Non-biological synthetic spike-in controls and the AMPtk software pipeline improve fungal high throughput amplicon sequencing data

Jonathan M. Palmer[1]*, Michelle A. Jusino[1]*, Mark T. Banik[1], and Daniel L. Lindner[1]

[1] Center for Forest Mycology Research, US Forest Service, Madison, WI 53726, USA

*authors contributed equally to this manuscript

Keywords:
Internal transcribed spacer, high throughput amplicon sequencing, fungi, environmental sequencing, non-biological synthetic mock community, spike-in control

Correspondence:
Jonathan Palmer jmpalmer@fs.fed.us
Daniel Lindner dlindner@fs.fed.us
1 Gifford Pinchot Drive
Madison, WI 53726

Total Word Count:	6,497
- Introduction:	1,026
- Methods:	1,774
- Results:	2,676
- Discussion:	989
- Acknowledgements:	32

Figures: 6
Tables: 1

Supplementary Figures: 3 color
Supplementary Tables: 5

Summary:

- High throughput amplicon sequencing (HTAS) of conserved DNA regions is a powerful technique to characterize biological communities from environmental samples. Recently, spike-in mock communities have been used to measure accuracy of sequencing platforms and data analysis pipelines. The fungal internal transcribed spacer (ITS) region is difficult to sequence due to its variability (length and sequence divergence) across the fungal kingdom.
- To assess the ability of sequencing platforms and data processing pipelines using fungal ITS amplicons, we created two ITS spike-in control mock communities composed of single copy plasmid DNA: a biological mock community (BioMock), consisting of cloned ITS sequences, and a synthetic mock community (SynMock), consisting of non-biological ITS-like sequences.
- Using these spike-in controls we show that pre-clustering steps for variable length amplicons are critically important and a major source of bias is attributed to initial PCR reactions. These data suggest HTAS read abundances are not representative of starting values.
- We developed AMPtk (amplicon toolkit), a versatile software solution equipped to deal with variable length amplicons featuring a method to quality filter HTAS data based on spike-in controls. While we describe herein a non-biological (synthetic) mock community for ITS sequences, the concept can be widely applied to any HTAS dataset.

Introduction:

High-throughput amplicon sequencing (HTAS) is a powerful tool that is frequently used for examining community composition of environmental samples. HTAS has proven to be a robust and cost-effective solution due to the ability to multiplex hundreds of samples on a single next-generation sequencing (NGS) run. However, HTAS output from environmental samples requires careful interpretation and appropriate and consistent use of positive and negative controls (Nguyen *et al.*, 2015). One of the major challenges in HTAS is to differentiate sequencing error versus real biological sequence variation. Considerable progress has been made in the last several years through improved quality of sequencing results through manufacturer upgrades to reagents as well as improved quality filtering and “clustering” algorithms. While most algorithm development in HTAS is focused on the prokaryotic microbiome, using the 16S subunit of the rRNA array (e.g. QIIME (Caporaso *et al.*, 2010), Mothur (Schloss *et al.*, 2009), UPARSE (Edgar, 2013), DADA2 (Callahan *et al.*, 2016)), many of these same tools have been adopted for use with other groups of organisms, such as fungi.

The internal transcribed spacer (ITS) region of the rRNA array has emerged as the molecular barcode for examining fungal communities in environmental samples (Schoch *et al.*, 2012). The ITS region is multi-copy and thus easily amplifiable via PCR even from environmental samples with low quantities of fungal DNA. The ITS region consists of two subunits, ITS1 and ITS2, and is generally conserved within a species yet possess enough variability to differentiate between species in many taxonomic groups. Because of its widespread use, several public databases are rich with reference fungal ITS sequences (Schoch *et al.*, 2012). However, there are several properties of the fungal ITS region that are potentially problematic for HTAS that include: i) fungi have variable cell wall properties making DNA extraction efficiency unequal for different taxa and/or cell types (hyphae, fruiting bodies, spores, etc) (Vesty *et al.*, 2017), ii) the number of nuclei per cell is variable between taxa (Roper *et al.*, 2011), iii) the number of copies of the rRNA array are different between taxa and in some cases isolates of the same taxa (Ganley & Kobayashi, 2007), iv) a single isolate can have multiple ITS sequences (intragenomic variability; (Simon & Weiss, 2008; Lindner & Banik, 2011), v) the ITS region is highly variable in length, vi) ITS sequences vary in GC content, and vii) there are a variable number of homopolymer repeats. Additionally, current read lengths of next-generation sequencing platforms (Illumina Miseq currently covers ~ 500 bp (2 x 300) and Ion Torrent is 450 bp) are not long enough to cover the entire length of the ITS region, which is typically longer than 500 bp. However, conserved priming sites exist to amplify either the ITS1 region or the ITS2 region, which has been shown to be sufficient for taxonomic identification.

While several studies have used the ITS1 region for HTAS, the ITS1 region contains introns in some taxa and thus to avoid potential bias it has been suggested that ITS2 region should be the preferred region for fungi (Taylor *et al.*, 2016) (Figure 1A).

