strong>Ligumia recta</strong></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>46</td>
<td>0</td>
</tr>
<tr>
<td><strong>Lasmigona complanata</strong></td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>0</td>
</tr>
</tbody>
</table>
3.2.2 Unique Sequences in eDNA Samples

The overwhelming majority of sequences in the eDNA samples were identified to Unionid species used in the mock mixture experiment. Table 15 shows the 3 instances in which a unique Unionid species was identified. Two of the species were identified by reverse primer 2 (2R), and one was identified by forward primer 1 (1F). All three were identified in 3 Rivers Quest samples. For one of the identifications, BLAST could not differentiate between two species of the same genera: *Pleurobema stabile* and *Pleurobema georgianum*.

Of these Unionid species, only one is native to the areas where sampling occurred: *Lampsilis cardium*. Within the genus *Lampsilis*, there are several species native to western Pennsylvania. Native *Lampsilis* species were aligned to each other in MEGA, and a measure of evolutionary distance within the group indicated a mean uncorrected nucleotide distance of 8.0%. *Villosa delumbis, Pleurobema stabile,* and *Pleurobema georgianum* do not occur naturally in western Pennsylvania and are native to the southeastern United States. Native species within the genera *Villosa* and *Pleurobema* were aligned in MEGA, and evolutionary distance was measured between the native species and the identified species. These computations revealed a mean uncorrected distance of 9.19% between *Villosa delumbis* and the *Villosa* species native to western PA. A mean uncorrected distance of 7.14% was found between *Pleurobema stabile/georgianum* and the *Pleurobema* species native to this region.
Table 15: Sequence counts for the 3 unique (non-mock mixture) Unionid species identified in this study. “Percent Identity” indicates the degree to which the identified sequence matches the top hit generated by BLAST. The “E Value” indicates the probability of identifying the given species by random chance.

<table>
<thead>
<tr>
<th>Species</th>
<th>E Value</th>
<th>Percent Identity (%)</th>
<th>Kiski-Leechburg Combined filters (12/15/2021) Primer 1F</th>
<th>Allegheny Lock&amp;Dam 5 Grade 1 filter (8/18/2022) Primer 2R</th>
<th>Kiski-Leechburg Grade 1 filter (12/15/2021) Primer 2R</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Villosa delumbis</em></td>
<td>8.00E-44</td>
<td>93.65</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>Lampsilis cardium</em></td>
<td>4.00E-52</td>
<td>97.62</td>
<td>0</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td><em>Pleurobema stabile/georgianum</em></td>
<td>3.00E-53</td>
<td>97.66</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
3.3 Discussion

Primer pairs 1 and 2 from the mock mixture experiment were used for the eDNA samples. Similar to the mock mixture results, the sequences identified from the eDNA samples were of high quality. Nearly 100% of the sequences were identified to Unionid species, with appropriately low e-values. In addition, the sequences generally matched our predicted amplicon length.

The identifications generated by BLAST showed the same patterns of uncertainty as in the mock mixture experiment. Unlike the mock mixture, however, we did not have prior knowledge of the species present in our samples. Therefore, uncertainty in GenBank greatly inhibits the ability to make reliable identifications. This further emphasizes the need to expand public databases like GenBank if eDNA studies are to be successful.

Nearly all of the sequences from the eDNA samples were identified to species used in the previous mock mixture experiment. In addition, all of the field blanks contained abundant sequences from mock mixture species. This suggests rampant contamination of the eDNA samples and prevents us from drawing any reliable conclusions about the presence of these species at the sites where the samples were taken. Despite this setback, total sequence counts suggest that primer pair 1 was more effective than primer pair 2 at amplifying Unionid DNA. In addition, the Grade 1 qualitative filters generally produced more sequences than the 0.7 um glass fiber filters, suggesting they may be more effective at DNA capture.

3.3.1 Unique Unionid Species
There were only three instances in which a unique (non-mock mixture) Unionid species was detected in the eDNA samples. All three were detected in 3 Rivers Quest samples. This makes sense, considering large rivers will likely contain greater density and diversity of Unionid species than smaller streams like Buffalo and Crooked Creek. Notably, two of the unique species were amplified by reverse primer 2 (2R), and one was amplified by forward primer 1 (1F). This illustrates the advantages of using more than one primer pair. Although primer pair 1 generated more sequences overall, it was primer pair 2 that successfully detected two of the three unique Unionid species.

