NEON Macroinvertebrate/ Zooplankton Metabarcoding Standard Operating Procedure, v1.1-GMCF

Prepared for: NEON

Prepared By:
Genomics and Microbiome Core Facility (GMCF)
Rush University Medical Center
1750 W. Harrison Street, Jelke 1255
Chicago, IL 60612

Date: August 18th, 2026

I. Version History

This is the first version of the Macroinvertebrate/Zooplankton Metabarcoding SOP to be prepared by the Genomics and Microbiome Core Facility (GMCF; Rush University).

Version	Effective Date	Reason for Revision
1.1	8-18-2026	Updated sequencing platform
1	6-13-2025	First version

II. Objective and Overview

Using the tailed primers described in **Table 2**, Cytochrome c oxidase subunit 1 gene fragments (cox1 or COI or CO1) will be amplified and prepared for sequencing. Genomic DNA will be extracted from field samples provided by the NEON Program and shipped to the Genomics and Microbiome Core Facility (GMCF). Multiple primer sets will be deployed, of which some have been previously used for macroinvertebrate and/or zooplankton characterization, and some novel primer pairs will be tested for efficacy by the GMCF. Previously established primer sets to be deployed in this SOP include: (a) F230 fragment of CO1 gene (Folmer et al. 1994; Gibson et al. 2015); (b) BR5 amplicon (Hajibabaei et al. 2012; Gibson et al. 2014); and (c) MLJG amplicon (Leray et al. 2013; Geller et al. 2013).

Environmental samples will be sent from NEON to the GMCF, and these samples will be homogenized and prepared for genomic DNA extraction. DNA extraction protocols will include simultaneous processing with multiple negative extraction controls (extraction blanks containing only extraction kit reagents). Subsequently, all samples will be processed for 1st and 2nd stage PCR, together with DNA extraction controls, PCR reagent blanks, and genomic DNA positive controls. After CO1 gene fragments are amplified using polymerase chain reaction (PCR; 1st stage), the PCR products are used as templates for a second amplification reaction using Illumina indexing primers (xGen™ Amplicon UDI Primers from Integrated DNA Technologies, IDT). All amplicons are then pooled in equal volume and purified using SPRI beads. A low-output quality control sequencing run is performed on the pool using an Illumina MiSeq i100+ sequencer (or equivalent low throughput sequencer). The original barcoded libraries will be re-pooled in variable volumes based on the output numbers of clusters. Purification of the pool will again be performed using SPRI beads. The final pool will be sequenced on an Illumina MiSeq i100+ sequencer employing 2x300 with a 50M cluster flow cell or similar. Alternative sequencing platforms include an Illumina NovaSeq X 1.5B flow cell with 600 cycle chemistry (i.e., 2x300). If necessary, longer read lengths (i.e., 2x500) are available on the Illumina MiSeq i100+. A schematic of this sample processing workflow is shown in **Figure 1** (sample processing and DNA extraction) and **Figure 2** (PCR amplification and sequencing).

