BUREAU OF RECLAMATION

Investigation of DNA metabarcoding for macroinvertebrate surveys and invasive species detection

Science and Technology Program
Research and Development Office
Final Report No. ST-2020-1831-01

U.S. Department of the Interior

September 24, 2020

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
24-09-2020	Research	2018 – 2020

4. TITLE AND SUBTITLE
Investigation of DNA metabarcoding for macroinvertebrate surveys and invasive species detection

5a. CONTRACT NUMBER
FA881 / RR.4888.FARD1803801

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
1541 (S&T)

6. AUTHOR(S)
Yale Passamaneck, Ph.D. (Bureau of Reclamation)

5d. PROJECT NUMBER
Final Report ST-2020-1831-01

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Technical Service Center
Bureau of Reclamation
U.S. Department of the Interior
Denver Federal Center PO Box 25007, Denver, CO 80225-0007

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Science and Technology Program
Research and Development Office
Bureau of Reclamation
U.S. Department of the Interior
Denver Federal Center
PO Box 25007, Denver, CO 80225-0007

10. SPONSOR/MONITOR'S ACRONYM(S)
Reclamation

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Final Report ST-2020-1831-01

12. DISTRIBUTION/AVAILABILITY STATEMENT
Final Report may be downloaded from https://www.usbr.gov/research/projects/index.html

13. SUPPLEMENTARY NOTES

14. ABSTRACT
DNA metabarcoding is a rapidly developing technique for rapidly surveying entire communities of organisms from individual environmental samples. Metabarcoding was tested for macroinvertebrate surveys and for detection of invasive species, including quagga mussel. Results were compared to traditional methods. Metabarcoding has significant potential, but this study demonstrated that carful planning and assay validation will be key to the successful adoption of this technique.

15. SUBJECT TERMS
Metabarcoding, DNA, macroinvertebrates

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*
U	U	U			

Mission Statements

The Department of the Interior (DOI) conserves and manages the Nation's natural resources and cultural heritage for the benefit and enjoyment of the American people, provides scientific and other information about natural resources and natural hazards to address societal challenges and create opportunities for the American people, and honors the Nation's trust responsibilities or special commitments to American Indians, Alaska Natives, and affiliated island communities to help them prosper.

The mission of the Bureau of Reclamation is to manage, develop, and protect water and related resources in an environmentally and economically sound manner in the interest of the American public.

Disclaimer

Information in this report may not be used for advertising or promotional purposes. The data and findings should not be construed as an endorsement of any product or firm by the Bureau of Reclamation, Department of Interior, or Federal Government. The products evaluated in the report were evaluated for purposes specific to the Bureau of Reclamation mission. Reclamation gives no warranties or guarantees, expressed or implied, for the products evaluated in this report, including merchantability or fitness for a particular purpose.

Acknowledgements

The Science and Technology Program, Bureau of Reclamation, sponsored this research. Molly Maier did much of the preliminary work for this project, performing DNA extractions, testing OCR strategies, and organizing reference lists of taxa. Sherri Pucherelli and Jacque Keele participated in development of the project, and collected field samples.

Investigation of DNA metabarcoding for macroinvertebrate surveys and invasive species detection

Final Report No. ST-2020-1831-01

prepared by

Technical Service Center
Yale Passamaneck, Ph.D., Biologist

Peer Review

Bureau of Reclamation
Research and Development Office
Science and Technology Program

Final Report No. ST-2020-1831-01

Investigation of DNA metabarcoding for macroinvertebrate surveys and invasive species detection

YALE PASSAMANECK

Digitally signed by YALE PASSAMANECK
Date: 2020.09.29 10:06:03 -06'00'

Prepared by: Yale J. Passamaneck, Ph.D.
Biologist, Ecological Research Laboratory, Hydraulic Investigations and Laboratory Services, 86-68560, Technical Service Center, Bureau of Reclamation

AARON MURPHY

Digitally signed by AARON MURPHY
Date: 2020.09.29 10:09:52 -06'00'

Reviewed by: Aaron Murphy
Biologist, Ecological Research Laboratory, Hydraulic Investigations and Laboratory Services, 86-68560, Technical Service Center, Bureau of Reclamation

Acronyms and Abbreviations

ASV	amplicon sequence variants
BLAST	Basic Local Alignment Tool
BOLD	Barcode of Life Data System
COI	*cytochrome c oxidase subunit I* gene
DNA	deoxyribonucleic acid
eDNA	environmental DNA
NCBI	National Center for Biotechnology Information
PCR	polymerase chain reaction
PE	paired-end
qPCR	quantitative PCR
RDP	Ribosomal Database Project
Reclamation	Bureau of Reclamation

Measurements

bp	base pairs
°C	degrees Celsius
Mbases	megabases
ng	nanogram
ng/µl	nanograms per microliter
µl	microliter
µM	micromolar

Contents

Executive Summary

Organismal surveys provide a critical tool for assessing the environmental impacts of Reclamation activities, as well as the efficacy of habitat restoration projects. For aquatic environments surveys are often focused on benthic macroinvertebrate communities, which are generally diverse and abundant, and which have well documented responses to environmental disturbances. Collection methods for macroinvertebrate surveys use standardized methods, allowing for comparisons across time points and between locations. However, sample processing and analysis can be time consuming, and taxonomic identification of samples can require specialized knowledge.

An alternative approach to these traditional methods for environmental community surveys has emerged with the development of DNA metabarcoding. DNA metabarcoding is an approach to surveying the whole community of organisms from an individual environment by using modern DNA sequencing technologies to capture thousands of sequences in parallel. Resultant sequences, and the taxa they are derived from, can then be identified through comparison to validated sequences in reference databases. DNA barcoding has the potential to be faster and less expensive than traditional surveys, and to cause less impacts to the sampled environment.

In this study metabarcoding was evaluated with samples collected near Folsom Dam, CA. Samples were also collected from Canyon Lake, AZ to test for applicability of metabarcoding to early detection of invasive species, including quagga mussels. Sequence data were obtained and analyzed, and sequences were identified as have come from a range of macroinvertebrate taxa.

DNA metabarcoding has significant potential for adoption for future projects. It is expected that as the technique gains more widespread usage, the challenges and limitations identified in this study will be overcome through standardization of practices.

Introduction

Aquatic macroinvertebrate surveys

Surveys of aquatic macroinvertebrate populations are widely used as an indicator of freshwater ecosystem health. Macroinvertebrates, in particular insects, are broadly distributed, diverse, and generally abundant in streams and other freshwater ecosystems. Because of these attributes, as well as the fact that aquatic macroinvertebrate populations may be affected by a variety of physical, chemical, and biological factors, these species are widely used as bioindicators. Temporal surveys of populations provide important indexes of the impact from environmental disturbances and the potential success of habitat restoration projects (Barbour, 1999). Macroinvertebrate surveys have been used to assess ecological impacts for a variety of Bureau of Reclamation projects (Carlisle et al., 2014; Nelson, 2005, 2009, 2011; Nelson & Wydowski, 2008).

Methods for sampling and sample processing of aquatic macroinvertebrates have been standardized by federal and state agencies responsible for environmental monitoring (Moulton et al., 2002; USEPA, 2013), as well as by international entities (Aqem Consortium, 2002; Canadian Aquatic Biomonitoring Network et al., 2012; Stark et al., 2001). These standardized protocols allow for the assessment of temporal changes to biodiversity and abundance at individual sites, as well as for comparisons of data from spatially distinct sampling locations and disjunct waterbodies.

Despite these attributes, there are disadvantages inherent to these traditional surveys for aquatic macroinvertebrates. Surveys generally target benthic communities, using kick net sampling to collect material from the substrate. Such sampling is necessarily disruptive, which can be an undesirable consequence of monitoring, particularly in habitats that may be occupied by threatened or endangered species. In addition, processing of samples can be time consuming and identification of organisms in the sample requires specialized taxonomic knowledge.

DNA metabarcoding

In the last decade there has been significant interest in using modern DNA sequencing technologies to develop an alternative method for conducting macroinvertebrate surveys. This approach, termed "DNA metabarcoding", relies on amplifying a conserved gene from an environmental sample, and then performing parallelized DNA sequencing and sequence analysis to identify organisms.

DNA barcoding

DNA barcoding is an approach that uses a sequence from a small fragment of DNA as a diagnostic identifier for the organism from which it was derived. For DNA barcoding the goal is to amplify a specific fragment of DNA from an unknown species using the polymerase chain reaction (PCR). The PCR product is then sequenced, and the resultant DNA sequence is compared to a database containing a large number of reference sequences. Comparison to the database should ideally result

in a perfect match, or identification of sequences of sufficient similarity such that taxonomic affiliation may be assigned, even if a specific species is not identified.

Target genes and universal primers

DNA barcoding is facilitated by the fact that many of the same genes are conserved in the genomes of widely diverse organism. For some genes, regions of sequence have been identified where there is sufficient conservation to design DNA primers that will amplify a fragment of the same gene from a variety of species. Depending upon the level of conservation of the primer recognition sites, and the design of the primers themselves, these so-called "universal" primers may have broad specificity. This allows amplification of the gene fragment of interest from an evolutionarily diverse range of species using the same primer set.

Cytochrome c oxidase I (COI)

For DNA barcoding to be useful, the targeted region must not only have sufficient conservation so that primers can be designed to successfully amplify from the organisms of interest, but the region of sequence between the primer recognition sites must also have sufficient variability such that each species possesses a unique sequence. The appropriate choice of target is dependent upon the taxonomic group of interest. In the case of metazoans (i.e. animals), and non-chordate metazoans (i.e. invertebrates) in particular, the *cytochrome c oxidase subunit I* gene is generally the target of choice (Andújar et al., 2018). *Cytochrome c oxidase subunit I* (abbreviated *COI* in this report, and also as *coxI* elsewhere) is a mitochondrial protein coding gene. *COI* has several attributes that contribute to its utility or barcoding. First, mitochondria, and their associated genomes, are numerous in most cells. Therefore, *COI* and other mitochondrial gene sequences are generally more abundant in tissue samples than single-copy genes in the nuclear genome, aiding amplification from limited tissue samples or environmental samples. Second, *COI* shows a high degree of sequence conservation due to the fact that it is required for oxidative metabolism. This functional constraint on sequence evolution allows for effective design of "universal" primers. However, despite this constraint, codon degeneracy allows for sufficient sequence evolution such that sequences of *COI* from invertebrates are generally species-specific. Finally, there is the fact that primers designed by Folmer et al. to amplify *COI* from a diverse group of taxa were among the first "universal" primers designed for invertebrate animals (Folmer et al., 1994). This led to the deposition of a large number of *COI* sequences in public databases in the decade between their publication and the first formal reference to sequence-based "barcoding", facilitating the adoption of *COI* as the marker of choice (Hebert et al., 1991).