Interestingly, only one of the three unique Unionid species is native to western Pennsylvania: *Lampsilis cardium*. The other two identifications (*Villosa delumbis* and *Pleurobema stabile/georgianum*) were for species native to the southeastern United States. There are, however, several species native to western Pennsylvania within the genera *Villosa* and *Pleurobema*. We suspect these misidentifications are due to the incompleteness of GenBank and the relatively small evolutionary distance with the genera. These obstacles also cast doubt on the *Lampsilis cardium* identification. Again, the continued expansion of reference databases is needed for next-generation sequencing to produce reliable species identifications.

3.3.2 Contamination

Widespread contamination in the eDNA samples prevented reliable interpretation of the sequencing results. Almost 100% of the sequences were attributed to mock mixture species, suggesting the samples were contaminated with DNA used during the mock mixture experiment. In addition, all field blanks were contaminated with mock mixture DNA. Field blanks consisted of a bottle of DI water stored alongside eDNA samples and subjected to the same filtration and
extraction processes. When PCR products were visualized using gel electrophoresis, the field blanks consistently produced visible bands. In fact, the bands associated with field blanks were often brighter than those for actual eDNA samples. The PCR negatives, however, did not show visible bands and generally contained few sequences according to the MiSeq results. This suggests that contamination occurred before amplification, most likely during extraction. This may also explain why the field blanks produced brighter bands than actual samples. Even trace amounts of DNA in the DI water of field blanks would be amplified more successfully than DNA in the actual samples, which contain PCR inhibitors (humic and fumic acids) and non-target DNA. Common protocol in eDNA studies is to discount any sequences identified in samples if their associated field blank is contaminated. If applied to this study, that practice would result in no usable data. Therefore, eliminating contamination must be a top priority of future studies employing eDNA techniques.

3.4 Conclusion and Future Directions

This experiment applied two previously validated primer pairs (1F-1R and 2F-2R) to eDNA samples taken from two major rivers (the Allegheny and the Kiskiminetas) and two local watersheds (Buffalo Creek and Crooked Creek). The sequencing results further validate the primers’ ability to amplify Unionid DNA. However, rampant contamination from the previous mock mixture experiment prevents us from drawing any reliable conclusions about which Unionid species are present in the samples. Only three unique taxa (*Lampsilis, Villosa, and Pleurobema*) were detected in the samples, and ambiguity in GenBank made it difficult to identify these sequences to species level.
Contamination is a seriously overlooked challenge facing eDNA studies. In this study, contamination was broadly pinpointed to the extraction step. Remnants of tissue DNA used in the mock mixture experiment was likely present in lab materials (pipets, reagents, microcentrifuge etc.) and possibly the atmosphere. Despite what we considered diligent cleaning protocols, this DNA was able to contaminate field samples at some point during extraction. If eDNA studies are to be reliable, reducing contamination must be a major endeavor. Researchers should not attempt these projects without first preparing protocols and equipment to address contamination. We suggest establishing checkpoints at which to test for contamination. If contamination is detected at a checkpoint, researchers can address the problem without wasting time and resources. Finally, the prevalence of contamination in eDNA studies will likely make them impracticable for certain research organizations. It may be necessary for these groups to send eDNA samples to labs with more advanced sterilization capabilities.
CHAPTER 4: CONCLUSIONS

In this thesis, eDNA-based methods were examined in an in-depth literature review. The two main approaches, single-species detection and metabarcoding, were compared. We argue that metabarcoding is generally superior to single-species detection, because it provides valuable taxonomic information for many species simultaneously. That being said, eDNA methods have several limitations, including the inability to provide reliable quantitative data, like individual counts or biomass. Despite challenges and the need for standardization, eDNA-based methods are powerful tools that can improve conservation by supplementing traditional survey methods.

In a mock mixture experiment, we validated new metabarcoding primers for the detection of Unionid species and explored the effects of amplification bias and stochasticity. We found that conducting multiple PCR replicates reduces bias by accounting for stochasticity. All of the primer pairs displayed amplification bias, explained by specific base pair mismatches between the primers and the various species included in the mock mixture. We advocate for greater use of mock mixture experiments, because they characterize primer bias and promote more reliable interpretation of sequencing data.