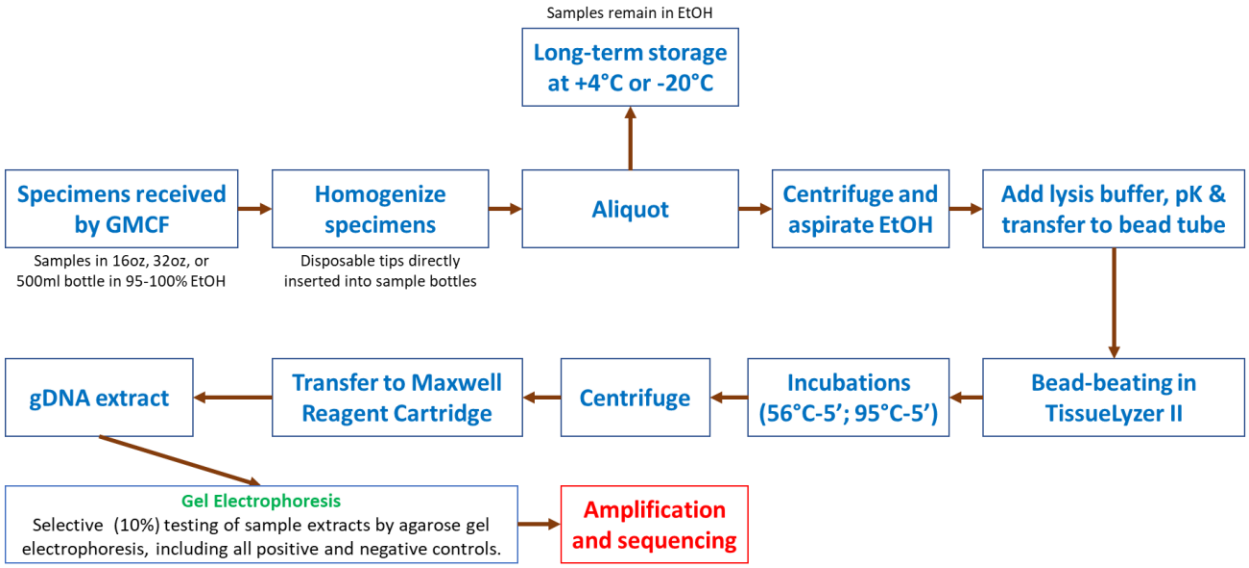

Figure 1: NEON macroinvertebrate and zooplankton DNA extraction workflow.

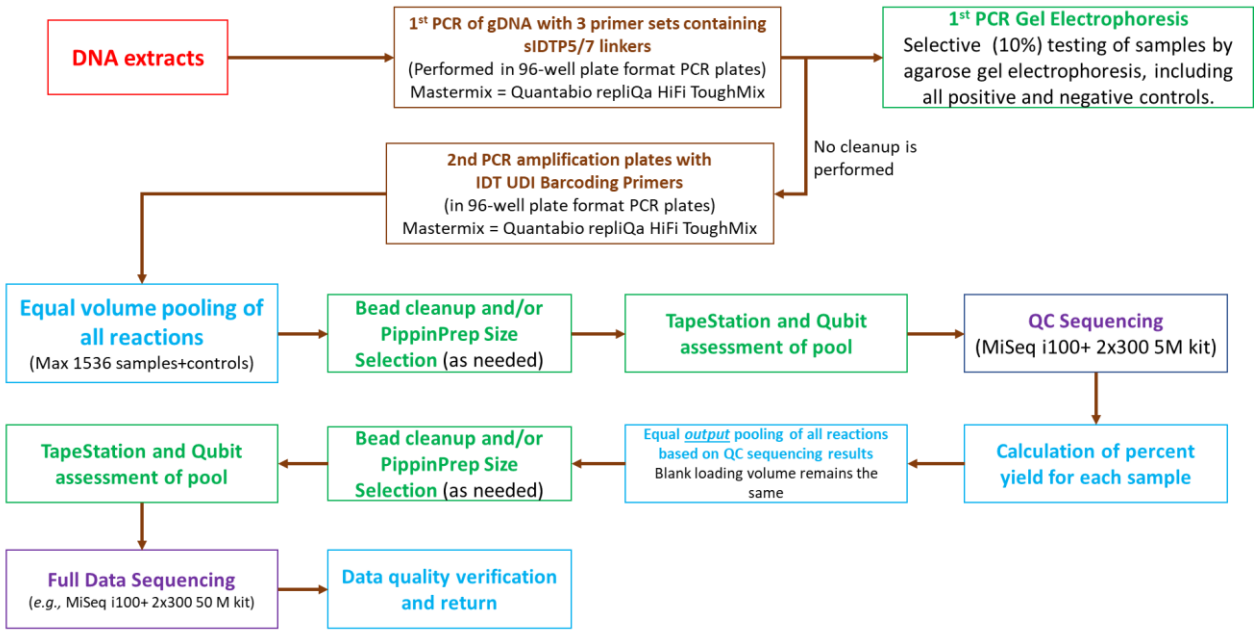

Figure 2: NEON macroinvertebrate and zooplankton CO1 marker gene sequencing library prep workflow. Red = nucleic acid extracts; brown = PCR; Green = cleaning and QC; Purple = Sequencing; Blue = bioinformatics and pooling.