Sequencing error is a known problem across NGS platforms used for HTAS. To address issues with sequencing error and reliability of results from HTAS, it has become increasingly common practice to use spiked-in “mock” community samples as positive controls for the parameterization and optimization of experimental workflows and data processing. Spike-in mock community controls for fungal ITS have been used (e.g., Amend *et al.*, 2010; Tonge *et al.*, 2014; Nguyen *et al.*, 2015; Taylor *et al.*, 2016; De Filippis *et al.*, 2017), and have consisted of fungal genomic DNA (gDNA) extracted from tissue from fruiting bodies, cultures, or spores of a number of taxa which are then (usually) combined in equimolar amounts. Mock communities composed of fungal gDNA from fruiting bodies, spores, and/or hyphae provide a measure of success of extraction, PCR, and sequencing and thus are useful in the HTAS workflow. However, such mock communities are of limited value if used to validate/parameterize data processing workflows due to intrinsic properties of the ITS region mentioned previously (variable copy number, intraspecific variation, variable length, etc.). Therefore, there is a need for fungal ITS spike-in control mock communities that function to validate laboratory experimental design, validate data processing steps, and compare results between sequencing runs and platforms.

HTAS is cost-effective due to the ability to massively multiplex environmental samples on a single sequencing run. This process depends on the attachment of a unique sequence identifier (referred to as a barcode, an index, or a tag, depending on sequencing platform) to each piece of DNA to be sequenced. In recent years, “index-bleed (“index hopping”, “tag jumping”, “barcode jumping”, “tag switching”, or “barcode switching”) has been noted to occur on Roche 454 platforms as well as Illumina platforms (Kircher *et al.*, 2011; Carlsen *et al.*, 2012; Degnan & Ochman, 2012; Philippe *et al.*, 2015; Schnell *et al.*, 2015). Index-bleed can lead to over-estimation of diversity in environmental samples (Philippe *et al.*, 2015; Schnell *et al.*, 2015) and mis-assignment of sequences to samples. It has been noted that spike-in mock communities may be useful to help detect index-bleed, and subsequent filters may be applied for use with the HTAS pipeline of choice (Degnan & Ochman, 2012; Philippe *et al.*, 2015).

In this study, we generated a biological mock community (BioMock) composed of cloned ITS sequences (single insert plasmids) from a diverse set of fungal taxa. We show how this BioMock can be used to parameterize a data processing workflow. Subsequently, we found that current “off-the-shelf” software solutions performed poorly with our BioMock community of fungal ITS sequences and thus developed AMPtk (amplicon toolkit), which produces improved results of variable length amplicons from HTAS. We demonstrate that read abundances are an unreliable proxy for measuring relative abundances in fungal communities on both Illumina and Ion Torrent sequencing platforms. Accurate measurement of index-bleed between samples can be accomplished by the use of a non-biological synthetic spike-in mock community consisting of ITS-like sequences (SynMock). Finally, we show how AMPtk paired with SynMock can be used to quality filter HTAS data by detecting and mitigating the effects of index-bleed among multiplexed samples.

Materials and Methods:

Biological mock community (BioMock)

To construct the Biomock we selected 26 identified fungal cultures (Supplementary Table S1) from the Center for Forest Mycology Research (CFMR) culture collection (US Forest Service, Madison, Wisconsin). These cultures were purposefully chosen to represent a taxonomic range of fungi, including paralogs, fungi with GC rich ITS regions, a variety of ITS lengths, and fungi with a variety of homopolymers in the ITS region. To measure the sensitivity of our bioinformatics approach, we also included two ITS sequences from *Leptoporus mollis* that were isolated from the same culture as an example of intragenomic variation in the fungal ITS region. These two sequences are more than 3% divergent (95.9% identical) and thus would typically represent separate operational taxonomic units (OTUs) in a clustering pipeline, despite being from the same fungal isolate. All cultures were grown on cellophane on malt extract agar, and DNA was extracted from pure cultures following (Lindner & Banik, 2008). Following extraction, the genomic DNA was PCR amplified using the fungal ITS specific primers ITS-1F (Gardes & Bruns, 1993) and ITS4 (White *et al.*, 1990). The PCR products were then cloned and Sanger sequenced using the ITS1-F primer following the protocol in (Lindner & Banik, 2011). Sequence identifications were verified via BLAST search and two clones of each isolate were selected and cultured in liquid LB (Luria-Bertani) media and incubated at 37 C for 24 hours. Plasmids were purified from the cultures in LB media using standard alkaline lysis. These plasmids will hereafter be termed “purified plasmids”. The purified plasmids were then Sanger sequenced with vector primers T7 and SP6 to verify the insertion of a single copy of the appropriate ITS fragment. We subsequently quantified the purified plasmid DNA concentration using a Qubit® 2.0 fluorometer and DNA concentrations were equilibrated to 10 nM using DNA-free molecular grade water. Following equilibration, 5 µl of each purified plasmid were combined to make an equimolar “biological mock” community of single-copy purified plasmids (BioMock).

PCR has known biases, which are related to different sequence characteristics and are hard to predict in mixed DNA communities of unknown composition. To illustrate the impact of initial PCR bias on the number of reads obtained from each member of a mixed DNA community, we generated individual HTAS-compatible PCR products from each Biomock plasmid which were subsequently mixed (post-PCR) in an equimolar ratio. This was accomplished by PCR amplifying each individual plasmid with the same barcoded primer set. PCR products were purified using E-gel® CloneWell™ 0.8% SYBR® Safe agarose gels (ThermoFisher), quantified using a Qubit® 2.0 fluorometer, and combined into an equimolar mixture post-amplification. This post-PCR combined mock community can be used to examine sequencing error on NGS platforms.