The final section of this thesis described the application of metabarcoding primers to eDNA samples taken from several rivers and streams in western Pennsylvania. This experiment revealed extensive contamination from the previous mock mixture experiment. The contamination likely occurred during the extraction step, inhibiting the amplification of any actual eDNA signals. Contamination is a major challenge facing eDNA studies and has often been overlooked. Those planning eDNA surveys should have clear strategies to address contamination. Furthermore, successful implementation of eDNA methods may not be feasible
for certain groups, such as those without sufficient sterilization capabilities. In those cases, research groups may consider sending eDNA samples to an external lab for analysis.

Unionid mussels are highly imperiled and difficult to identify morphologically, making them prime candidates for eDNA-based methods. Despite their ecological importance, Unionids remain understudied. This study validated metabarcoding primers for the detection of Unionid species, which have the capacity to improve ecological study and conservation. However, the application of these methods is inhibited by the lack of reference sequences for many species. Databases like GenBank are severely lacking when it comes to Unionids, making it difficult to identify sequences to species level. Metabarcoding studies of Unionids and other taxonomic groups will benefit from greater efforts to sequence the DNA of neglected species.

Environmental DNA has radically transformed ecology, especially in the study of aquatic organisms. As methods continue to improve, eDNA will undoubtedly yield many breakthroughs in the study of the natural world.
1. Lacoursière-Roussel, A.; Deiner, K., Environmental DNA is not the tool by itself. Journal of Fish Biology 2019, n/a (n/a).


19. Andújar, C.; Arribas, P.; Yu, D. W.; Vogler, A. P.; Emerson, B. C., Why the COI barcode should be the community DNA metabarcode for the metazoa. Molecular Ecology 2018, 27 (20), 3968-3975.


36. del Carmen Gomez Cabrera, M.; Young, J. M.; Roff, G.; Staples, T.; Ortiz, J. C.; Pandolfi, J. M.; Cooper, A., Broadening the taxonomic scope of coral reef palaeoecological studies using ancient DNA. Molecular Ecology 2019, 28 (10), 2636-2652.


55. Clare, E. L.; Economou, C. K.; Faulkes, C. G.; Gilbert, J. D.; Bennett, F.; Drinkwater, R.; Littlefair, J. E., eDNAir: proof of concept that animal DNA can be collected from air sampling. PeerJ 2021, 9, e11030.


82. Bylemans, J.; Gleeson, D. M.; Hardy, C. M.; Furlan, E., Toward an ecoregion scale
evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of
Genner, M. J.; Mariani, S., Non-specific amplification compromises environmental DNA
84. Clarke, L. J.; Soubrier, J.; Weyrich, L. S.; Cooper, A., Environmental metabarcodes for
insects: In silico PCR reveals potential for taxonomic bias. Molecular Ecology Resources 2014,
14 (6), 1160-1170.
85. Deagle, B. E.; Jarman, S. N.; Coissac, E.; Pompanon, F.; Taberlet, P., DNA
metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. Biology
Letters 2014, 10 (9).
86. Krehenwinkel, H.; Wolf, M.; Lim, J. Y.; Rominger, A. J.; Simison, W. B.; Gillespie,
R. G., Estimating and mitigating amplification bias in qualitative and quantitative arthropod
metabarcoding. Scientific Reports 2017, 7 (1), 17668.
Degenerate PCR Primers from Alignments or Consensus Sequences. Evolutionary
Bioinformatics 2012, 8, EBO.S8870.
88. Kebschull, J. M.; Zador, A. M., Sources of PCR-induced distortions in high-throughput
sequencing data sets. Nucleic Acids Res 2015, 43 (21), e143-e143.


93. Capo, E.; Spong, G.; Königsson, H.; Byström, P., Effects of filtration methods and water volume on the quantification of brown trout (Salmo trutta) and Arctic char (Salvelinus alpinus) eDNA concentrations via droplet digital PCR. Environmental DNA 2020, 2 (2), 152-160.


96. Jo, T. S., Correlation between the number of eDNA particles and species abundance is strengthened by warm temperature: simulation and meta-analysis. Hydrobiologia 2022.
100. Capo, E.; Spong, G.; Norman, S.; Königsson, H.; Bartels, P.; Byström, P., Droplet digital PCR assays for the quantification of brown trout (Salmo trutta) and Arctic char (Salvelinus alpinus) from environmental DNA collected in the water of mountain lakes. PLoS ONE 2019, 14 (12).


112. Jo, T.; Minamoto, T., Complex interactions between environmental DNA (eDNA) state and water chemistries on eDNA persistence suggested by meta-analyses. Molecular Ecology Resources 2021, 21 (5), 1490-1503.