III. Materials

Table 1. Materials and Sources

Material	Manufacturer	Catalog Number
Sample processing		
Disposable homogenizer tips, 7x110 mm plastic	Fisher Scientific	15340176
Sample storage tubes, Matrix 2D, 12ml or similar	ThermoFisher	3775
DNA extraction		
Maxwell RSC Fecal Microbiome DNA Kit	Promega	AS1700
ZR BashingBead Lysis Tubes (0.1 & 0.5 mm)	Zymo	S6012-50
E-Gel 48 SYBR SAFE (1%)	ThermoFisher	G820801
Sample storage tubes, Matrix 2D, 1 ml or similar	ThermoFisher	3741-BR
PCR amplification, QC and Cleanup		
repliQa HIFi ToughMix	QuantaBio	95200-500
KAPA HyperPure Beads	Roche	8963878001
Qubit dsDNA HS Assay Kit	ThermoFisher	Q32851
E-Gel 48 SYBR SAFE (2%)	ThermoFisher	G820842
TapeStation D1000 Screen Tape	Agilent	5067-5582/3
TapeStation HS D1000 Screen Tape	Agilent	5067-5584/5
TapeStation HS D5000 Screen Tape	Agilent	5067-5592/3
Blue PippinPrep Gels (1.5% agarose)	Sage Scientific	BDF1510
NEON-supplied community standard	NDCS	
Primers, Lab-ready, standard	IDT	Custom
Nuclease-free water	IDT	11-04-02-01
xGen Amplicon UDI Primers	IDT	10009846, 10009851, 10009852, 10009853
Sequencing		
KAPA Library Quantification Kit (Illumina)	Roche	KK4923
MiSeq i100+, 5M 600 cycle kit	Illumina	20126566
MiSeq i100+, 25M 600 cycle kit	Illumina	20115696
MiSeq i100+, 50M 600 cycle kit	Illumina	20141597
MiSeq i100+, 25M 1000 cycle kit	Illumina	20148254
NovaSeq X, 1.5B 600 cycle kit	Illumina	20145964

IV. Procedure

A. Sample Requirements

Upon receipt of samples, staff members will perform a sample audit to verify that all samples are present and correspond to names present in NEON sample submission forms. Staff members will fill out and upload shipping receipt forms according to NEON instructions. If samples are received in a compromised condition, GMCF will alert Battelle to problems with the shipment. An electronic receipt form emailed with the shipping information will be completed to document the condition of samples upon receipt. Each completed receipt form will be uploaded to the NEON Data Portal per instructions provided in **Attachment 2a**.

Unprocessed samples (stored in ethanol, EtOH) will be maintained at 4°C until homogenization. After homogenization, ~10 ml of ethanol-sample homogenate will be aliquoted to 2D-matrix barcoded sample tubes and will be retained at the GMCF until samples results are reviewed and approved by NEON. After approval by NEON, remaining homogenate will either be returned to NEON or disposed. In addition, a 1-ml aliquot of homogenate will be transferred to a microfuge tube, and centrifuged at maximum g (>15,000g), and ethanol decanted. These ethanol-decanted samples will be frozen and eventually processed for DNA extraction as described below. Approximately 150 to 250 mg of homogenized material is suggested for extractions, but samples are expected to have high levels of total material. After extraction, DNA concentrations of each sample will be quantified with fluorimetry. A limited selection of samples, including negative and positive controls, will be evaluated by gel electrophoresis (**Figure 1**). Subsequently, genomic DNA samples will be aliquoted into 96-well plates for PCR amplification, and remaining DNA will be stored at -80°C. After successful PCR amplification and sequencing of samples, the original large volume of sample will be disposed of. The residual 10 ml of ethanol-sample homogenate will be maintained at -20°C for at least 180 days and returned to NEON upon request. DNA extracts will be frozen (-20°C) for 30 days and then stored at -80°C indefinitely or returned to NEON upon request.

Extracted genomic DNA will be processed using a two-stage PCR amplification protocol to generate sequencer-ready libraries. No specific minimum input requirement is indicated by the protocol, but DNA extracts with concentrations below 1 ng per microliter may create difficulties due to low input DNA levels. There is no minimum number of reactions or maximum number of controls as the approach is highly flexible. A marker gene sequencing run consists of up to 1536 possible samples and controls (i.e., 1536 unique barcodes), though generally many fewer samples are processed simultaneously. In general, a simple formula of the number of samples available for processing x twice the depth of sequencing desired per sample is used to select the appropriate depth sequencing flow cell. The GMCF will sequence all DNA extraction blanks generated during sample processing. In addition, a minimum of five PCR reagent blanks will be amplified and sequenced for each lane of sequencing performed. GMCF will also perform a minimum of three technical replicates of a positive control provided by NEON (‘NEON DNA Community Standard’, NDCS) per sequencing run. Upon receipt of the positive control received from NEON, GMCF staff will aliquot the material into single use aliquots and store at -80°C. For each library preparation setup, a single tube will be removed from the freezer and thawed and used as template for PCR amplification. This tube will be discarded after use. Each aliquoted

tube will contain sufficient volume and DNA concentration for 15 reactions (i.e., three technical replicates x three primer sets + additional volume for pipetting losses). NEON and GMCF may add additional standards in future runs, but NDCS standards will continue to be processed for all sequencing runs.