Non-biological mock community (SynMock)

We used the well-annotated ribosomal RNA (rRNA) sequence from *Saccharomyces cerevisiae* as a starting point to design ITS-like synthetic sequences. The ITS adjacent regions of small subunit (SSU) and large subunit (LSU) of *S. cerevisiae* were chosen as anchoring points because of the presence of conserved priming sites ITS1/ITS1-F and ITS4. A 5.8S sequence was designed using *S. cerevisiae* as a base but nucleotides were altered so it would be compatible with several primers in the 5.8S region, including ITS2, ITS3, and fITS7. Random sequences were generated with constrained GC content and sequence length for the ITS1 and ITS2 regions. Twelve unique sequences were synthesized (Genescript) and cloned into pUC57 harboring ampicillin resistance. The SynMock sequences and the script to produce them are available in the OSF repository ((https://osf.io/4xd9r/) as well as packaged into AMPtk distributions. Each plasmid was purified by alkaline lysis, quantified, and an equimolar mixture was created as a template for HTAS library prep.

Preparation of HTAS libraries and NGS Sequencing

HTAS libraries were generated using a proofreading polymerase, Pfx50 (ThermoFisher), and thermocycler conditions were as follows: initial denaturation of 94°C for 3 min, followed by 11 cycles of [94°C for 30 sec, 60°C for 30 sec (drop 0.5°C per cycle), 68°C for 1 min], then 26 cycles of [94°C for 30 sec, 55°C for 30 sec, and 68°C for 1 min], with a final extension of 68°C for 7 minutes. PCR products were cleaned using either E-gel® CloneWell™ 0.8% SYBR® Safe agarose gels (Life Technologies) or Zymo Select-a-size spin columns (Zymo Research). All DNA was quantified using a Qubit® 2.0 fluorometer with the high-sensitivity DNA quantification kit (Life Technologies).

A single step PCR reaction was used to create Ion Torrent compatible sequencing libraries, and primers were designed according to manufacturer's recommendations. Briefly, the forward primer was composed of the Ion A adapter sequence, followed by the Ion key signal sequence, a unique Ion Xpress Barcode sequence (10-12 bp), a single base-pair linker (A), followed by the fITS7 primer (Ihrmark *et al.*, 2012). The reverse primer was composed of the Ion trP1 adapter sequence followed by the conserved ITS4 primer (White *et al.*, 1990). Sequencing on the Ion Torrent PGM was done according to manufacturer's recommendations using an Ion PGM™ Hi-Q™ OT2 Kit, an Ion PGM™ Hi-Q™ Sequencing Kit, an Ion PGM™ sequencing chip (316v2 or 318v2), and raw data were processed with the Ion Torrent Suite v5.0.3 with the "--disable-all-filters" flag given to the BaseCaller. Libraries for Illumina MiSeq were generated by a two-step dual indexing strategy. All samples were PCR amplified with Illumina-fITS7 and Illumina-ITS4 primers. PCR products were cleaned and then dual-barcoded using an 8 cycle PCR reaction using the Illumina Nextera Kit and subsequently sequenced using 2 x 300 bp sequencing kit on the Illumina MiSeq at the University of Wisconsin Biotechnology Center DNA Sequencing Facility. All primers utilized in this study are available via the OSF repository (https://osf.io/4xd9r/).

Data processing using AMPtk

AMPtk is publically available at https://github.com/nextgenusfs/amptk. All primary data and data analysis done in this manuscript is available via the Open Science Framework (https://osf.io/4xd9r/). AMPtk is written in Python and relies on several modules: edlib (Šošic & Šikic, 2017), biopython (Cock *et al.*, 2009), biom-format (McDonald *et al.*, 2012), pandas (McKinney), numpy (van der Walt *et al.*, 2011), and matplotlib modules (Hunter, 2007). External dependencies are USEARCH v9.1.13 (Edgar, 2010) or greater and VSEARCH v2.2.0 (Rognes *et al.*, 2016) or greater. In order to run the DADA2 (Callahan *et al.*, 2016) method R is required along with the shortRead (Morgan *et al.*, 2009) and DADA2 packages. The major steps for processing HTAS data are broken down into i) pre-processing reads, ii) clustering into OTUs, iii) filtering OTU table, and iv) assigning taxonomy.

Pre-processing reads – Data structures from Roche 454 and Ion Torrent are similar where reads are in a single file and have a unique barcode at the 5' end of the read followed by the gene-specific priming site; therefore, AMPtk processes reads from these two platforms very similarly. As a preliminary quality control step, only reads that have a valid barcode and forward primer are retained. Next, reverse primer sequences are removed and reads are trimmed to a user-defined maximum length. Data from Illumina is processed differently because reads are most often paired-end reads and most sequencing centers provide users with de-multiplexed by sample paired-end data (i.e. output of 'bcl2fastq'). AMPtk first merges the paired end reads using USEARCH or VSEARCH, phiX spike-in control is removed with USEARCH, forward and reverse primers are removed if found, and all data are combined into a single file. Pre-processing reads in AMPtk from any of the sequencing platforms results in a single output file that is compatible with all downstream steps.