121. Gleason, J. E.; Elbrecht, V.; Braukmann, T. W. A.; Hanner, R. H.; Cottenie, K., Assessment of stream macroinvertebrate communities with eDNA is not congruent with tissue-based metabarcoding. Molecular Ecology 2021, 30 (13), 3239-3251.
122. Dickie, I. A.; Boyer, S.; Buckley, H. L.; Duncan, R. P.; Gardner, P. P.; Hogg, I. D.;
Holdaway, R. J.; Lear, G.; Makiola, A.; Morales, S. E.; Powell, J. R.; Weaver, L., Towards
robust and repeatable sampling methods in eDNA-based studies. Molecular Ecology Resources
123. Evans, N. T.; Shirey, P. D.; Wieringa, J. G.; Mahon, A. R.; Lamberti, G. A.,
Comparative Cost and Effort of Fish Distribution Detection via Environmental DNA Analysis
Scientific and budgetary trade-offs between morphological and molecular methods for deep-sea
biodiversity assessment. Integrated Environmental Assessment and Management 2022, 18 (3),
655-663.
125. Aquatic Systems, I. Invoice Night Electrofishing Field Work; Aquatic Systems, Inc.:
2010.
126. Machtinger, E. T., Native Freshwater Mussels. In Natural Resources Conservation
Management Leaflet, 2007; Vol. 46.
mussels/.
Allen, D.; Bogan, A. E.; Cordeiro, J.; Cummings, K. S.; Cuttelod, A.; Darrigran, G.; Darwall,
Aldridge, D. C.; Aravind, N. A.; Budha, P. B.; Clavijo, C.; Van Tu, D.; Gargominy, O.;


158. Coghlan, S. A.; Currier, C. A.; Freeland, J.; Morris, T. J.; Wilson, C. C., Community eDNA metabarcoding as a detection tool for documenting freshwater mussel (Unionidae) species assemblages. Environmental DNA 2021.

159. Klymus, K. E.; Richter, C. A.; Thompson, N.; Hinck, J. E.; Jones, J. W., Metabarcoding assays for the detection of freshwater mussels (Unionida) with environmental DNA. Environmental DNA 2021, 3 (1), 231-247.


assessment: methods and results for the Western Palearctic (European sub-region).
Hydrobiologia 2021, 848 (12), 2931-2950.


171. Institute, W. V. W. R. What is 3RQ.


173. FISHBIO Electrofishing Rentals. https://fishbio.com/electrofishing-rentals/#:--text=Backpack%20Electrofisher%20Rental%3A%20FISHBIO%20also%20has%20Smith-Root%20LR-24,to%20rent%20for%20241%2C500%20per%20week%20plus%20shipping.


175. Cabela's Frabill Minnow Seines. https://www.cabelas.com/shop/en/frabill-minnow-seines?ds_e=MICROSOFT&ds_c=Cabelas%7CShopping%7CSmart%7CCoolers%7CGeneral%
176. Amazon.com Amazon. [https://www.amazon.com/](https://www.amazon.com/).

177. Commission, P. F. a. B. Buy a Fishing License. [https://www.fishandboat.com/Fish/FishingLicense/Pages/BuyAFishingLicense.aspx](https://www.fishandboat.com/Fish/FishingLicense/Pages/BuyAFishingLicense.aspx).


180. College, A.-B. T. C. Laboratory Use & Equipment Rental. [https://abtech.edu/programs/workforce/bionetwork/laboratory-use-equipment-rental#~:text=Laboratory%20Rental%20and%20Basic%20Supply%20Fees.%20%2420%20per%20the%20use%20of%20specific%20equipment%20as%20listed%20below](https://abtech.edu/programs/workforce/bionetwork/laboratory-use-equipment-rental#~:text=Laboratory%20Rental%20and%20Basic%20Supply%20Fees.%20%2420%20per%20the%20use%20of%20specific%20equipment%20as%20listed%20below).


Appendix

References for Cost Analysis of eDNA and Traditional Methods

*Backpack Electrofishing*

Electrofisher (batteries, charger, etc.) 172, 173

Seine and dip nets 174, 175

Bubblers, tape measure, collection/processing buckets, processing nets 176

Fishing permits 177

*Environmental DNA (eDNA) Methods*

Collection bottles, filtration equipment, lab consumables 178

Extraction kit 179

Lab usage/rental 180

qPCR assays 181

qPCR services 182, 183

Next-generation sequencing services 183, 184