DNA sequence databases

For taxonomic identification through DNA barcoding to be successful, a large and reliable database of reference sequences is required, against which sequences can be compared. The National Center for Biotechnology Information's (NCBI) GenBank database is one of the world's largest repositories of DNA sequence data, and as of 2018 held over 2.5 million *COI* sequences (Porter & Hajibabaei, 2018a). Although concerns have been raised about the accuracy of taxonomic assignments for DNA sequences housed in NCBI, a recent study has found that the proportion of errors is quite low for metazoan barcoding markers (Leray et al., 2019). In addition to GenBank, numerous independent databases have developed to focus on specific taxa and marker genes. For *COI*, the Barcoding of

Life Data System (BOLD) database (www.boldsystems.org) is the most widely used (Ratnasingham & Hebert, 2007). The Bold database currently houses over 2 million publicly available *COI* sequences from over 136,000 species. Both GenBank and BOLD provide tools for querying sequences of interest through their online portals to identify matching or similar sequences.

DNA metabarcoding

While DNA barcoding originally referred to analysis of DNA sequences derived directly from organismal tissue samples, a complementary approach termed "metabarcoding" has been developed to investigate environmental samples. DNA metabarcoding has emerged as a synthesis of DNA barcoding and a second approach, environmental DNA (eDNA). eDNA refers to extra-organismal DNA released from living or dead individuals and present in the environment. Sampling and analysis of environmental samples for eDNA can be used as a tool when looking for the presence of species of interest, particularly where traditional sampling methods may be of limited efficacy, cost and/or labor intensive, or undesirable due to negative impacts.

DNA metabarcoding involves 5 main steps:

1. Environmental sample collection
2. eDNA extraction and isolation
3. DNA target amplification
4. DNA sequencing
5. DNA sequence analysis

Like species-specific eDNA assays, DNA metabarcoding relies on performing PCR amplification of the target gene from a whole environmental sample. However, DNA metabarcoding uses universal primers to amplify all the DNA fragments in the sample recognized by the primers, ideally representing the range of species present in the sampled environment. DNA metabarcoding has been facilitated by the increasing availability, and decreasing cost, of next-generation DNA sequencing, which allows for sequencing of thousands or millions of individual DNA fragments in parallel.

Following sequencing, metabarcoding data requires much more intensive sequence analysis methods than does conventional barcoding, where individual DNA sequences are generated. Details of sampling, laboratory methods, and sequence analysis as pursued in this study are presented below.

In this report the term "eDNA" will be used to refer to genetic material isolated directly from an environmental sample, while "DNA" will be used to refer to the product of any downstream application, such as polymerase chain reaction (PCR) amplification or sequencing.

Methods

Sample collection

Two distinct types of samples were collected for DNA analysis, bulk-tissue samples and eDNA samples. Bulk-tissue refers to samples collected from the benthos and sediment, containing the actual organism (insects and other taxa) targeted for identification, along with other organic and inorganic material. The entire sample is then processed for DNA extraction and isolation. In the current study this refers to samples collected by kick-net, using the same sampling techniques used to collect material for traditional (visual) taxonomic identification. Environmental DNA (eDNA) is used to refer to collection of water samples from the environment of interest, from which DNA is extracted and isolated.

Bulk-tissue samples

Bulk-tissue samples were collected by kick-net sampling near Folsom Dam on 2/28/2019. Samples were analyzed from sites MI-8 (38.69476, -121.11303) and WC-1 (38.64637, -121.18687). Paired samples were collected, with one sample preserved for DNA analysis, and a second sample, collected at the same location and time, sent to BSA Environmental Services (Beachwood, OH) for traditional taxonomic identification and analysis. Samples designated for DNA analysis were preserved with a final concentration of 70% ethanol or isopropanol.

eDNA samples

Environmental DNA (eDNA) samples were collected from the marina at Canyon Lake, AZ (33.536307, -111.423006) on 11/06/2019. Samples were collected using a plankton tow net with a 64 micrometer pore size. Sample collection and preservation were performed following the Technical Service Center's Ecological Research Laboratory standard procedures.

DNA extraction

Bulk-tissue DNA samples

For kick-net bulk-tissue samples, the supernatant was removed from the sample and the solid portion of the sample was blended at high speed to homogenize it. 250 micrograms of the resultant slurry of organic material was placed in a clean 2.0 milliliter tube and allowed to dry in a vented hood. For each sample set a lab blank was also created as a negative control. For the lab blank, 250 microliters of ethanol was placed in a clean tube, and processed in parallel with field samples in all subsequent steps. DNA isolation for environmental samples and lab blanks was performed using the DNeasy PowerSoil Kit (QIAGEN, Waltham, MA), following the manufacturer's protocol. Samples were processed with the OneStep PCR Inhibitor Removal Kit (Zymo Research, Irvine, CA) to reduce the presence of organic compounds, which were found to inhibit PCR amplification during early testing.

eDNA samples

eDNA samples and field blanks collected by tow net were processed using Ecological Research Laboratory standard operating procedures (https://www.usbr.gov/mussels/). Briefly, a 40 ml aliquot of the samples was centrifuged, and a 250 ul subsample from the resultant pellet was used for DNA extraction and isolation. DNA extraction and isolation was performed using the Fisher BioReagents SurePrep Soil DNA Isolation Kit (Thermo Fisher Scientific, Waltham, MA), following the manufacturer's protocol. A lab blank containing deionized water was also created as a negative control and was processed in parallel with eDNA samples.

Mock community

Mock communities are those for which the composition is known a priori. Mock communities are generated by combining tissue or DNA from multiple organisms of known identity and processing these samples in parallel with unknown environmental samples. Mock communities provide an internal control for evaluating the fidelity of downstream steps such as PCR amplification and DNA sequencing, and may be used to evaluate the potential for biomass quantification from environmental metabarcoding samples (Braukmann et al., 2019; Lamb et al., 2019).

To generate mock communities, DNA isolation was performed on individual organisms, which had been isolated and taxonomically identified by BSA Environmental Services (Beachwood, OH). DNA isolation for organismal tissue samples was performed using the DNeasy Blood & Tissue kit (QIAGEN, Waltham, MA), following the manufacturer's protocol. DNA concentrations were measured using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific, Waltham, MA) and a Qubit 4 fluorometer (Thermo Fisher Scientific, Waltham, MA). Samples used to generate the mock community sample are listed in Table 1. Samples were selected to cover a range of taxonomic groups and concentrations were adjusted to cover four orders of magnitude in the mock community.

Table 1 Tissue DNA isolations included in the mock community sample

DNA source *	DNA concentration in mock community (ng/µl)
Procambrus sp.	18.14
Physidae	5.85
Aedes sp.	2.38
Callibaetis sp.	1.01
Notonecta sp.	0.34
Turbellaria	0.19
Cyzicus californicus	0.14
Daphnia sp.	0.10
Enochrus californicus	0.52
Cymbiodyta sp.	0.02

* Taxonomic identification of the source organisms from BSA Environmental Services

PCR amplification

Primer selection

Numerous oligonucleotide primer pairs have been proposed for amplification of *COI* for metabarcoding. Almost all these primer pairs target a subregion of the fragment amplified by the Folmer primers. A smaller fragment is desirable because DNA sequencing with the Illumina sequencing platform (detailed below) produces sequences too short to cover the entire Folmer fragment. Ideally, the sequenced fragment should be short enough such that the entire sequence can be covered by paired-end (PE) sequencing, and the complementary reads (raw DNA sequence data) overlap sufficiently so as to allow joining following denoising (removal of regions of DNA sequence of low quality; see the descriptions of fastq files and data analysis below).

For the current study, primers published by Elbrecht and Leese (Elbrecht & Leese, 2017) were selected based upon their reported ability to amplify from a wide diversity of freshwater macroinvertebrate taxa. The forward primers BF1 and BF2 were each used in combination with the reverse primer BR2 in independent reactions (Table 2).

Table 2 Oligonucleotide primers used for eDNA and mock community PCR amplification

Primer name	Primer sequence	Source
BF1	5′-ACW GGW TGR ACW GTN TAY CC-3′	Elbrecht & Leese, 2017
BF2	5′-GCH CCH GAY ATR GCH TTY CC-3′	Elbrecht & Leese, 2017
BR2	5′-TCD GGR TGN CCR AAR AAY CA-3′	Elbrecht & Leese, 2017
BF1-IlluminaF *	5′-aca ctc ttt ccc tac acg acg ctc ttc cga tct ACW GGW TGR ACW GTN TAY CC-3′	current study
BF2-IlluminaF *	5′-aca ctc ttt ccc tac acg acg ctc ttc cga tct GCH CCH GAY ATR GCH TTY CC-3′	current study
BR2-IlluminaR *	5′-gac tgg agt tca gac gtg tgc tct tcc gat ctT CDG GRT GNC CRA ARA AYC A-3′	current study
IlluminaF	5′-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT-3′	Genewiz
IlluminaR	5′-GAC TGG AGT TCA GAC GTG TGC TCT TCC GAT CT-3′	Genewiz

* For clarity, the Illumina adapted portion of fusion primers is show in lowercase

PCR amplification

Multiple approaches were tested to optimize polymerase chain reaction (PCR) of the *COI* fragments from eDNA and bulk-tissue samples. Because Illumina sequencing requires that fragments have standardized adapter sequences at their ends, a 2-step amplification protocol was developed. In the first step, templates were amplified with the fusion primers (primer pairs: BF1-IlluminaF + BR2-IlluminaR or BF2-IlluminaF + BR2-IlluminaR). In the second step, 2 microliters of product from the first-round reaction was used is a reaction with the Illumina adapters, IlluminaF and IlluminaR,

as primers. For simplicity descriptions of amplifications in the remainder of this report will reference only the names of the metabarcoding portions of the primers used (BF1, BF2, and BR2), unless otherwise noted.