Clustering reads into OTUs – AMPtk is capable of running several different clustering algorithms including UPARSE, DADA2, UNOISE2, UNOISE3, and reference-based clustering. The algorithms all start with quality filtering using expected errors trimming and are modified slightly in AMPtk to build OTU tables using the original de-multiplexed data; therefore read counts represent what was in the sample prior to quality filtering. This is an important distinction, as expected errors quality trimming (Edgar & Flyvbjerg, 2015) can be rather stringent if long read lengths are used and the amplicons are of variable length.

Index-bleed filtering of OTU tables – Filtering in AMPtk works optimally when a spike-in mock community is sequenced in the dataset. While by default AMPtk is setup to work with the SynMock described herein, any spike-in mock community can be used. AMPtk identifies which OTUs belong to the mock community and calculate index-bleed of that mock community into other samples as well as bleed into the mock community from samples. This calculated index-bleed percentage is then used to filter the OTU table. Filtering is done on a per OTU basis, such that read counts in each OTU that are below the index-bleed threshold are set to zero as they fall within the range of data that could be attributed to index-bleed.

Assigning taxonomy - AMPtk is pre-configured with databases for fungal ITS, fungal LSU, arthropod mtCO1, and prokaryotic 16S; however custom databases are easily created with the 'amptk database' command. Several tools are available for taxonomy assignment in AMPtk including remote blast of the NCBI nt database, RDP Classifier (Wang *et al.*, 2007), global alignment to a custom sequence database, UTAX Classifier (RC Edgar, http://drive5.com/usearch/manual9.2/cmd_utax.html), and the SINTAX Classifier (Edgar, 2016). The default method for taxonomy assignment in AMPtk is a "hybrid" approach that uses classification from global alignment, UTAX, and SINTAX. The best taxonomy is then chosen as follows: i) if global alignment percent identity is > 97% then the top hit is retained, ii) if global alignment percent identity is < 97%, then the best confidence score from UTAX or SINTAX is used, iii) if there is disagreement between taxonomy levels assigned by each method then a least common ancestor (LCA) approach is utilized to generate a conservative estimate of taxonomy. AMPtk also can take a QIIME-like mapping file that can contain all the metadata associated with the HTAS study; the output is then a multi-fasta file containing taxonomy in the headers, a classic OTU table with taxonomy appended, and a BIOM file incorporating the OTU table, taxonomy, and metadata. The BIOM output of AMPtk is compatible with several downstream statistical and visualization software packages such as PhyloSeq (McMurdie & Holmes, 2013).

Accessory scripts in AMPtk - AMPtk has several additional features that will aid the user in analyzing HTAS data. For instance, AMPtk contains a script that will prepare data for submission to the NCBI SRA archive by formatting it properly and outputting a ready-to-submit SRA submission file. The FunGuild (Nguyen *et al.*, 2016) package which assigns OTUs to an annotated database of functional guilds is also incorporated directly into AMPtk. Additionally, users can draw a heatmap of an OTU table as well as summarize taxonomy in a stacked histogram.

Results:

In silico analysis of the fungal ITS region

To gain baseline data on potential amplicons of the ITS1 or ITS2 regions, the ITS1 and ITS2 regions were extracted using priming sites specific for each region (ITS1: ITS1-F and ITS2; ITS2: fITS7 and ITS4) from the UNITE+INSD v7.2 database (Abarenkov *et al.*, 2010) consisting of 736,375 ITS sequences. For comparison, the commonly sequenced V3-V4 region was extracted from prokaryotic 16S sequences from the Silva v128 database (Quast *et al.*, 2013). A length histogram for each dataset as well as summary statistics were generated (Figure 1B; Table 1), indicating that all three of these molecular barcodes have an average length of ~ 250 bp (Table1); however, there was considerable variation in the lengths of the ITS region in comparison to the V3/V4 region of 16S (Figure 1B). Stretches of homopolymer sequences can also be problematic for some NGS platforms (454 and PGM), and thus the number of sequences in this dataset that contained homopolymer stretches greater than 6 nucleotides were calculated (Table 1). Given the small percentage of ITS1 and ITS2 regions that are greater than 450 bp (the current upper limit of the Ion Torrent PGM platform), the number of taxa in the reference database that are unlikely to sequence on the Ion Torrent due to amplicon length is relatively small (Table 1). Illumina MiSeq is now capable of paired end 300 bp read lengths (2 x 300); however, reads need to overlap for proper processing in NGS software platforms and thus a ~ 500 bp size limit would also be able to sequence most taxa in the reference database.

Table 1. Summary statistics of the fungal ITS molecular barcode in comparison to bacterial 16S.

Region	Num Seqs	Avg Length (bp)	% HP[1] > 6	% HP[1] > 8	% > 450 bp
ITS Full Length	696 704	488	55.07%	8.66%	-
ITS1	685 399	247	36.58%	5.60%	3.27%
ITS2	535 200	264	44.19%	5.54%	0.83%
16S (V3/V4)	627 247	253	23.74%	1.02%	-

