NEON ITS Marker Gene Sequencing Standard Operating Procedure, v.2.1-GMCF

Prepared for:
NEON

Prepared By:
Genomics and Microbiome Core Facility (GMCF)
Rush University Medical Center
1750 W. Harrison, Jelke 444
Chicago, IL 60612

Date: August 29, 2024

I. Version History

Version	Effective Date	Reason for Revision
2.1	7/12/2024	Updates to clarify pooling strategy and adjust positive control strategy.
2a	9/22/2023	Changes from the first version include shifting of the PCR amplicon barcoding from Fluidigm Access Array for Illumina to IDT xGen™ Amplicon UDI Primers and shifting from MiSeq to NovaSeq sequencers for final data generation.
1		First version

II. Objective and Overview

Using the tailed primers described in Tables 1 and 2, fungal internal transcribed spacer (ITS) region 1 (ITS1) fragments are amplified and prepared for sequencing. gDNA extracts of soil and aquatic field samples will be provided by the NEON Program and shipped to the Genomics and Microbiome Core Facility (GMCF). These primers (ITS1F and ITS2R), initially described by Gardes and Bruns (1993) and White et al. (1990), were used initially for next-generation amplicon sequencing with Roche 454 pyrosequencing (Smith and Peay, 2014), but have been updated for

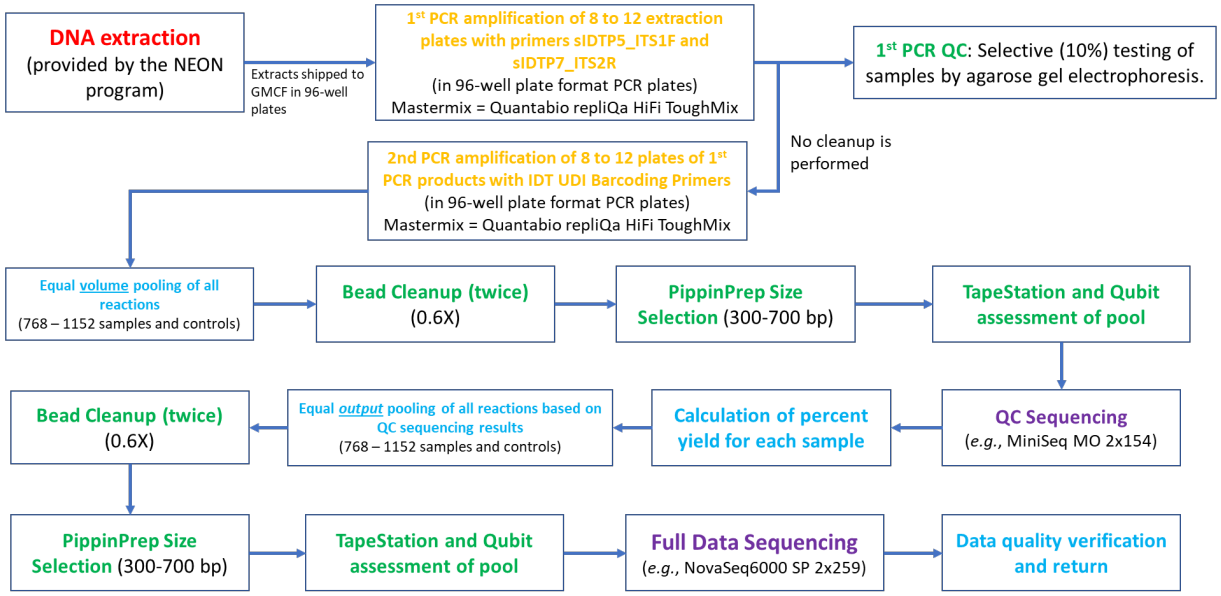

Figure 1: NEON marker gene sequencing library prep workflow. Red = nucleic acid extracts; Yellow = PCR; Green = cleaning and QC; Purple = sequencing; Blue = bioinformatics and pooling.