The two-step strategy was selected to maximize yield to ensure sufficient product for sequencing, while minimizing the number of PCR cycles. PCR bias (preferential priming and amplification of a subset of templates) increases with greater numbers of cycles, so minimizing the number of cycles is ideal (Kanagawa, 2003). PCR amplification programs were as follows:

PCR amplification round 1:

	Temperature	Length (minutes:seconds)
Initial denaturation	94°C	2:00
Amplification (25 cycles)		
Denaturation	94°C	0:30
Annealing	50°C	0:30
Extension	72°C	1:00
Final Extension	72°C	2:00

PCR amplification round 2:

	Temperature	Length (minutes:seconds)
Initial denaturation	94°C	2:00
Amplification (20 cycles)		
Denaturation	94°C	0:30
Annealing	65°C	0:30
Extension	72°C	1:00
Final Extension	72°C	2:00

Both first and second-round PCR amplification reactions were performed with Invitrogen Platinum II Hot-Start Master Mix (2X) (Thermo Fisher Scientific, Waltham, MA). All reactions contained 1x Platinum II Master Mix, and 0.5 µM of both forward and reverse primers. For first round reactions, the total volume was 20 µl, and the amount of DNA template (isolated environmental DNA or controls) added was adjusted such that each was 25 ng. For second round reactions the total volume was 50 µl, with 2 µl of product from the first-round reaction added as template.

PCR products were run on a 1.5% agarose gel with TAE buffer, and stained with Invitrogen SYBR Safe DNA Gel Stain (Thermo Fisher Scientific, Waltham, MA) to confirm produce size. The expected product sizes were: 381 base pairs (bp) for products with the BF1 and BR2

primers and Illumina adapters (316 bp of informative sequence exclusive of primers), and 486 bp for products with the BF2 and BR2 primers and Illumina adapters (421 bp of informative sequence exclusive of primers).

PCR reactions which showed a strong single band of product of the expected size were purified using MinElute PCR Purification Kit (QIAGEN, Waltham, MA). Sample elution was performed with 10 mM Tris buffer. PCR product concentration was measured using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific, Waltham, MA) and a Qubit 4 fluorometer (Thermo Fisher Scientific, Waltham, MA).

DNA sequencing

Illumina high-throughput DNA sequencing was performed by Genewiz (South Plainfield, NJ), using their Amplicon-EZ service. The Amplicon-EZ service provides library preparation and 2x250 bp Illumina PE sequencing with a target output of approximately 50,000 reads per sample. The cost per sample was $50 (for products with Illumina adapters), and the turn-around time from sample receipt to data delivery is less than two weeks.

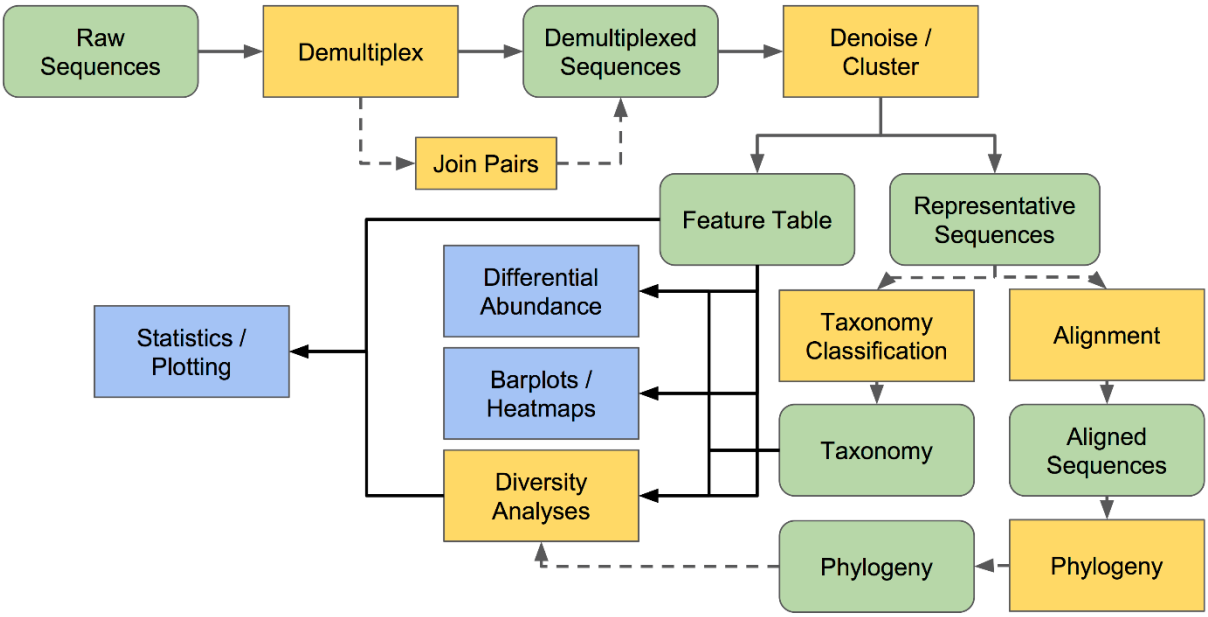
For all samples, excluding negative controls, sample volume and concentration was normalized to 20ul of 25ng/µl (500 ng total). Samples were shipped overnight to Genewiz for processing.

Data analysis

Despite the significant interest in the use of *COI* for metabarcoding, particularly of invertebrates, a literature search did not identify a software package specifically tailored to the analysis of such datasets. Most studies published to date on this topic use software originally designed for metabarcoding analysis of microbial populations and/or customized software pipelines. QIIME 2 is one software package that was originally designed for microbial studies, but which has been extensively used for analysis of a range of datasets (Bolyen et al., 2019). QIIME 2 is a modular collection of programs each designed to perform specific task. A simplified diagram of QIIME 2 data analysis is shown in Figure 1. The precise set of commands and options used is dependent upon the specific datasets being analyzed and the research questions being pursued. QIIME 2 provides a powerful and flexible set of tools, including data visualization and metadata files that include provenance, facilitating analysis refinement and reproducibility. QIIME 2 provides extensive documentation and tutorials through the project website (www.qiime2.org). The developer website also hosts an active and moderated forum for technical inquiries and support.

QIIME 2 does not natively support installation under Windows operating systems. However, QIIME 2 is available in a VirtualBox Image, and detailed installation instructions are provided on the www.qiime2.org website. QIIME 2 2020.08 was installed and run on VirtualBox 6.1 (www.virtualbox.org) under Ubuntu 18.04.5 LTS (www.ubuntu.com).

Figure 1 QIIME 2 workflow (from https://docs.qiime2.org/2020.8/tutorials/overview/ used with permission)



Because all metabarcoding data are intrinsically similar, in that they are composed of large collections of presumably homologous short nucleotide sequences, regardless of the source material, targeted taxa, or amplification methods, the flexibility of QIIME 2 allows it to be applied to a wide variety of studies.

However, because QIIME 2 was designed with microbial data in mind, not all features are directly or easily applied to other datasets. In particular, taxonomic assignment from metazoan *COI* sequences is not directly supported. Although a custom database of reference sequences can be developed and imported, this study utilized other software tools, as detailed below.

Results

Sequencing and initial quality control

Sequences were retrieved from Genewiz in .fastq format. A summary of the sequencing statistics, including the number of reads and quality statistics, is shown in Table 3.

Table 3 Summary statistics for sequencing results

Sample	Forward primer	Reverse primer	# Reads	Yield (Mbases)	Mean Quality Score	% Bases >= 30
Folsom MI-8	BF1	BR2	26133	13	35.84	89.53
Folsom MI-8	BF2	BR2	28869	14	31.86	73.2
Folsom WC-1	BF1	BR2	27314	14	33.88	80.8
Folsom WC-1	BF2	BR2	31143	16	32.05	74.27
Canyon 3	BF1	BR2	37782	19	37.68	97.35
Canyon 3	BF2	BR2	36339	18	35.47	88.53
Mock community	BF1	BR2	35309	18	37.21	95.41
Mock community	BF2	BR2	37809	19	35.22	87.27
Field blank	BF1	BR2	32054	16	25.41	43.37
Field blank	BF2	BR2	36440	18	25.27	43.1
Lab blank	BF1	BR2	37138	19	24.28	38.27
Lab blank	BF2	BR2	37861	19	24.38	39.05

Fastq is a widely used format for high-throughput DNA sequence data, which includes nucleotide sequence data as well as quality scores. Quality scores in fastq files are referred to as Q Scores, and are summarized in Table 3. They are a measure of the probability that individual base calls are correct. Q Scores are presented on a scale of 0 to 40, and are derived from the formula:

$$\mathbf{Q = -10\log_{10}(p)}$$

where **p** is the estimated probability of an individual base call being incorrect. Therefore, a Q Score of 30 equates to a 99.9% probability that the base call is correct, while a Q Score of 40 equates to a 99.99% probability that the base call is correct.

Q Scores averaged above 30 for all sequence sets from eDNA and mock community samples. Field and extraction blanks showed lower quality scores, with average Q Scores below 30, and the percentage of bases with Q Scores 30 or higher was less than 50%.

More detailed analysis of sequence quality was performed on individual fastq files using FastQC, a Java-based software program that provides a variety of quality statistics (Andrews, 2010). FastQC analysis showed several patterns of sequence length, depending upon the sample and primers used. For samples from Canyon Lake, taken from filtered water, FastQC showed the vast majority of sequences had the targeted length of 250 bp (Figure 2A). For the field and laboratory blank datasets almost all of sequences were shorter than 40 bp (Figure 2B). This is consistent with low levels of contamination, with the majority of reads likely derived from the Illumina library preparation reaction either not finding a template or indexing PCR primer that was carried over during the reaction clean-up. An unexpected pattern was observed for sequences from the Folsom sites WC-1 and MI-8, which were derived from macerated organic material collected in kick-net samples. In these samples there was a significant peak in sequence lengths around 140 bp, particularly from samples amplified with the BF2 and BR2 primer pair (Figure 2C). For these samples all sequences were expected to be 250 bp in length, given that the targeted *COI* fragments are over 250 bp and their lengths are highly conserved across invertebrates. The presence of this lower peak in sequence length was suggestive of contamination in the PCR products used for sequencing.