[1] HP: homopolymer stretches

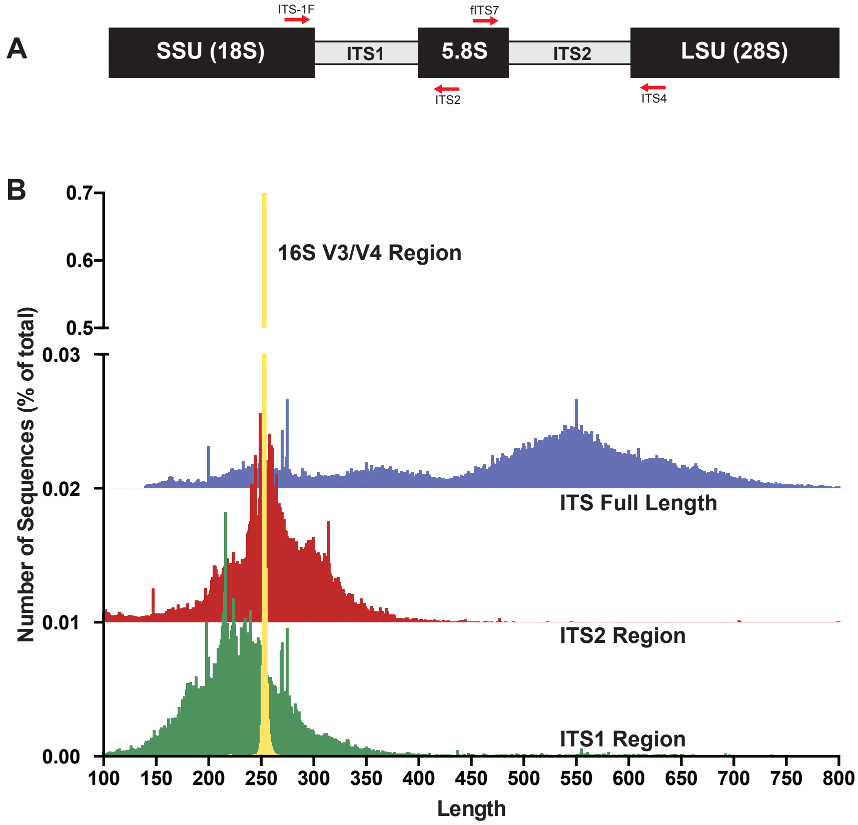


Figure 1. The fungal internal transcribed spacer (ITS) region of the rRNA array is highly variable in length. (A) A schematic of the rRNA array highlights the conserved priming sites commonly used to amplify either the ITS1 or ITS2 region. (B) Size distribution of full length ITS (blue), ITS1 (green), ITS2 (red) sequences in the UNITE v7.2 curated databases shown in comparison to the bacterial 16S V3/V4 amplicon from the Silva v128 database. Current sequencing technologies do not have read lengths long enough to capture full-length ITS sequences, and thus ITS1 or ITS2 regions are used for fungal environmental community analysis. 16S V3/V4 in yellow; ITS full length in blue, ITS2 in red, and ITS1 in green.

Creation of a representative artificial fungal mock community (BioMock)

Given the results from analysis of the UNITE datasets, we set out to create a representative ITS mock community to be used as a spike-in sequencing control to determine the quantitative nature of ITS HTAS and to measure the performance between the commonly used Illumina MiSeq platform versus the Ion Torrent PGM. To circumvent the problematic issues associated with the ITS region, we reasoned that cloned ITS sequences in plasmid form would allow for accurate quantification and pooling, thus providing a means to test the accuracy of the sequencer platforms and data processing workflows. Therefore, we cloned known ITS sequences from 26 cultures from the CFMR culture collection that varied in length (237 bp to 548 bp), ranged in GC content (43.8% - 68.4%), and contained sequences with homopolymer stretches with one sequence containing two 9 bp stretches (Figure 4). These plasmids were combined into a BioMock and BioMock-standards as described in materials and methods section. The value of the BioMock-standards is that the library was combined after PCR, and thus the standards are free from PCR-induced artifacts that may arise from PCR of a mixed community.

Existing data processing workflows perform poorly with fungal ITS sequences

Clustering amplicons into operational taxonomic units (OTUs) is common practice in molecular ecology and there are many software solutions/algorithms (such as QIIME (Caporaso *et al.*, 2010), UPARSE (Edgar, 2013), Mothur (Schloss *et al.*, 2009), and DADA2 (Callahan *et al.*, 2016)) that have been developed to deal appropriately with errors associated with next-generation sequencing platforms. Many studies using 16S amplicon data have focused on comparing clustering methods (Edgar, 2013; Callahan *et al.*, 2016), while others have focused on quality filtering reads prior to clustering (Edgar & Flyvbjerg, 2015). Therefore, we chose not to compare the different software algorithms here, but will briefly mention that when we did run our data through QIIME, the number of OTUs was highly over-estimated and the error rates were very high (Table S2). We were unable to run our data through Mothur due to the inability to do a multiple sequence alignment and subsequent distance matrix of the ITS region. The best performing clustering pipeline was UPARSE; however there were several issues with how the reads were pre-processed and quality filtered that lead to suboptimal results (Table S3 and Table S4). It is important to note that all of these software solutions have been built with 16S amplicons in mind and several have been optimized for Illumina data.