Illumina sequencers in this SOP. DNA extracts will be prepared simultaneously with negative extraction controls (NEC, extraction blanks containing extraction kit reagents) and PCR reagent blanks. After ITS regions are amplified using polymerase chain reaction (PCR), 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 and PippinPrep size selection. A low-output quality control sequencing run is performed on the pool using an Illumina MiniSeq (mid-output) or MiSeq (Nano) sequencer. Based on the output – numbers of clusters and the percentage of reads that are fungi for each sample – the original barcoded libraries are re-pooled in variable volumes. Purification of the pool is again performed using SPRI beads and PippinPrep size selection. The final pool is sequenced on an Illumina NovaSeq6000 (SP flow cell with 2x259 base sequencing). A schematic of this library prep workflow is shown in Figure 1.

III. Recommended Materials

Table 1. Recommended materials and sources.

Material	Manufacturer	Catalog #
repliQa HiFi ToughMix	QuantaBio	95200-500
KAPA HyperPure Beads kit (4X60ml)	Roche	8963878001
Qubit dsDNA HS Assay Kit	ThermoFisher	Q32851
E-GEL 48 SYBR SAFE 2, 4X8 GEL	Thermo	G820842
High Sensitivity D5000 ScreenTape/Reagent	Agilent	5067-5592/3
High Sensitivity D1000 ScreenTape	Agilent	5067-5584/5
D1000 ScreenTape	Agilent	5067-5582/3
Blue PippinPrep 1.5% agarose gels	Sage Scientific	BDF1510
KAPA Library Quantification Kits for Illumina platforms	Roche	KK4923
MiSeq Reagent Nano Kit v2 (500 cycle)	Illumina	MS-103-1003
PhiX v3	Illumina	15017397
NovaSeq 6000 SP Reagent Kit v1.5 (500 cycle)	Illumina	20028402
MiniSeq Mid Output kit (300 cycle)	Illumina	FC-420-1004
ZymoBIOMICS Microbial Community DNA Standard	Zymo	D6306
sIDTP5_ITS1F Primer	IDT	Custom
sIDTP7_ITS2R Primer	IDT	Custom
Nuclease-free water	IDT	11-04-02-01 (2ml) 11-05-1-14 (300 ml)
xGen™ Amplicon UDI Primers	IDT	10009846, 10009851, 10009852, 10009853

IV. Procedure

A. Sample Requirements

A marker gene sequencing run consists of up to 1536 possible samples and controls (*i.e.*, 1536 unique barcodes). There is no minimum number of reactions or maximum number of controls as the approach is highly flexible. However, for best outcomes, soil samples and aquatic samples will be analyzed in independent sequencing runs. The GMCF will sequence all DNA extraction blanks provided by NEON. 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 four technical replicates of a positive control per sequencing run if positive controls are provided by NEON. Standards such as the ZymoBIOMICS Microbial Community DNA Standard have not amplified well with fungal ITS primers employed in this work. NEON may add additional standards in future runs. Furthermore, eight 'true' samples will be selected for technical replication assessment. Briefly, these eight samples will be PCR amplified independently three times and each replicate will be barcoded with a unique barcode. Finally, eight composite samples will be generated from earlier NEON samples and amplified and sequenced with unique barcodes. These eight composite samples will be independently amplified for each next-generation sequencing run to allow for assessment of run-to-run variability.

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. Samples will be stored at -80°C until PCR amplification. DNA concentrations of samples will not be verified prior to amplification. 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-amplification is necessary. Assessment of PCR amplification across all samples will be performed using the 'quality control' (QC) sequencing run on MiniSeq or MiSeq sequencers. Prior to sequencing each NovaSeq6000 lane, NEON will be requested to approve proceeding with deep sequencing based on results from the QC sequencing run.

B. 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 proposed by Walters et al., 2015 (Table 3) and using primers described by the Earth Microbiome Project (Table 2; https://earthmicrobiome.org/protocols-and-standards/its/). The first stage PCR amplifies the ITS1 target region 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 concentration of each primer is 300 micromolar. Reaction volumes are 10 microliters, with 1 microliter of DNA for each sample. The thermocycler is run using the conditions in Table 3.

Table 2. Primers to be used in first stage PCR amplification. 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.