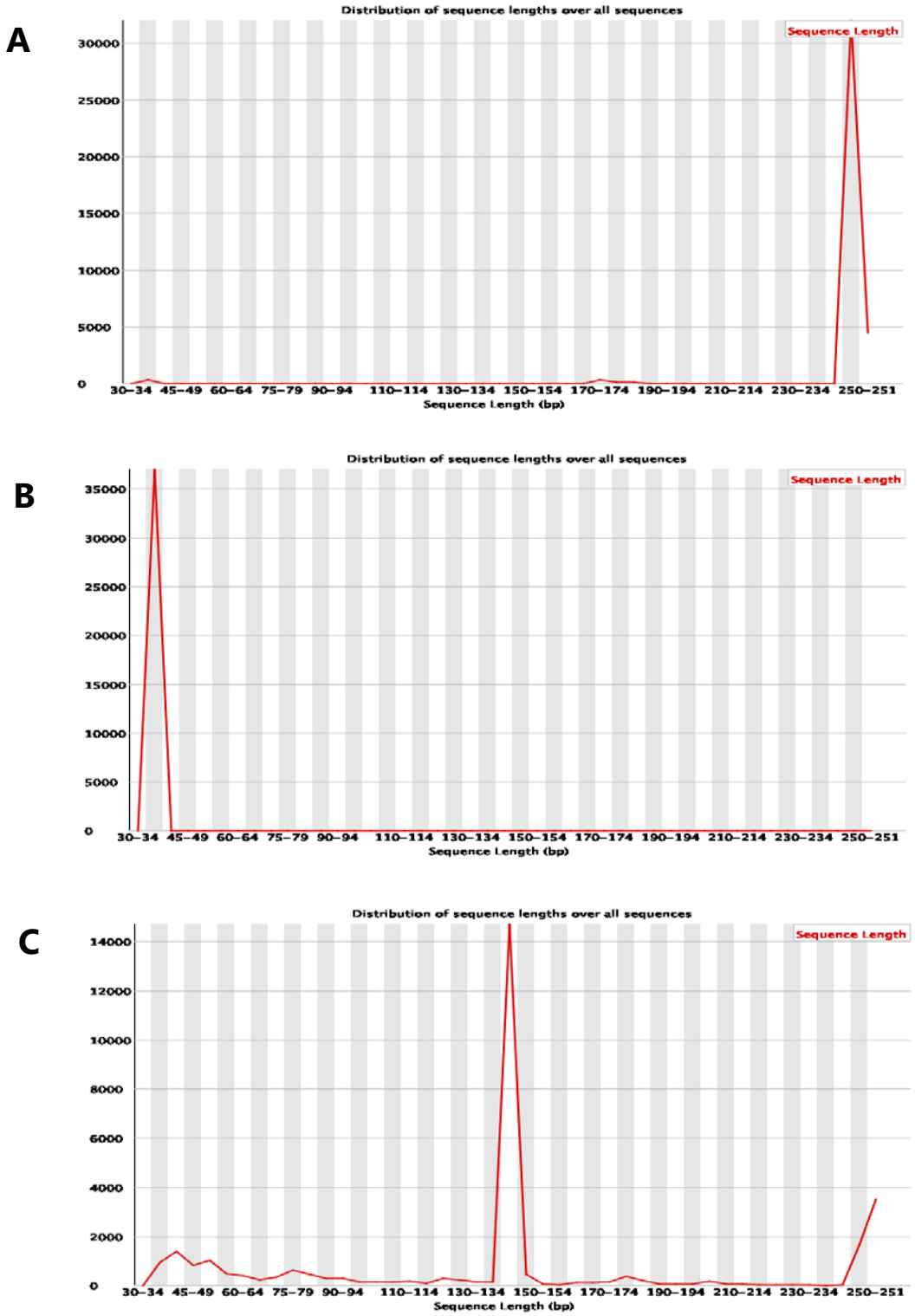


Figure 2 Sequence length size distributions from FastQC. (A) Canyon 3 filtered sample amplifies with BF1 and BR2 derived primers. (B) Lab extraction blank amplified with BF1 and BR2 derived primers. (C) Folsom WC1 kick-net bulk-tissue sample amplified with BF2 and BR2 derived primers.

Data analysis

QIIME 2 quality control and inference of ASVs

All initial sample processing and analysis was performed in QIIME 2 2020.08 (Bolyen et al., 2019). Briefly, raw sequence data in fastq files were imported to QIIME 2, and demultiplexed using q2-demux. Demultiplexing is the operation of sorting sequences from different samples that were pooled in the same reaction for sequencing. Individual sample sets can be recovered based upon unique marker tag sequences added during sequencing library construction. Initial demultiplexing was performed by GENEWIZ, and sequence data for each amplicon submitted for sequencing were retrieved as unique fastq files. Regardless of this, the demultiplexing step is a key step in data processing in QIIME, and is required for sample tracking and subsequent joining of complementary paired-end reads.

Summarized read quality statistics were checked visually, and low-quality regions were noted for removal. Denoising was performed with DADA2 (Callahan et al., 2016), using parameters determined from visual inspection of the data. An exemplar of QIIME 2 commands used to perform data analysis is shown in Appendix 1.

DADA2 joins forward and reverse paired-end reads, and clusters identical sequences into features referred to as amplicon sequence variants (ASVs). The number of identical sequences in the joined dataset that match a particular ASV is referred to as the feature frequency. Unique sequences that appear only once in the joined dataset are considered to potentially have resulted from a sequencing error and are not included in ASVs.

Sequence length variance

As discussed above, fastq sequence libraries displayed variance in sequence lengths, with an unexpectedly high number of intermediate length sequences, particularly from bulk-tissue samples. Although PCR amplification was intended to target invertebrate taxa, initial efforts at taxonomic assignment (detailed below) evidenced a relatively high degree of non-specific amplification. Preliminary analysis of full ASV datasets returned a range of taxonomic groupings, including bacteria and non-metazoan eukaryotes. Inspection of sequence lengths demonstrated that most of these "non-specific" ASVs diverged from the expected target sequence lengths (316 bp of informative sequence from primer pair BF1 and BR2, and 421 bp of informative sequence from primer pair BF2 and BR2). Given this, QIIME 2 was used to filter the reference sequence datasets (Figures 3 & 4; Appendix B). For primer pair BF1 and BR2, sequences of lengths 310 bp to 320 bp were retained, while for primer pair BF2 and BR2, sequences of lengths 416 bp to 425 bp were retained. For samples where fastq files were trimmed beyond the first 20 nucleotides corresponding to the BF1, BF2, or BR2 primer sequences during denoising with DADA2, filtered sequence lengths were adjusted accordingly (e.g. 36 bases for forward and reverse sequences from the MI-8 bulk-tissue amplified with BF1 and BR2; see Appendix 1).

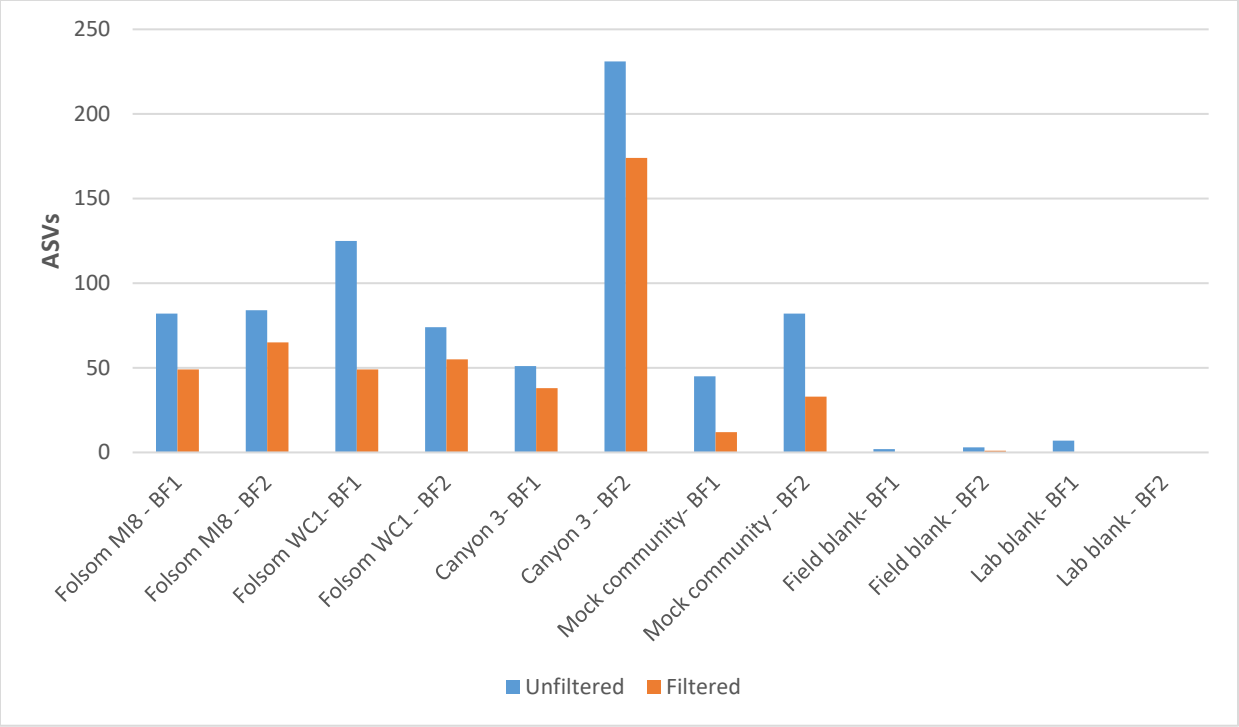


Figure 3 Chart of total ASVs for unfiltered and filtered datasets

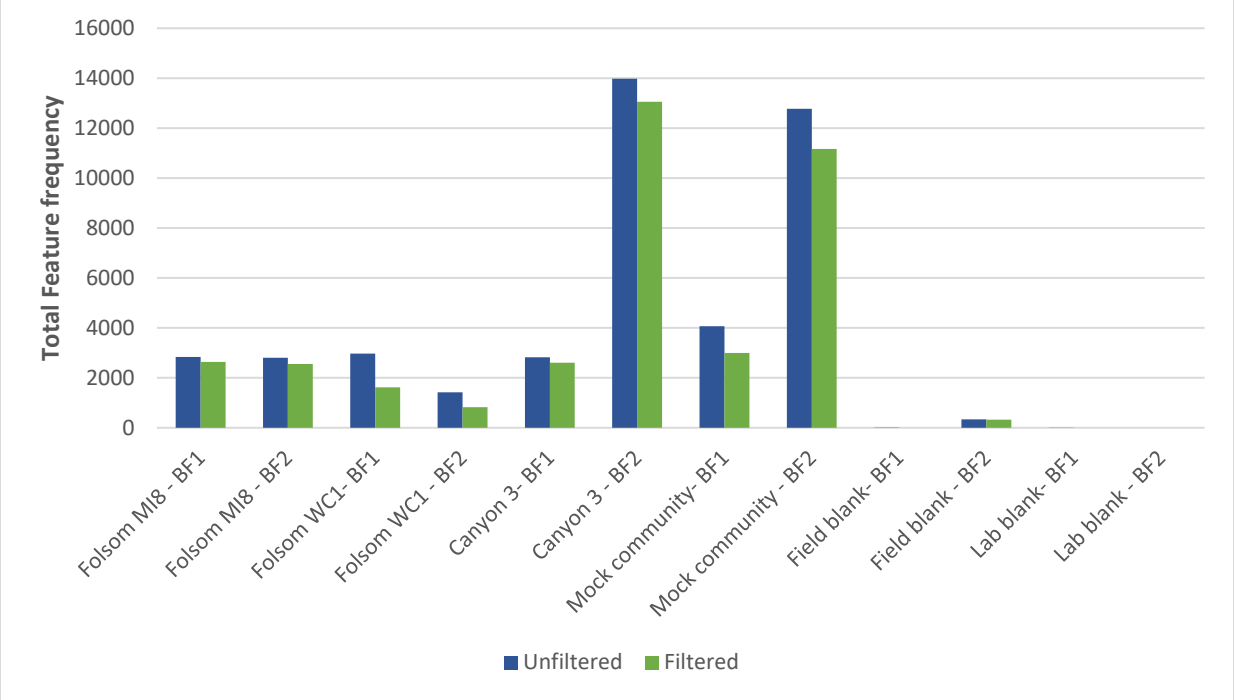


Figure 4 Chart of total feature frequencies of ASVs for unfiltered and filtered datasets