From each 96-well plate of PCR reactions, a single column (eight samples out of 96) will be selected for PCR amplification assessment after the 1st stage PCR using agarose gel electrophoresis. Analysis of these gels will evaluate the proper amplification of the target region and will be used to determine if there are samples that are PCR amplifying poorly or not at all. If more than 10% of samples show no amplification, GMCF will notify NEON to discuss whether re-extraction is necessary. Assessment of PCR amplification across all samples and controls will be performed using a 'quality control' (QC) sequencing run on a MiSeq i100+ sequencer (5M clusters, 600 cycle chemistry). Prior to deep sequencing runs, NEON will be requested to approve based on results from the QC sequencing run.

B. **Sample processing and genomic DNA extraction**

Received samples will be stored in 95% ethanol. These samples will initially be homogenized while still in the original container and in the ethanol using a Fisherbrand 850 Homogenizer (catalog #15340169) with motor stand assembly (catalog #15340172). Homogenization will be performed with disposable generator probes (**Table 1**). Homogenization will be performed in four sequential pulses of 30 seconds for a total of 120 seconds cumulative homogenization time. Single use tips will be disposed of once homogenization is complete. If there is excessive plant material in samples which prevents the proper homogenization of the sample, the GMCF staff will contact NEON for instructions.

Aliquots of the homogenized material will be taken for storage (~10 ml aliquot) and for immediate DNA extraction (1 ml aliquot). The ~10 ml aliquot will be stored at +4°C or -20°C and will be stored at the GMCF until completion of project and returned or disposed according to NEON instructions. The remaining large volumes of material will be disposed of after sequence data has been generated. The 1 ml aliquot will be centrifuged at a minimum of 15,000g for 10 minutes to pellet material. The ethanol supernatant will be decanted, and the remaining biomass will be frozen at -20°C. The biomass will be thawed prior to genomic DNA extraction processing.

Genomic DNA will be extracted from a maximum of 250 mg of the residual pellet using a Maxwell RSC Fecal Microbiome DNA kit (Promega), with much of the workflow implemented on a Maxwell RSC48 device. Briefly, 1ml of lysis buffer will be added to the pellet or portion thereof, and the pellet will be resuspended. The buffer and material will be transferred to a bead-beating tube containing 0.1- and 0.5-mm glass beads (Zymo; **Table 1**) as well as 40 µL Proteinase K (provided by the Maxwell kit). Samples in bead tubes with lysis buffer and proteinase K will be placed in a TissueLyser II device (Qiagen) and bead-beaten. Bead-beating will be performed for 2 rounds of 5 minutes @30Hz, with a 2-minute pause in between rounds. After bead-beating, tubes will be incubated in a heat-block at 56°C for 5 minutes and then incubated in a separate heat-block at 95°C for 5 minutes. Tubes will be cooled to room

temperature and then centrifuged at maximum g (>15,000g) for 5 minutes. Maxwell RSC reagent strips will be prepared for sample processing by: (a) adding 300 µL of Binding Buffer to well #1 of the reagent cartridge, (b) adding 20 µL RNase A to well #3 of the reagent cartridge, and (c) transferring 300 µL sample supernatant into well #1 of the reagent cartridge. The Maxwell RSC device will finish purification of genomic DNA from the sample, and gDNA will be eluted in 100 µL of elution buffer. At least 1 extraction control (i.e., lysis buffer in a tube without a pellet) will be processed with each set of samples processed on a Maxwell RSC48 run. A selection of samples, including negative and NEON-provided positive controls, will be checked by gel electrophoresis. All DNA extracts will be quantified using Qubit fluorimetry. DNA extracts are anticipated to have concentrations above 1 ng/µL and NTCs are expected to have concentrations well below 1 ng/µL. When negative control extractions yield blank DNA concentrations greater than 1 ng/µL, this is an immediate issue requiring discussion with NEON. Samples with measurable DNA concentrations below 1 ng may still be viable, but NEON should be notified which samples fail to meet the 1 ng/µL criterion. DNA extracts will be stored indefinitely at -80°C in 2D barcoded tubes (LN2-safe) and returned to the NEON Biorepository at ASU with dry ice shipment upon communication. Shipments will include a hard-copy shipping manifest listing all samples in the shipment following the instructions in Attachment 3a for completing the template in Attachment 3b. At the time of shipment, GMCF will submit an electronic version of the shipping manifest to the archive facility and others as described in Attachment 3a.