Target gene	Primer	Oligonucleotide Sequence (5'-3')	Tm (°C)
ITS	sIDTP5_ ITS1f	CTACACGACGCTCTTCCGATCT **CTTGGTCATTTAGAGGAAGTAA**	58.0
	sIDTP7_ ITS2r	CAGACGTGTGCTCTTCCGATCT **GCTGCGTTCTTCATCGATGC**	64.0

Table 3. Thermocycler conditions for ITS region first stage PCR.

Temperature	Duration	Cycles
98°C	2 minutes	1
98°C	10 seconds	28
52°C	1 second	
68°C	1 second	
4°C	∞	Hold

C. 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 NovaSeq6000 sequencers are subject to 'index hopping' – an unfortunate process that leads incorrect assignment of sequences on a small percentage of reads. By incorporating

unique dual indices (UDIs), mis-assigned sequences are removed from the dataset. IDT has 16 plates of UDI primers for a total of 1536. The second stage PCR amplifies the products of the first stage PCR by targeting the linker sequences. Reactions are performed using the same mastermix as for the first stage PCR, Quantabio 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% isopropyl alcohol. 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 4.

Table 4 – Thermocycler conditions for second stage PCR

Temperature	Duration	Cycles
98°C	2 minutes	1
98°C	10 seconds	8
60°C	1 second	
68°C	1 second	
4°C	∞	Hold

D. 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 Nano or MiniSeq mid-output 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.

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 expected to be in the range of 350-600 bp, depending on fungal lineages present in the samples. If small fragments (< 150 bp) are detected the pool, the pool will undergo an additional SPRI bead cleanup. Additional purification using a Blue PippinPrep device, and employing 1.5% agarose gels, is performed for best results. Size selection is performed from 300-700 bp, according to the manufacturer's instructions. The pool is then analyzed using Qubit prior to loading the QC sequencing run (below).

E. Running the QC Sequencing Run

The pooled libraries are loaded onto a "quality control" QC sequencing run. For example, if loading onto a MiSeq instrument, the MiSeq cartridge is thawed and gently mixed according to the manufacturer's instructions. The final pooled library is then denatured with fresh 0.2N NaOH and diluted to a final DNA concentration of 3.8-4.0 pM. The exact concentration can vary from instrument to instrument and may need to be optimized; recommended cluster density is in the range of 400-700K/mm^2. Next, approximately 30% phiX spike-in is added to the diluted, denatured library and this mixture is loaded on the MiSeq Nano cartridge. 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. QC sequencing may also be performed on a MiniSeq instrument; exact loading conditions will vary depending on which instrument is used for QC sequencing. We have observed that high phiX (~30%) is required for proper sequencing of ITS amplicon libraries.

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

Following the QC sequencing run, the number of clusters for each sample is used to calculate a percent of pass-filter clusters. The percentage of each library is 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 (*e.g.*, 1/960th) and divided by the measured relative abundance of each sample as calculated as a portion of pass-filter, barcoded reads. If the performance of some of the samples precludes pooling to the desired relative abundance, then two independent re-pools will be constructed. These pools, called "low" and "high" represent pools of samples with either lower or higher than average number of QC sequencing reads. Within each pool, samples will be normalized to yield an equal depth of sequencing relative to all samples within each pool. Subsequently, each pool will undergo bead-cleaning, as performed initially for the QC pool. Finally, DNA concentrations will be measured for both pools, and the concentrations of the pools will be equalized. Finally, the two DNA pools will be combined in proportion to the number of samples in each pool.

Repooling. As described above, samples are repooled 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). A liquid handling robotic instrument is highly desirable as each sample will require a different volume; manual pipetting for this 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. Additional purification using a PippinPrep device employing 1.5% agarose gels will be performed on each pool to isolate fragments in the 300-

700 bp range. The pools are then analyzed using Qubit to measure library concentration. The concentrations of the two pools are then equalized by diluting the more concentrated pool. Finally, the two pools are mixed in ratios proportional to the number of samples in each pool. An aliquot of the final pool of libraries is then shipped overnight on blue ice packs to the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign for NovaSeq6000 sequencing. An approximately 30% phiX spike-in is used for sequencing these libraries.