While targeted *COI* fragments are expected to be highly constrained in length across invertebrate taxa, the ranges of sequence lengths listed above were selected as a result of observed variation in the dataset among ASVs identified as being from invertebrate taxa. Whether these differences in sequence length represented true biological signal or were a stochastic result of errors in PCR

replication or sequencing was not investigated further in this study. Filtered datasets were used for all subsequent taxonomic identification.

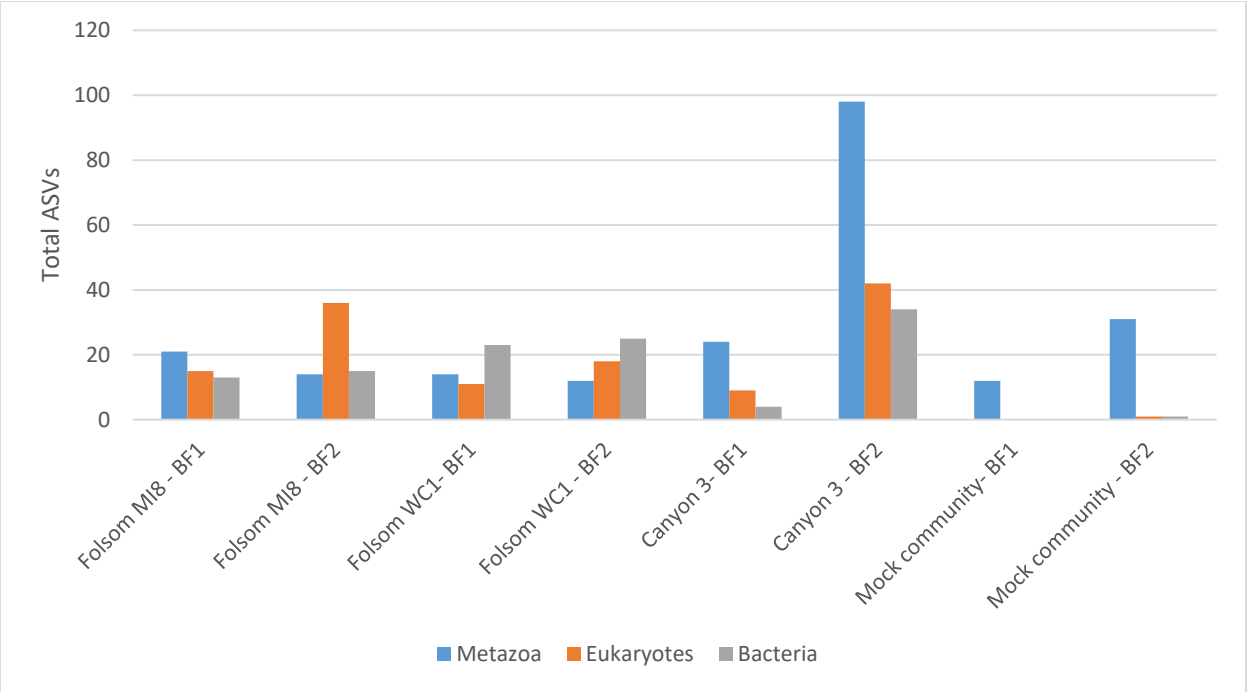


Figure 5 Total ASVs identified for major taxonomic groups in filtered datasets

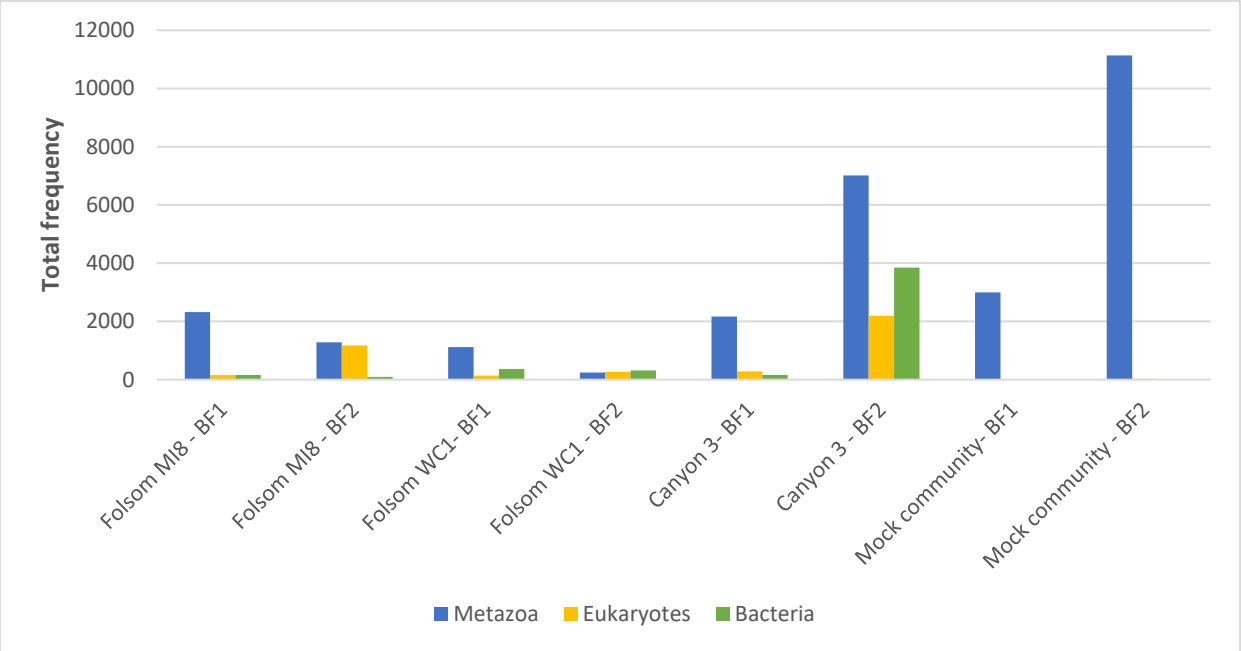


Figure 6 Total feature frequencies of ASVs identified for major taxonomic groups in filtered datasets

Taxonomic identification (as described below) demonstrated that even after filtering a considerable proportion of ASVs appeared to be derived from bacteria or non-metazoan eukaryotes (Figures 5 &

6; Appendix C). The proportion of non-metazoan ASVs was higher in the Folsom datasets than the Canyon samples.

Taxonomic identification

Resultant ASVs from DADA2 were exported in .fasta format files and taxonomic assignment was performed with reference files from the BOLD database as collected in the CO1Classifier trained dataset (Porter & Hajibabaei, 2018b), using Ribosomal Database Project (RDP) classifier software (Wang et al., 2007). The RDP classifier is a naïve Bayesian classifier which can provide more accurate and reliable taxonomy assignment as compared to traditional methods such as BLAST, which rely upon sequence similarity (Porter et al., 2014). Output from the RDP classifier includes a hierarchy of taxonomic assignments for each ASV, with bootstrap support for each level. Bootstrap support, which is presented on a scale of 0 to 1 (1 being 100% support), provides a measure of confidence in each level of taxonomic assignment. Bootstrap support of 0.5 is generally considered the minimum for an assignment to be reliable, although the more conservative value of 0.8 is frequently used to ensure that assignments are robust and well supported. For many sequences, reliable bootstrap may only be obtained at higher taxonomic levels, rather than at the level of genus or species.

The RDP classifier can provide a more nuanced understanding of the data than is available for a BLAST search which only provides percent similarity to the closest matches in the queried database. The BOLD Identification Engine tool (http://www.boldsystems.org/index.php/IDS_OpenIdEngine) utilizes a more sophisticated search algorithm, pairing an initial BLAST search with a subsequence Hidden Markov Model analysis of the protein translation for the sequence. However, because the BOLD Identification Engine aims to provide species level matches, it uses a very high threshold of 99% similarity to assign a match. Queries against the BOLD species level barcode database will return no match if a reference sequence of sufficient similarity is not found. However, a query against all barcode records on BOLD can return records for matches with lower similarity. For such results the percent similarity is reported, as well as a simple neighbor-joining based phylogenetic tree.

Comparison of the different approaches to taxonomic classification found that no one method may be considered best, and the choice of one or more tools may depend on the study objects. Use of the RDP classifier with the CO1Classifier dataset is likely the most robust and least prone to error. However, this approach may also be considered overly conservative in some contexts.

For example, one sequence that was queried using the RDP classifier was found to have bootstrap support of only 0.23, with the next highest support being 0.8 for placement in the phylum Arthropoda. This would suggest relatively low confidence for taxonomic assignment below the level of phylum. In contrast, a query of the same sequence with the BOLD Identification Engine provided a top match with 100% similarity to a reference sequence for the ostracod species *Cypridopsis vidua*. BOLD Identification Engine matches for samples labeled as *Cypridopsis vidua* ranged from 100% similarity down to 81.67% similarity. This degree of divergence would normally be expected between, rather than within, species, and could suggest misidentification or "clumping" of multiple species under a single identifier. However, *Cypridopsis vidua* displays an unusual reproductive strategy, apomictic parthenogenesis, and clonal lines have been documented to display high degrees of sequence divergence (Cywinska & Hebert, 2002; Havel & Hebert, 1989). The low bootstrap value recovered from the RDP classifier may have been influenced by this sequence diversity.