The major difference in 16S amplicons versus those of ITS1/ITS2 is that the lengths of 16S amplicons are nearly identical, while ITS1/ITS2 amplicons vary in length (Figure 1B). This distinguishing feature makes ITS sequences from diverse taxa impossible to align (Schoch *et al.*, 2012) and thus represents a major limitation in data processing. To illustrate the importance of properly pre-processing ITS data, we clustered using UPARSE the ITS1 and ITS2 regions using the UNITE reference database (Figure 2). Using the full length ITS1/ITS2 sequences as a benchmark, we then explored the outcome of trimming/truncating the sequences to different length thresholds, a common practice in OTU clustering pipelines. The UPARSE algorithm uses global alignment and as such terminal mismatches count in the alignment (as opposed to local alignment where terminal mismatches are ignored); thus the UPARSE pipeline expects that reads are truncated to a set length. UPARSE achieves this by truncating all reads to a set length threshold and discards reads that are shorter than the length threshold. Therefore real ITS sequences are discarded (Figure 2). We then came up with two potential solutions to fix this unintended outcome: i) truncate reads that were longer than the threshold and keep all shorter reads (full length), and ii) truncate longer reads and pad the shorter reads with N's out to the length threshold (padding). Using the UNITE v7.2 database of curated sequences (general release June 28th, 2017) as input, both "full-length" and "padding" improved UPARSE results with the "full length" method recovering more than 99% of the expected OTUs (Figure 2).

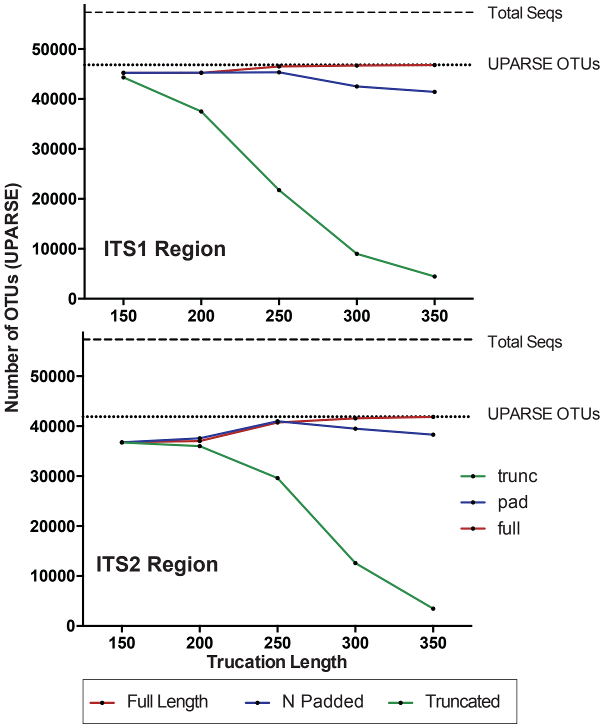


Figure 2. Pre-processing ITS sequences is critically important to accurately recover OTUs using the curated UNITE v7.2 reference database. ITS1 and ITS2 sequences were extracted from the UNITE v7.2 general fasta release database using ‘AMPtk database’. Identical sequences were collapsed (dereplication) and remaining sequences were clustering using UPARSE (‘cluster_otus) to generate the total number of UPARSE OTUs expected for the ITS1 and ITS2 regions. The data was then processed to five different lengths (150, 200, 250, 300, and 350 bp) and then clustered (UPARSE ‘cluster_otus’) using i) default UPARSE truncation (longer sequences are truncated and shorter sequences are discarded), ii) padding with ambiguous bases (longer sequences truncated and shorter sequences padded with N’s to length threshold), and iii) full-length sequences (longer sequences are truncated and shorter sequences are retained if reverse primer is found). Full-length and padding pre-processing sequences outperforms default UPARSE truncation.

Due to the intrinsic nature of the variable length ITS amplicons, we needed a data processing solution that would be flexible enough to maintain the full length of the reads, trim reads without data loss, prepare sequencing reads for downstream clustering algorithms, and support all major NGS platforms. Using the BioMock artificial communities as a means to validate the results of all data processing steps, we wrote a flexible series of scripts for processing Illumina, Ion Torrent, as well as Roche 454 data that are packaged into AMPtk (amplicon tool kit). A flow diagram of AMPtk is illustrated in Figure 3 and a more thorough description of AMPtk is provided in the material and methods section. A manual for AMPtk is available at http://amptk.readthedocs.io/en/latest/. After data is pre-processed with AMPtk via a platform specific method, AMPtk then functions as a wrapper for several popular algorithms including UPARSE, DADA2, UNOISE2, and UNOISE3. All data presented in this manuscript were processed with AMPtk v1.0.1.