Once the volumes are calculated, an Excel script is used to re-pool the libraries using a robotic liquid handler (*e.g.*, Opentrons). A liquid handling robotic instrument is highly desired as each sample will require a different volume; manual pipetting for this many samples is difficult. The final 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. Additional purification using a PippinPrep device employing 1.5% agarose gels is performed again to isolate fragments in the 300-700 bp range. The pool is then analyzed using Qubit and qPCR to measure library concentration. If qPCR and Qubit results are close (e.g., within 25% of each other), loading will proceed using library concentration calculated by qPCR. If qPCR and Qubit results diverge (i.e., >25% difference), the estimate that has the highest concentration will be used to load the sequencing run. This ensures that the run will not be overloaded. An aliquot of the pooled libraries will then be shipped overnight on blue ice packs to the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign for NovaSeq6000 sequencing. An approximately 30% phiX spike-in will be used for sequencing these libraries.

G. Loading a 500-cycle kit on a NovaSeq6000 instrument

The final pool will be held at +4°C at the receiving facility until the sequencer is loaded. Each pool will be loaded onto a single lane of an Illumina NovaSeq 6000 V1.5 flow cell, according to the manufacturer's instructions. The final pooled library will be denatured with fresh 0.2N NaOH and diluted to a final DNA concentration of 0.9 nM. The exact concentration can vary from instrument to instrument and may need to be optimized. Next, approximately 30% phiX spike is added to the diluted, denatured library and this mixture is prepared with EXAmp reagents and added to the flowcell lane following manufacturer's instructions. The target percentage cluster pass filter is 70-80%. The sequencing run will be performed to generate additional read-length to improve overlap between forward and reverse reads. Given the 20 bases of UDIs, an extra 16 bases are available. Thus, sequencing will be run in a paired-end 2x259 base mode, and these parameters will be set up following the prompts on the instrument.

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, fungal-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 remain after quality trimming and chimera removal. Biologically, the number of reads that can be annotated as fungi 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 (~30% 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. We have observed that a 30% phiX spike-in is necessary for routine successful sequencing of this primer set. However, phiX levels above 30% do not affect quality negatively but reduce overall data yield. Due to expected high output on NovaSeq6000 SP 2x259 sequencing, data loss due to high phiX is not expected to be problematic.

Secondary metrics are performed on a per-sample basis. Per sample-level data acceptance criteria: Each sample is expected to produce at least 10,000 clusters after forward and reverse read merging, primer trimming, and quality trimming. We anticipate substantially higher yields on the NovaSeq6000, likely exceeding >100,000 clusters/sample, though some samples may fail to generate sufficient sequence data. We anticipate that >90% of samples will produce at least 10,000 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 ITS sequencing data table as ‘Fail’.

Overall run success will also be evaluated by examining the amount of data generated for samples, replicates (biological and technical), and negative controls (DNA extraction controls and PCR amplification controls will be assessed independently; ideally will be <1% of the sample average). We will also assess the distribution of clusters generated for each sample, by

sequencing run. After NEON has performed annotation of the sequence data, GMCF will calculate the number of samples, per run, that do not meet the minimum criteria of 10,000 **fungal** sequences after read-merging, primer trimming, quality trimming, and annotation. 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 10,000 sequences after read-merging, primer trimming, quality trimming and annotation (data coming from NEON).

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 BOX folder. Raw sequence data, as multiplexed files, will be uploaded as archive sequences files to a separate designated BOX folder 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 Microbial Marker Gene Sequencing."

Sample name conventions will be used as follows:

[lab id]_[internal sample id]_[sequencing run id]_[marker]_[fwd/rev read]

For example, "RUSH_19S_12_1061_RUN03_ITS_R1/2"

If the sample is a replicate, this will be indicated after the internal sample, as follows: "RUSH_19S_12_1061_REP1_RUN03_ITS_R1/2". The sample name will also indicate whether a sample is a control. For example, on negative control replicate could be named "RUSH_RUN03_NEG01_RUN03_ITS_R1/2".

VI. References

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.

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.

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.