Analysis of all datasets was initially performed with the RDP classifier and CO1Classifier trained dataset based upon the fact that this approach provided bootstrap support for taxonomic assignments. Levels of taxonomic assignment for ASVs identified as being of metazoan origin are shown for each dataset Figures 7 & 8. Source values for these charts are presented in Appendix D. For ASVs identified as being from metazoan taxa, the proportion that could be confidently assigned to the level of species (bootstrap ≥ 0.5) ranged from 20% to 90% among the datasets.

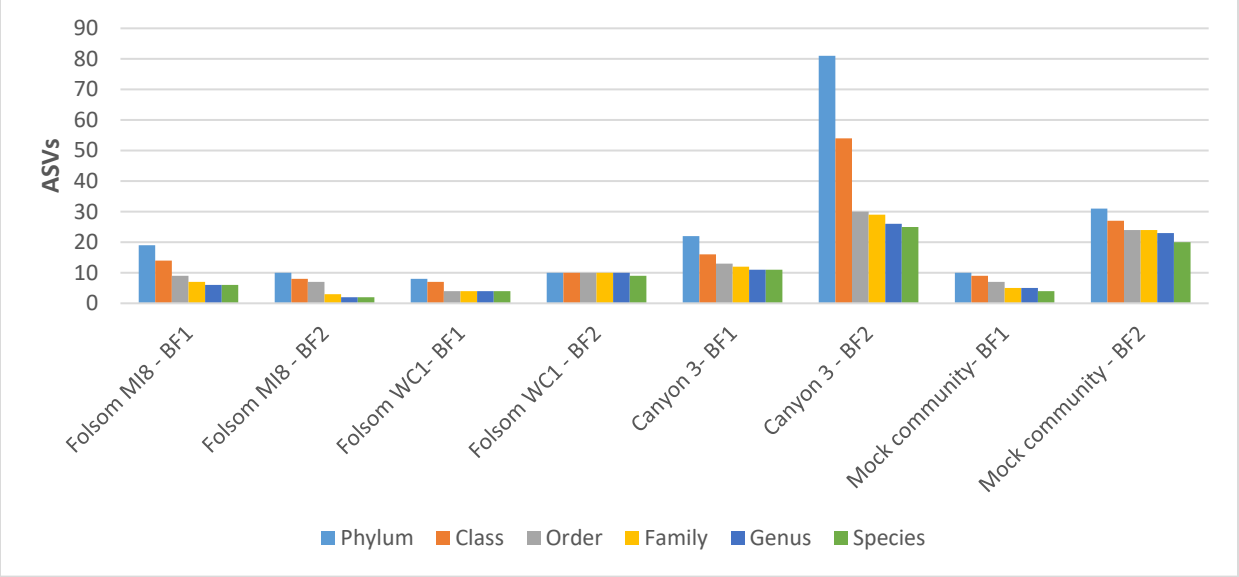


Figure 7 Total ASV numbers identified as metazoan to the level of phylum, class, order, family, genus, and species with bootstrap values ≥ 0.5

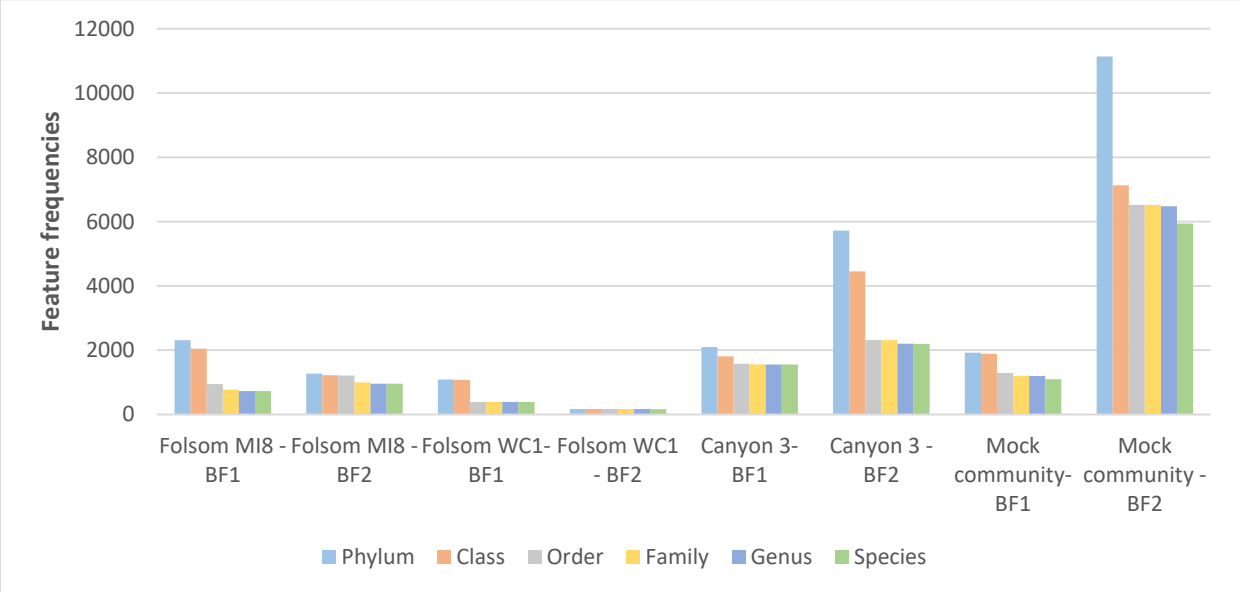


Figure 8 Total feature frequencies of ASVs identified as metazoan to the level of phylum, class, order, family, genus, and species with bootstrap values ≥ 0.5

While the direct comparability of ranked hierarchical taxonomic assignments between taxa should be viewed with caution (de Queiroz & Gauthier, 1994), these assignments and the associated bootstrap values provide a useful placeholder for assessing the level of taxonomic specificity that may be confidently attached to an ASV.

Metabarcoding versus traditional taxonomy

For samples from sites Folsom MI-8 and Folsom WC-1, samples collected in parallel were sent to BSA Environmental Services for visual taxonomic identification. BSA Environmental Services identified 5 taxa from the MI-8 sample, and 6 taxa in the WC-1 sample (Tables 4-7). All were identified to the level of species or genus. Filtered metabarcoding data resulted in comparable numbers of ASVs identified to the genus or species level, depending upon the primer pair used (Appendix D). Comparison of the datasets demonstrated the metabarcoding datasets recovered only a subset of the taxa identified by visual inspection of the samples (Tables 4-7). Other ASVs recovered from DADA2 were found to match taxa not identified by BSA Environmental Services or were not reliably identified beyond the level of phylum from the RDP classifier (Figures 7 & 8).

Table 4 Visual taxonomy versus RDP Classifier identification for sample Folsom MI-8 amplified with primers BF1 and BR2

Taxonomy visual	Count visual	RDP Classifier best match	RDP bootstrap	ASV frequency
Daphnia sp.	9396	*Daphnia pulex*	0.95	602
Crangonyx sp.	82	none	n/a	n/a
Procambarus sp.	6	none	n/a	n/a
Physidae	6	*Physella acuta*	1	113
Lumbriculidae	1	none	n/a	n/a

Table 5 Visual taxonomy versus RDP Classifier identification for sample Folsom MI-8 amplified with primers BF2 and BR2

Taxonomy visual	Count visual	RDP Classifier best match	RDP bootstrap	ASV frequency
Daphnia sp.	9396	*Daphnia pulex*	1	958
Crangonyx sp.	82	none	n/a	n/a
Procambarus sp.	6	none	n/a	n/a
Physidae	6	none	n/a	n/a
Lumbriculidae	1	*Allonais_paraguayensis*	0.07	29

Table 6 Visual taxonomy versus RDP Classifier identification for sample Folsom WC-1 amplified with primers BF1 and BR2

Taxonomy visual	Count visual	RDP Classifier best match	RDP bootstrap	ASV frequency
Daphnia sp.	3070	*Daphnia pulex*	1	282
Chironomus sp.	199	*Chironomus_maturus*	1	64
Crangonyx sp.	81	none	n/a	n/a
Aedes sp.	7	*Ochlerotatus_increpitus*	0.87	45
Musculium sp.	6	none	n/a	n/a
Physidae	1	none	n/a	n/a

Table 7 Visual taxonomy versus RDP Classifier identification for sample Folsom WC-1 amplified with primers BF2 and BR2

Taxonomy visual	Count visual	RDP Classifier best match	RDP bootstrap	ASV frequency
Daphnia sp.	3070	*Daphnia_pulex*	0.97	54
Chironomus sp.	199	*Chironomus_maturus*	1	84
Crangonyx sp.	81	none	n/a	n/a
Aedes sp.	7	*Ochlerotatus_increpitus*	0.94	24
Musculium sp.	6	none	n/a	n/a
Physidae	1	none	n/a	n/a

Invasive species and taxonomic diversity

Canyon Lake sample 3 was used for metabarcoding to evaluate the potential to use this tool for early detection of invasive species. Canyon Lake, AZ has an established population of quagga mussel (*Dreissena rostriformis bugensis*). Prior to metabarcoding, both primer pairs (BF1 and BR2; BF2 and BR2) were tested using DNA extracted from quagga mussel tissue and found to amplify PCR products of the expected sizes. Canyon Lake sample 3 was also tested by species-specific quantitative PCR (qPCR) and quagga mussel DNA was detected in the sample. Despite this, no ASVs matching quagga mussel were found in the two datasets, based upon both RDP taxonomy assignment and BLAST searches against the datasets.

Although ASVs matching to quagga mussel were not recovered, ASVs matching to another invasive species, the water flea *Daphnia lumholtzi*, were identified in Canyon Lake datasets from both primer pairs. *Daphnia lumholtzi* has previously been identified from visual surveys as being present in Canyon Lake (https://nas.er.usgs.gov/queries/factsheet.aspx?SpeciesID=164).

The primer pairs used did amplify sequences who's ASVs were identified as matching to a wide range of invertebrate taxa, including annelids, bryozoans, rotifers, and tardigrades. Such diversity of sequences could aid in broadening both ecological surveys and early detection of invasive species.