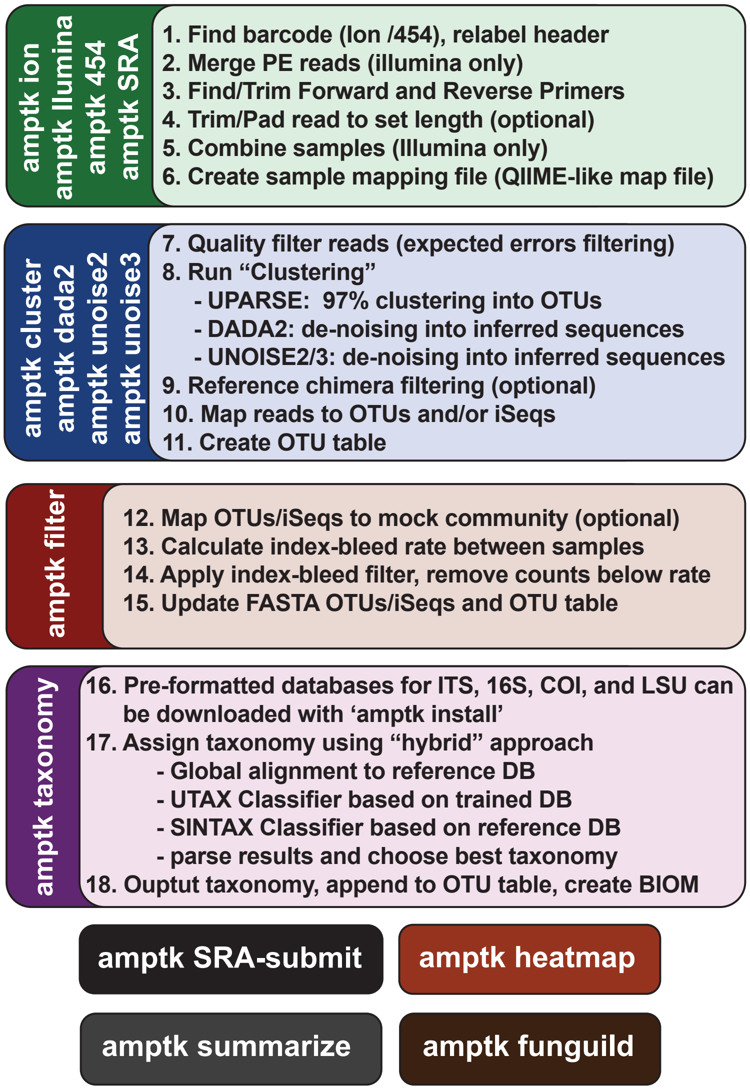
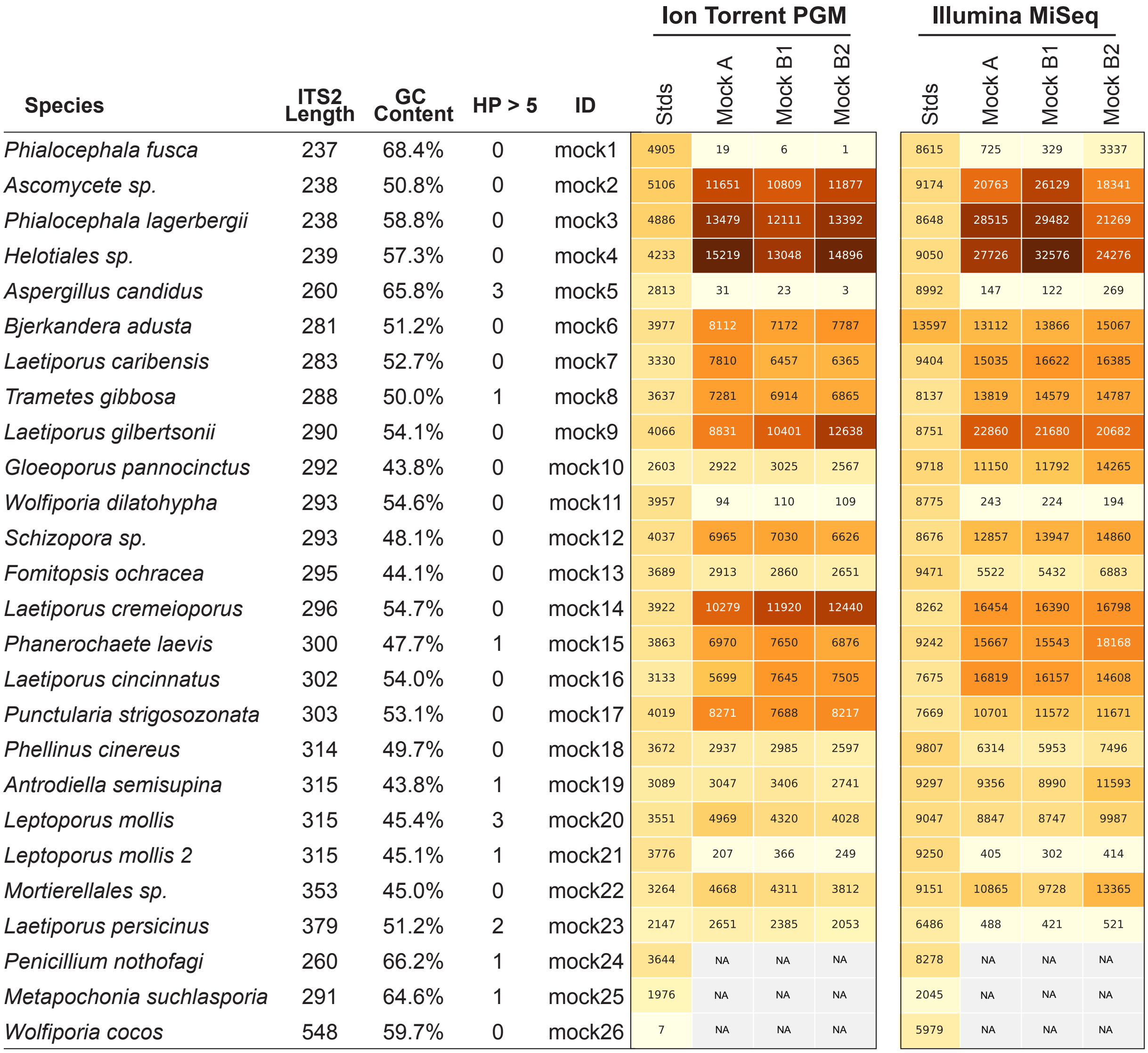