C. **First Stage Amplification**

Library preparation processing largely follows the two-stage PCR amplification workflow described in Naqib et al. 2018, with PCR conditions modified from those previously proposed. Primer sequences are shown in **Table 2**. PCR conditions for first stage PCRs are shown in **Table 3**. PCR conditions for second stage PCRs are also shown in **Table 3**. The first stage PCR amplifies portions of the CO1 genes using QuantaBio repliQa HiFi ToughMix mastermix. All PCR prep work is conducted in AirClean® Systems AC600 Series PCR Workstations with ISO 5 HEPA-filtered air. Prior to work, the workstation is decontaminated by wiping all surfaces with 10% bleach followed by 70% ethanol. A germicidal UV light is turned on for a minimum of 10 minutes. The PCR master mix is prepared according to the manufacturer's instructions using the primers in **Table 2**; final concentrations of each primer is 300 micromolar. Reaction volumes are 10 microliters, with 4 microliters of DNA for each sample. Gel electrophoresis will be performed to evaluate amplification of positive and negative control samples and a limited selection (~10%) of samples. If the DNA extraction control (NEC) produces positive PCR results, GMCF will troubleshoot accordingly (e.g., identify sources of contamination, re-extract or re-PCR as necessary). Similarly, if PCR negative controls (NTCs) produce positive PCR results that are of the proper expected amplicon size, troubleshoot source of contamination and re-run samples as necessary. If the NDCS templates do not amplify or produce only primer-dimers or similar non-specific, off-target amplification products, new aliquots should be selected and re-amplified to troubleshoot.

Table 2. 1st stage PCR primers for CO1 gene amplification

Primer name	Oligonucleotide sequencing (5'-3')	Tm (°C)	Published Amplicon Size (bp)	Observed amplicon size (bp; with linkers)	Original Annealing Temp (°C) **	Total Cycles	Citation
F230 amplicon							
LCO1490	CTACACGACGCTCTTCCGATCTGGTCAACAAATCATAAAGATATTGG	59.2	239	~300	46°C	35	Folmer et al. 1994
230_R	CAGACGTGTGCTCTTCCGATCTCTTATRTTRTTTATICGIGGRAAIGC	62.6-67.7					Gibson et al. 2015
BR5 (BE) amplicon							
BR5	CTACACGACGCTCTTCCGATCTCCIGAYATRGCITTYCCICG	65.1-71.9	224-310	~350-400	46°C	35	Hajibabaei et al. 2012
R5_R	CAGACGTGTGCTCTTCCGATCTGTRATIGCICCIGCIARIACIGG	72.8-76.1					Gibson et al. 2014
MLJG amplicon							
mlCOIintF	CTACACGACGCTCTTCCGATCTGGWACWGGWTGAACWGTWTAYCCYCC	63.8-68.7	313	~400	62-46°C (1°/cycle touchdown)	41	Leray et al. 2013
jgHCO2198	CAGACGTGTGCTCTTCCGATCTTAIACYTCIGGRTGICCRAARAAYCA	66.9-73.5					Geller et al. 2013

Bold indicates genomic DNA target region of the primers while underlining indicates IDT linkers. Source: Walters et al., 2015. Melting temperature was calculated using IDT's OligoAnalyzer, using 300 nM primers, 50 mM Na^+, 2 mM Mg^{2+} and 0.2 mM dNTPs. Tm is shown only for the target-specific portion of the primer, not including the linkers. ** With linkers, an annealing temperature of 50°C without touchdown has been successfully employed for all primer sets.

Table 3. Thermocycler conditions

PCR stage	Temperature	Duration	Number of cycles
F230, BR5 (BE) and MLJG amplicons			
Initial denaturation	98°C	2 min	1
Cycling-denaturation	98°C	10 s	30
Cycling-annealing	50°C*	1 s	
Cycling-elongation	68°C	1 s	
Hold	4°C	Hold	1
* PCR conditions are altered from original manuscripts due to the incorporation of linker sequencers in all primers and a different PCR mastermix. A 50°C annealing temperature was successfully evaluated for all primer sets and differs from annealing temperatures initially reported for amplicons.			
Second Stage PCR (all amplicons)			
Initial denaturation	98°C	2 min	1
Cycling-denaturation	98°C	10 s	8
Cycling-annealing	60°C	1 s	
Cycling-elongation	68°C	1 s	
Hold	4°C	Hold	1