Negative controls

Field and lab controls had few or zero joined sequences, consistent with very low levels of contamination in these samples (Table 4). For the field blank there was one ASV that passed size filtering. This ASV was found to match sequence for *Cyphomella cornea*, a chironomids or non-biting midge. The larva of chironomids are aquatic, while the adults are terrestrial and can fly. DNA in the filed blank could therefore represent either contamination on the net from a previous field sample collection, or could have been derived from an adult inadvertently caught in the tow net when the field blank was captured. Interestingly, DNA from this organism was amplified with the BF2 and BR2 primer pair, but not with the BF1 and BR2 primer pair.

Conclusions

The present study aimed to initiate establishment of metabarcoding techniques in the Technical Service Center's Ecological Research Laboratory, and to evaluate its applicability to diversity surveys and invasive species early detection. Metabarcoding was found to have potential for use in ecological surveys and early detection of invasive species. A number of lessons from this pilot study should be given careful consideration for future studies and prior to adoption of this technique beyond the realm of research projects.

Experimental design

Sampling strategy

For the current study, two sampling methods were tested, bulk-tissue collection and processing, and tow net filtration. Many metabarcoding studies use a third strategy, where a volume of water (usually 1 to 2 liters) is collected at or near the surface and run through a filter with a small pore size. Direct comparison of different methods was not conducted in this study. Choice of sampling method could affect the range and diversity of taxa identified. The presence of sediment and detritus in the bulk-tissue samples processed from the Folsom sites could have contributed to the high proportion of ASVs assigned to non-metazoan eukaryote taxa and bacteria in these datasets. A direct comparison of different methods for sample collection from the same site would aid in selection of the most appropriate approach for future metabarcoding studies.

Primer selection

The primers BF1, BF2, and BR2 (Elbrecht & Leese, 2017) were selected based upon their expected ability to amplify *COI* fragments from a wide variety of metazoan invertebrate taxa. These primers did appear to perform to this degree, with ASVs from field samples matching to annelids, bryozoans, rotifers, and tardigrades, in addition to a range of arthropods.

However, a large proportion of ASVs did not match to metazoan *COI* sequences, but rather appeared to be derived from non-metazoan eukaryotes or bacteria. For non-metazoan eukaryotes these sequences did appear to derive from the *COI* gene. Bacteria do not have a direct ortholog for *COI*, and these sequences appeared to derive from amplification of other regions of bacterial genomes. This off-target amplification could be due in part to the high degree of degeneracy in the selected primers (128-fold for BF1; 216-fold for BF2; 192-fold for BR2). While these high levels of degeneracy were intended to expand the diversity of invertebrate metazoans recognized by the primers, they appear to have also contributed to off-target amplification. There is an inherent trade-off in that while increasing primer degeneracy can broaden the range of taxa of interest that may be recognized and amplified, it also increases the probability that off-target taxa or genomic regions will be amplified.

For invertebrate metazoans a wide range of primers have been proposed for metabarcoding of *COI*. As yet, no single primer pair has seen widespread adoption for invertebrate metabarcoding studies. In addition, a number of markers other than *COI*, including *cytB*, *16S rRNA*, and *18S rRNA*, have been investigated and proposed for use. This is in contrast to the situation for some other groups of organisms, such as for fish and bacteria, where select primer pairs are used for the majority of studies. A number of efforts are currently under way to carefully evaluate metabarcoding methodologies for freshwater aquatic invertebrates, including marker and primer selection. Consensus on community standards for these studies is expected to greatly facilitate both adoption of the technique and comparability of data between studies.

Sequencing depth

A sequencing depth of 50,000 reads per amplicon was targeted in this study, based upon a review of the current literature and the availability of an affordable and streamlined sequencing service with this output from the selected vendor, Genewiz. The number of recovered reads was lower than this target, with an average of 33,683 reads per sample. This discrepancy may have been due to the quality and quantity of the amplicons submitted for sequencing and/or to shortfalls during library construction and sequencing by the vendor. After demultiplexing and denoising steps in QIIME 2 (but before filtering based on feature length), this resulted in an average of 5,459 joined features for the environmental and mock community samples. Future efforts should optimize the number of reads generated and the number features recovered from data processing. Power tests should also be employed on pilot data to estimate the necessary sequencing depth required to recover that range of targets, and resultant amplicons, predicted to be present in the sample.

Selection of more specific primers may help to increase the effective sequencing depth, given the proportion of features that were removed after filtering for length or were identified as matching to non-invertebrate sources.

Data analysis

As with primer selection, no standard methods currently exist for analysis of invertebrate metabarcoding data. A high degree of analysis customization will always be necessary, given the size and complexity of metabarcoding datasets, and the broad range of research designs and objectives it will be used for. However, to date a wide range of software programs and analytical approaches are currently employed for invertebrate data. Many, such as QIIME 2, were originally designed for

datasets from other taxa, such as surveys of bacterial communities. While commonalities across metabarcoding datasets from different taxa allow such flexibility, the use of some tool, such as ready-made reference sequence libraries, may not be available. More tailored software packages may become available in the future, and would likely aid in more widespread adoption of metabarcoding approaches to surveys of aquatic invertebrate communities.

The absence of quality reference sequence libraries formatted for use in available software packages was identified as a particular issue in this study. Most studies currently use bespoke sequence libraries, often drawn from available data in public repositories such as NCBI GenBank and BOLD. Given the geographic and taxonomic diversity of invertebrate taxa, this may continue to be the best approach. Ideally, a reference library would be developed from organisms collected at the sample site and individually barcoded. Analysis could then be supplemented with searches against publicly available datasets to identify sequences not matching to the site-specific library. While such an approach adds to the effort required it would ensure the most reliable interpretation of the data.

The choice of identification strategies was also seen to have a significant impact on the outcome of the analysis. Where sequenced features are a perfect match to a sequence in the reference library it is expected that disparate approaches should converge on the same results, however exceptions to this expectation were observed. When relying on publicly available data, matches with lower similarity may be the best that is obtained. Ideally, assignment to a taxonomic level other than species would still be resolved with some reasonable degree of confidence attached to it. The use of RDP classifier, built on a naïve Bayesian classifier, is appealing for its potential to provide such output, however its performance in this study was variable. In the near term the use of more than one dentification method on a single dataset may be desirable, with congruent identifications from more than one approach taken to be the most reliable.

Detection of invasive species

Conversations with the scientific community and land managers during this project evidenced interest in utilization of metabarcoding as a potential tool for early detection of invasive species. This approach has also been championed in several recent publications (Borrell et al., 2017; Klymus et al., 2017; Mychek-Londer et al., 2020; Westfall et al., 2020). The idea is that a broad range of potential invaders could be screened for with one or a few assays, rather than each requiring an individual test as is currently the standard with species-specific qPCR assays. While this approach is appealing, the current study evidences that it should be approached with great care. While laboratory tests suggested that the selected primers should amplify quagga mussel sequences, and the Canyon Lake sample was known to contain quagga mussel DNA, matching sequences were not recovered in the metabarcoding datasets. Whether this was attributable to primer specificity/bias, sequencing depth, or some other experimental conditions bears further investigation. At a minimum it is suggestive that metabarcoding approaches to invasive species detection require a degree of validation comparable to that considered necessary to vet single-species qPCR assays, including testing from known positive environments, rather than relying on laboratory tests of mock communities. For taxa where standards for sample processing and data analysis are well developed, such as for fish, metabarcoding technology has already reached a point where it may be directly applied to surveys for invasive species. It is expected that tools and methodologies applicable to surveys for invasive invertebrates will likewise become more standardized in the near future.

As discussed above, development of a carefully curated reference sequence library will be critical to using metabarcoding for detecting invasive species. Such a library should include only sequences validated as being amplified by the primers utilized. This is necessary to minimize misinterpretations of non-detection results.

Visual taxonomy and metabarcoding

Metabarcoding has been proposed to have several potential benefits over traditional visual taxonomic identification. Among these are a decreased requirement for specialized taxonomic knowledge, decreased sample handing and analysis times, decreased costs as multiple samples may be processed and analyzed in parallel, and decreased disturbance of fragile environments during sample collection. While all of these arguments have merit, metabarcoding is perhaps best viewed as a complementary approach to traditional taxonomy, rather than a replacement for it. For example, in large-scale temporal studies it may be advantageous to initially perform traditional identification, as this can provide ground truthing for subsequent metabarcoding datasets, and can provide material for development of a site-specific reference library.

Use of metabarcoding at the Bureau of Reclamation

Metabarcoding is a rapidly developing approach that has been shown to have a wide variety of applications. It is already a standard approach to the investigation of microbial communities, and it is expected that its use for surveying and studying macro-organisms will increase rapidly in the coming years. This study investigated the tools available for sample collection and processing, data collection, and data analysis. Knowledge gained during this project has facilitated development of a project to utilize metabarcoding to survey for the presence of invasive fish. It is expected that the utility of this approach for surveys of aquatic macroinvertebrate communities will increase as methodologies and tools become more standardized in the near future.

References

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, *27*(20), 3968–3975. https://doi.org/10.1111/mec.14844

Aqem Consortium. (2002). *Manual for the application of the AQEM system* (A Comprehensive Method to Assess European Streams Using Benthic Macroinvertebrates, Developed for the Purpose of the Water Framework Directive. Version 1(02)).

Barbour, M. T. (1999). *Rapid bioassessment protocols for use in wadeable streams and rivers: Periphyton, benthic macroinvertebrates and fish*. US Environmental Protection Agency, Office of Water.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Borrell, Y. J., Miralles, L., Do Huu, H., Mohammed-Geba, K., & Garcia-Vazquez, E. (2017). DNA in a bottle—Rapid metabarcoding survey for early alerts of invasive species in ports. *PLOS ONE*, *12*(9), e0183347. https://doi.org/10.1371/journal.pone.0183347

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., de Waard, J. R., Sones, J. E., Zakharov, E. V., & Hebert, P. D. N. (2019). Metabarcoding a

diverse arthropod mock community. *Molecular Ecology Resources*, *19*(3), 711–727. https://doi.org/10.1111/1755-0998.13008

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Canadian Aquatic Biomonitoring Network, Carter, L., Canada, & Environment Canada. (2012). *Canadian Aquatic Biomonitoring Network, field manual—Wadeable streams*. Environment Canada. http://epe.lac-bac.gc.ca/100/201/301/weekly_checklist/2012/internet/w12-31-U-E.html/collections/collection_2012/ec/En84-87-2012-eng.pdf

Carlisle, D. M., Nelson, S. M., & Eng, K. (2014). Macroinvertebrate community condition associated with the severity of streamflow alteration: Macroinvertebrate condition and streamflow alteration. *River Research and Applications*, *30*(1), 29–39. https://doi.org/10.1002/rra.2626

Cywinska, A., & Hebert, P. D. N. (2002). Origins of clonal diversity in the hypervariable asexual ostracode Cypridopsis vidua. *Journal of Evolutionary Biology*, *15*(1), 134–145. https://doi.org/10.1046/j.1420-9101.2002.00362.x

de Queiroz, K., & Gauthier, J. (1994). Toward a phylogenetic system of biological nomenclature. *Trends in Ecology & Evolution*, *9*(1), 27–31. https://doi.org/10.1016/0169-5347(94)90231-3

Elbrecht, V., & Leese, F. (2017). Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, *5*. https://doi.org/10.3389/fenvs.2017.00011

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol*, *3*(5), 7.

Havel, J. E., & Hebert, P. D. N. (1989). Apomictic parthenogenesis and genotypic diversity in Cypridopsis vidua (Ostracoda: Cyprididae). *Heredity*, *62*(3), 383–392. https://doi.org/10.1038/hdy.1989.53

Hebert, P. D., Wilson, C. C., Murdoch, M. H., & Lazar, R. (1991). Demography and ecological impacts of the invading mollusc *Dreissena polymorpha*. *Canadian Journal of Zoology*, *69*(2), 405–409.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, *96*(4), 317–323.

Klymus, K. E., Marshall, N. T., & Stepien, C. A. (2017). Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. *PLOS ONE*, *12*(5), e0177643. https://doi.org/10.1371/journal.pone.0177643

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, *28*(2), 420–430. https://doi.org/10.1111/mec.14920

Leray, M., Knowlton, N., Ho, S.-L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences*, *116*(45), 22651–22656. https://doi.org/10.1073/pnas.1911714116

Moulton, S. R. I., Kennen, J. G., Goldstein, R. M., & Hambrook, J. A. (2002). *Revised Protocols for Sampling Algal, Invertebrate, and Fish Communities as Part of the National Water-Quality Assessment Program* (US Geological Survey Open-File Report No. 02–150; Open-File Report).

Mychek-Londer, J. G., Chaganti, S. R., & Heath, D. D. (2020). Metabarcoding of native and invasive species in stomach contents of Great Lakes fishes. *PLOS ONE*, *15*(8), e0236077. https://doi.org/10.1371/journal.pone.0236077

Nelson, S. M. (2005). *Stream Macroinvertebrate Surveys in the Cle Elum and Bumping River Watersheds* (PN-YDFP-002; Technical Series, p. 51). Bureau of Reclamation.

Nelson, S. M. (2009). *Biological indicators of conditions below dams in the Western United States*. 29th Annual USSD Conference, Nashville, TN.

Nelson, S. M. (2011). Response of stream macroinvertebrate assemblages to erosion control structures in a wastewater dominated urban stream in the southwestern U.S. *Hydrobiologia*, *663*(1), 51–69. https://doi.org/10.1007/s10750-010-0550-y

Nelson, S. M., & Wydowski, R. (2008). *San Diego River invertebrate monitoring program—Final report* (Technical Memorandum No. 86-68220-08–06). Bureau of Reclamation.

Porter, T. M., Gibson, J. F., Shokralla, S., Baird, D. J., Golding, G. B., & Hajibabaei, M. (2014). Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome *c* oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. *Molecular Ecology Resources*, n/a-n/a. https://doi.org/10.1111/1755-0998.12240

Porter, T. M., & Hajibabaei, M. (2018a). Over 2.5 million COI sequences in GenBank and growing. *PLOS ONE*, *13*(9), e0200177. https://doi.org/10.1371/journal.pone.0200177

Porter, T. M., & Hajibabaei, M. (2018b). Automated high throughput animal CO1 metabarcode classification. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-22505-4

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

Stark, J., Boothroyd, I., Harding, J., Maxted, J., & Scarsbrook, M. (2001). *Protocols for sampling macroinvertebrates in wadeable streams* (New Zealand Macroinvertebrate Working Group Report No. 1). Cawthron Institute.

USEPA, U. (2013). *National Rivers and Streams Assessment 2013/14: Field operations manual wadable.* EPA-841-B-12-009b. Office of Water and Office of Environmental Information ….

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Westfall, K. M., Therriault, T. W., & Abbott, C. L. (2020). A new approach to molecular biosurveillance of invasive species using DNA metabarcoding. *Global Change Biology*, *26*(2), 1012–1022. https://doi.org/10.1111/gcb.14886

Appendix A

Commands for QIIME 2 processing of samples from Folsom site MI8 amplifies with BF1 and BR2 based primers:

```
cd ~/Documents/Folsom/BF1/MI8_BF1

qiime tools import \
  --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path MI8_BF1_manifest \
  --output-path MI8_BF1_PE_demux.qza \
  --input-format PairedEndFastqManifestPhred33V2

qiime demux summarize \
  --i-data ./MI8_BF1_PE_demux.qza \
  --o-visualization ./MI8_BF1_PE_demux.qzv

#check MI8_BF1_PE_demux.qzv

qiime tools view MI8_BF1_PE_demux.qzv

#go to interactive plot and note regions of low quality - up to
position 35 on the left and position 249 for the reverse on the
right in this case

qiime dada2 denoise-paired \
  --i-demultiplexed-seqs MI8_BF1_PE_demux.qza \
  --p-trim-left-f 36 \
  --p-trim-left-r 36 \
  --p-trunc-len-f 250 \
  --p-trunc-len-r 249 \
  --p-trunc-q 10 \
  --o-table MI8_BF1_q10_table.qza \
  --o-representative-sequences MI8_BF1_q10_rep-seqs.qza \
  --o-denoising-stats MI8_BF1_q10_denoising-stats.qza

#visualization

qiime feature-table summarize \
  --i-table MI8_BF1_q10_table.qza \
  --o-visualization MI8_BF1_q10_table.qzv \
  --m-sample-metadata-file MI8_BF1_metadata.tsv

qiime tools view MI8_BF1_q10_table.qzv

qiime metadata tabulate \
```

```
  --m-input-file MI8_BF1_q10_denoising-stats.qza \
  --o-visualization MI8_BF1_q10_denoising-stats.qzv

qiime tools view MI8_BF1_q10_denoising-stats.qzv

qiime feature-table tabulate-seqs \
  --i-data MI8_BF1_rep-seqs.qza \
  --o-visualization MI8_BF1_q10_rep-seqs.qzv

qiime tools view MI8_BF1_q10_rep-seqs.qzv

#filtering

qiime feature-table filter-seqs \
    --i-data MI8_BF1_q10_rep-seqs.qza \
    --m-metadata-file MI8_BF1_q10_rep-seqs.qza \
    --p-where 'length(sequence) > 274' \
    --o-filtered-data MI8_BF1_q10_rep-seqs_over_275.qza

qiime feature-table filter-seqs \
    --i-data MI8_BF1_q10_rep-seqs_over_275.qza \
    --m-metadata-file MI8_BF1_q10_rep-seqs_over_275.qza \
    --p-where 'length(sequence) < 285' \
    --o-filtered-data MI8_BF1_q10_rep-seqs_275-285.qza

qiime feature-table tabulate-seqs \
  --i-data MI8_BF1_q10_rep-seqs_275-285.qza \
  --o-visualization MI8_BF1_q10_rep-seqs_275-285.qzv

qiime tools view MI8_BF1_q10_rep-seqs_275-285.qzv
```

Appendix B

Table 8 ASV numbers and feature frequencies with and without filtering for target sequence length

Sample	Forward primer	Reverse primer	ASVs unfiltered	Total frequency unfiltered	ASVs filtered	Total frequency filtered
Folsom MI8	BF1	BR2	82	2834	49	2635
Folsom MI8	BF2	BR2	84	2804	65	2554
Folsom WC1	BF1	BR2	125	2970	49	1619
Folsom WC1	BF2	BR2	74	1420	55	822
Canyon 3	BF1	BR2	51	2824	38	2608
Canyon 3	BF2	BR2	231	13977	174	13055
Mock community	BF1	BR2	45	4064	12	2995
Mock community	BF2	BR2	82	12776	33	11168
Field blank	BF1	BR2	2	11	0	0
Field blank	BF2	BR2	3	340	1	321
Lab blank	BF1	BR2	7	1	0	0
Lab blank	BF2	BR2	0	0	0	0

Appendix C

Table 9 Total ASV numbers and feature frequencies assigned to Metazoa, non-metazoan eukaryotes, and bacteria

Sample	Forward primer	Reverse primer	Metazoa ASVs	Metazoa frequency	Eukaryote ASVs	Eukaryote frequency	Bacteria ASVs	Bacteria frequency
Folsom MI8	BF1	BR2	21	2322	15	155	13	158
Folsom MI8	BF2	BR2	14	1284	36	1174	15	96
Folsom WC1	BF1	BR2	14	1118	11	139	23	362
Folsom WC1	BF2	BR2	12	244	18	264	25	314
Canyon 3	BF1	BR2	24	2169	9	281	4	158
Canyon 3	BF2	BR2	98	7015	42	2193	34	3847
Mock	BF1	BR2	12	2995	0	0	0	0
Mock	BF2	BR2	31	11138	1	27	1	3

Appendix D

Table 10 Total ASV numbers identified as metazoan to the level of phylum, class, order, family, genus, and species with bootstrap values ≥ 0.5

Sample	Phylum ASVs	Class ASVs	Order ASVs	Family ASVs	Genus ASVs	Species ASVs
Folsom MI8	19	14	9	7	6	6
Folsom MI8	10	8	7	3	2	2
Folsom WC1	8	7	4	4	4	4
Folsom WC1	10	10	10	10	10	9
Canyon 3	22	16	13	12	11	11
Canyon 3	81	54	30	29	26	25
Mock	10	9	7	5	5	4
Mock	31	27	24	24	23	20

Table 11 Total feature frequencies of ASVs identified as metazoan to the level of phylum, class, order, family, genus, and species with bootstrap values ≥ 0.5

Sample	Phylum frequencies	Class frequencies	Order frequencies	Family frequencies	Genus frequencies	Species frequencies
Folsom MI8	2309	2045	946	771	728	728
Folsom MI8	1269	1223	1211	998	958	958
Folsom WC1	1086	1079	391	391	391	391
Folsom WC1	170	170	170	170	170	165
Canyon 3	2099	1807	1579	1559	1556	1556
Canyon 3	5721	4453	2317	2312	2204	2199
Mock	1921	1890	1290	1198	1198	1097
Mock	11138	7127	6524	6524	6477	5941