D. Second Stage PCR

The purpose of the second stage PCR is to attach dual indices (barcodes) and Illumina sequencing adapters into amplicons from each sample so they can be loaded together on Illumina sequencers. After completion of the first stage PCR, the second stage of PCR processing follows the workflow described in Naqib et al. 2018 but employs the xGen™ Amplicon UDI Primers from IDT. Illumina sequencers with patterned flow cells, including the MiSeq i100+, are subject to 'index hopping' – an unfortunate process that leads to incorrect assignment of sequences on a small percentage of reads. By incorporating unique dual indices (UDIs), mis-assigned sequences

are removed from datasets. IDT has 16 96-well plates of UDI primers for a total of 1536 indices. The second stage PCRs amplify the products of the first stage PCR by targeting the linker sequences of the amplified products; genomic DNA is no longer targeted. Reactions are performed using the same mastermix as for the first stage PCR, Quanta repliQa HiFi ToughMix. All PCR prep work is conducted in AirClean® Systems AC600 Series PCR Workstations with ISO 5 HEPA-filtered air. Prior to work, the workstation is decontaminated by wiping all surfaces with 10% bleach followed by 70% ethanol. A germicidal UV light is turned on for a minimum of 10 minutes. Reaction volumes are 10 microliters, with 1 microliter of PCR product from the 1st reaction used as input template for each sample. Two microliters of IDT UDI primers are used for each reaction; each well receives a unique primer pair from the xGen™ Amplicon UDI Illumina primer plates. The thermocycler is run using the conditions in **Table 3**.

E. Pooling of Libraries for 'Quality Control' sequencing run

In this SOP, PCR amplicons from individual reactions are not purified. Rather, a small equal volume of each sample is pooled and purified for the QC run. A low output (*e.g.*, MiSeq i100+ 5 M cluster flow cell) sequencing run is used to measure the relative abundance of amplicons from each sample and to allow for accurate re-pooling for the final sequencing run. At least 4 PCR NTCs should be run per run, with the expectation of generating fewer than 3,000 merged and quality filtered, primer-trimmed, and length-filtered reads (i.e., negative controls can sometimes generate large numbers of reads that are primer dimers and are not actual contamination). A maximum of 1,536 amplicons can be multiplexed per run, but no more than 384 amplicons are recommended for sequencing per 50 M cluster flow cell to ensure depth of sequencing is adequate for minimum requirements. Many more amplicons can be sequenced concurrently on a NovaSeq X 1.5 cluster flow cell (2x300), though turn-around time will be longer.

After completion of the second stage PCR, amplification products of each sample are pooled in equal volume using a multi-channel pipettor or Opentrons OT-2 Lab Robot. Accurate pipetting during pooling is important. The pool of amplicons is purified twice sequentially, using KAPA HyperPure beads according to the manufacturer's instructions, with a 0.6X ratio to remove fragments shorter than 300 bp. The DNA fragments in the pooled libraries are analyzed using an Agilent TapeStation device, with PCR products (after 2nd PCR) expected to be approximately 350 - 400 bp. If small fragments (< 150 bp) are detected in the pool, the pool will undergo additional SPRI bead cleanups. The pool will then be analyzed using a Qubit (**Table 1**) prior to loading the QC sequencing run (below).

Running the QC Sequencing Run. The pooled libraries will be loaded at 100 pM concentration onto a small (5M cluster) MiSeq i100+ sequencing run (600 cycle), according to the manufacturer's instructions. No thawing of reagents is required as MiSeq i100+ reagents are stored at room temperature. A 10% phiX spike-in will be provided for all sequencing runs. The sequencing run is set up following the prompts on the instrument. Custom sequencing primers are not needed when using the xGen™ Amplicon UDI Primers.

F. Assessing QC sequencing run yield and re-pooling of samples

Following the QC sequencing run, the number of clusters for each sample will be used to calculate a percent of pass-filter clusters. The percentage of each library will be used to calculate a new volume to pool in order to balance the reads per sample. Since an equal volume of each sample was pooled for the QC run, a simple formula is used in Excel to calculate a new volume based on the percentage of clusters for each sample. Briefly, the volume of each sample used for the QC sequencing run is multiplied by the desired relative abundance and divided by the measured relative abundance of each sample as calculated as a portion of pass-filter, barcoded reads.

Re-pooling and final sequencing. As described above, samples are re-pooled into independent pools based on QC sequence data. Once the needed volumes for normalization have been calculated, an Excel table is used to re-pool the libraries using a robotic liquid handler (e.g., Opentrons OT-2). A liquid handling robotic instrument is highly desirable as each sample will require a different volume; manual pipetting for these many samples is difficult. Each pool of amplicons is purified twice sequentially as previously performed for the original pool, using SPRI beads, with a 0.6X ratio to remove fragments shorter than 300 bp. The pools are then analyzed using Qubit to measure library concentration. The final re-pooled library will be sequenced again on a MiSeq i100+ instrument with a larger flow cell as appropriate for the number of samples processed. For 120 samples expected annually, a 50M cluster flow cell will be employed.

V. Sequence Run Quality Review

Run-level data acceptance criteria include both primary and secondary metrics. The primary metrics refer to sequencing run quality (cluster density, Q30, total clusters, and percent Q30 data). The secondary metrics refer to processed data to identify viable sequence data for analysis, and include, on a per-sample basis: merged reads, primer trimmed reads, quality trimmed reads, post-chimera removal reads, macroinvertebrate/zooplankton-annotated reads. The metric for sequencing success is the number of reads per sample that merge properly using the software package PEAR (Zhang et al. 2014), that contain both forward and reverse primers in the proper orientation (*i.e.*, forward primer in the forward orientation and reverse primer as an inverse complement of the primer sequence), that pass size filtering, and that remain after quality trimming and chimera removal. Once the data are determined to meet the minimum criteria from Table 6, data will be evaluated using the open-source free software, FastQC. Biologically, the number of reads that can be annotated as phylum Arthropoda or Rotifera are the most valuable.

Two primary quality metrics are monitored during the run (with their associated acceptance criteria in parentheses):

1) Q30, the percentage of sequenced bases with Phred-equivalent quality scores of at least 30 (greater than 70%)
2) Percent phiX aligned (~10% of the phiX spike-in)

Deviations from these metrics will cause the run to be flagged and reported to NEON and may require reprocessing the sequencing run. However, it may be possible for a successful run to be achieved even if some of these criteria are not met. For example, the Q30 metric can vary from run to run, depending on the length of amplicons. Similarly, lower cluster density is likely to yield high quality data, but the total number of reads will be less than desired. Conversely, higher cluster density may still yield an overall acceptable run, but average quality is likely to be lower.

Secondary metrics are performed on a per-sample basis. Per sample-level data acceptance criteria: Each sample is expected to produce at least 25,000 clusters after forward and reverse read merging, primer trimming, and quality trimming. We anticipate that >90% of samples will produce at least 25,000 clusters. In general, despite the best efforts of library QC sequencing and re-pooling, some samples still fail to generate enough data, likely as a result of poor amplification due to low DNA, presence of inhibitors, or both. These metrics can be calculated from the standard bioinformatics pipeline using the software package PEAR for read merging, and CLC genomics workbench (or similar) for primer trimming and quality trimming. Any samples that do not meet these criteria will be flagged by the GMCF in the qaqcStatus field of the sequencing data table as 'Fail'.

Overall run success will also be evaluated by examining: (a) amount of data generated for negative controls (DNA extraction controls and PCR amplification controls will be assessed independently; ideally will be <1% of the sample average or of the positive control samples) and (b) amount of data generated for positive control samples (NDCS). We will also assess the distribution of clusters generated for each sample, by sequencing run. After NEON has performed annotation of the sequence data, the GMCF will calculate the number of samples, per run, that do not meet the minimum criteria of 25,000 merged sequences after read-merging, primer trimming, quality trimming, and annotation and will verify that the positive control yields >25,000 sequences. Technical reproducibility will be evaluated for select samples that have been sequenced multiple times with unique barcodes. These "overall run" results will be evaluated with NEON to determine whether sample processing needs to be repeated. GMCF will maintain a historical list of run metrics, including: (a) loading density, (b) pass filter rate, (c) percent of bases with >Q30; (d) total number of PF clusters; (e) total number of samples; and (f) total number of samples with minimum 25,000 sequences after read-merging, primer trimming, quality trimming and annotation (data coming from NEON).

VI. Data Return

For each sequencing run, sequence data will be returned to NEON according to requirements provided, including data ingest files for sequencing, PCR amplification and raw data files. Raw sequence data, as demultiplexed FASTQ files, will be deposited in a designated cloud-based collaboration platform (e.g., SharePoint, BOX, etc) folder. Raw sequence data, as multiplexed files, will be uploaded as archive sequences files to a separate designated cloud-based collaboration platform or other location specified by NEON if needed due to file size. Procedures

for data submission and ingest file submission will be performed according to document "Uploading Data to the NEON Data Portal NEON INV/ZOO DNA Metabarcoding."

Return extraction metadata for each sample by completing the data return ingest attachment 2c ('inv_dnaEtraction_in) and uploading to the NEON data portal. Data for both macroinvertebrate and zooplankton samples can be returned on the same ingest. The 'dnaSampleID' in the metadata table is created by the lab using the parent sampleID provided by NEON with the "-DNA1" suffix if the first extraction proceeds to analysis and the "-DNA2" suffix if the sample required a second extraction due to low DNA concentration. The dnaExtraction ingest table must be uploaded to the NEON data portal prior to loading any downstream data from PCR amplification and sequencing.

In addition, metadata describing the PCR, sequencing, and raw data files will be uploaded to NEON's data portal. Data for macroinvertebrate and zooplankton samples can be returned in the same data file. Instructions for uploading data files are provided in Attachment 2a. Descriptions for each data field and acceptable choices for data fields where a defined list of values must be used for the field are provided in Attachment 2b. Templates for the data return are provided in the attachments noted in Table 8 of the NEON SOW. Raw sequence files (both demultiplexed and multiplexed) and fastQC (or equivalent) sequence quality reports will be uploaded to a Google Cloud Platform (GCP) bucket that will be set up by Battelle. Contractor must notify the NEON team by email as each deliverable is returned to the NEON data portal or GCP.

Data ingest template	Ingest table names	Description	Frequency	Upload location
Attachment 2d	inv_pcrAmplification_in	Metadata and quality metrics for the PCR amplification	Per sample extraction + target subfragment (BE or F230)	NEON data portal
Attachment 2e	inv_markerGeneSequencing_in	Metadata for sequencing	Per sample extraction + target subfragment	NEON data portal
Attachment 2f	inv_rawDataFiles_in	List of samples and associated raw data files that were loaded to the GCP folder (Section 2.5.2). This ingest includes GCP file locations that make the sequence data discoverable	Per sample extraction + target subfragment	NEON data portal
NA	NA	Demultiplexed, compressed raw data files with file names matching the name entries in inv_rawDataFiles_in - rawDataFileName	Per dnaSampleID + target subfragment	GCP
NA	NA	Multiplexed raw data files as returned by the sequencer with file names matching the name entries in inv_rawDataFiles_in - archiveDataFileName	Per sequencing run	GCP
NA	NA	fastQC (or equivalent) sequencing quality report	Per sequencing run	GCP

VI. References

Folmer, O., Black, M., Hoeh, W., Lutz, R. and Vrijenhoek, A.R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotechnol. Mol Mar Biol Biotechnol, 3(5), pp.294-9.

Gardes M, Bruns T (1993) ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. Mol Ecol 2: 113–118.

Geller, J., Meyer, C., Parker, M. and Hawk, H., 2013. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular ecology resources, 13(5), pp.851-861.

Gibson, J.F., Shokralla, S., Curry, C., Baird, D.J., Monk, W.A., King, I. and Hajibabaei, M., 2015. Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. PloS one, 10(10), p.e0138432.

Gibson, J., Shokralla, S., Porter, T.M., King, I., Van Konynenburg, S., Janzen, D.H., Hallwachs, W. and Hajibabaei, M., 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proceedings of the National Academy of Sciences, 111(22), pp.8007-8012.

Hajibabaei, M., Spall, J.L., Shokralla, S. and van Konynenburg, S., 2012. Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. BMC ecology, 12, pp.1-10.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. and Machida, R.J., 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in zoology, 10, pp.1-14.

Naqib, A., Poggi, S., Wang, W., Hyde, M., Kunstman, K. and Green, S.J., 2018. Making and sequencing heavily multiplexed, high-throughput 16S ribosomal RNA gene amplicon libraries using a flexible, two-stage PCR protocol. In Gene expression analysis (pp. 149-169). Humana Press, New York, NY.

Parada, A.E., Needham, D.M. and Fuhrman, J.A., 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environmental microbiology, 18(5), pp.1403-1414.

Smith, Dylan P., and Kabir G. Peay. 2014. “Sequence Depth, Not PCR Replication, Improves Ecological Inference from Next Generation DNA Sequencing.” PLOS ONE 9 (2): e90234.

Walters, William, Embriette R. Hyde, Donna Berg-Lyons, Gail Ackermann, Greg Humphrey, Alma Parada, Jack A. Gilbert, Janet K. Jansson, J. Gregory Caporaso, and Jed A. Fuhrman. 2016. “Improved Bacterial 16S RRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys.” *Msystems* 1 (1): e00009–15.

White TJ, Bruns TD, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gefland DH, Sninsky JJ, White TJ, editors. PCR protocols: a guide to method and applications. San Diego, Academic Press. pp. 315–322.

Zhang, J., Kobert, K., Flouri, T. and Stamatakis, A., 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics, 30(5), pp.614-620.

Zhang, W., Fan, X., Shi, H., Li, J., Zhang, M., Zhao, J. and Su, X., 2023. Comprehensive Assessment of 16S rRNA Gene Amplicon Sequencing for Microbiome Profiling across Multiple Habitats. Microbiology Spectrum, pp.e00563-23.