amptk ion
amptk illumina
amptk 454
amptk SRA
1. Find barcode (Ion /454), relabel header
2. Merge PE reads (illumina only)
3. Find/Trim Forward and Reverse Primers
4. Trim/Pad read to set length (optional)
5. Combine samples (Illumina only)
6. Create sample mapping file (QIIME-like map file)
amptk cluster
amptk dada2
amptk unoise2
amptk unoise3
7. Quality filter reads (expected errors filtering)
8. Run "Clustering"
- UPARSE: 97% clustering into OTUs
- DADA2: de-noising into inferred sequences
- UNOISE2/3: de-noising into inferred sequences
9. Reference chimera filtering (optional)
10. Map reads to OTUs and/or iSeqs
11. Create OTU table
amptk filter
12. Map OTUs/iSeqs to mock community (optional)
13. Calculate index-bleed rate between samples
14. Apply index-bleed filter, remove counts below rate
15. Update FASTA OTUs/iSeqs and OTU table
amptk taxonomy
16. Pre-formatted databases for ITS, 16S, COI, and LSU can be downloaded with 'amptk install'
17. Assign taxonomy using "hybrid" approach
- Global alignment to reference DB
- UTAX Classifier based on trained DB
- SINTAX Classifier based on reference DB
- parse results and choose best taxonomy
18. Ouptut taxonomy, append to OTU table, create BIOM
amptk SRA-submit
amptk heatmap
amptk summarize
amptk funguild

Figure 3. Overview of the commands in AMPtk. AMPtk is built to be compatible with multiple sequencing platforms as well as contains several clustering algorithms.

Read abundances do not represent community abundances: PCR introduces bias

Next-generation sequencing platforms are quantitative if the library to be sequenced is unbiased, as is typically the case with RNA-sequencing and whole genome sequencing library prep protocols. However, PCR of mixed communities has long been shown to introduce bias in next-generation sequencing workflows (Aird *et al.*, 2011; Pinto & Raskin, 2012; Kebschull &

Zador, 2015). For HTAS this is an important caveat, as molecular ecologists are interested in diversity metrics of environmental communities as well as their relative abundance. Through the use of mock communities, several studies have pointed out that read abundance from fungal HTAS are not representative of relative biological abundance (Amend *et al.*, 2010; De Filippis *et al.*, 2017). However, it was recently reported that for a fungal ITS mock community of 8 members, abundances were meaningful (Taylor *et al.*, 2016) and many studies continue to use abundance-based metrics to analyze HTAS, without giving any consideration to presence/absence-based metrics. We reasoned we could investigate this issue using the ITS BioMock artificial community, which would not suffer from bias associated with DNA extraction, ITS copy numbers, and intraspecific variation. We compared the relative read abundances of BioMock-standards to 3 different combinations of BioMock on both the Ion Torrent PGM and Illumina MiSeq platforms (Figure 4). The BioMock-standards consist of an equimolar mixture of 26 PCR products, while the BioMock communities consist of an equimolar mixture of 23 single-copy plasmids. These data show that even in an extreme example of an equally mixed community of cloned ITS sequences, read abundance does not represent actual abundance in the mock community (Figure 4). The majority of the bias is introduced at the initial PCR step, as the pre-PCR combined BioMock-standards result in a more equal distribution of reads, albeit not a perfect distribution. We also tested PCR conditions, DNA concentrations, and sample reproducibility on the Ion Torrent PGM (Supplemental Figure S1). While the bias via PCR is consistent between sequencing platforms, there is no obvious correlation between length of the read, GC content, nor stretches of homopolymers affecting efficient PCR amplification. For example, *Wolfiporia dilatophya* (mock11) contains no homopolymer stretches larger than 5, has GC distribution of 54.6%, and is near the median in length, yet it does not PCR amplify well in the BioMock community (Figure 4). These data also show a size limitation in the Ion Torrent PGM workflow, as *Wolfiporia cocos* (mock 26) sequences very poorly due to its long ITS2 region (Figure 5). Three members of the original 26 members of the BioMock community were dropped (mock24, mock25, mock26) due to persistent problems getting them to amplify/sequence in repeated HTAS on the Ion Torrent platform (Supplemental Figure S1).



					Ion Torrent PGM				Illumina MiSeq			
Species	ITS2 Length	GC Content	HP > 5	ID	Stds	Mock A	Mock B1	Mock B2	Stds	Mock A	Mock B1	Mock B2
Phialocephala fusca	237	68.4%	0	mock1	4905	19	6	1	8615	725	329	3337
Ascomycete sp.	238	50.8%	0	mock2	5106	11651	10809	11877	9174	20763	26129	18341
Phialocephala lagerbergii	238	58.8%	0	mock3	4886	13479	12111	13392	8648	28515	29482	21269
Helotiales sp.	239	57.3%	0	mock4	4233	15219	13048	14896	9050	27726	32576	24276
Aspergillus candidus	260	65.8%	3	mock5	2813	31	23	3	8992	147	122	269
Bjerkandera adusta	281	51.2%	0	mock6	3977	8112	7172	7787	13597	13112	13866	15067
Laetiporus caribensis	283	52.7%	0	mock7	3330	7810	6457	6365	9404	15035	16622	16385
Trametes gibbosa	288	50.0%	1	mock8	3637	7281	6914	6865	8137	13819	14579	14787
Laetiporus gilbertsonii	290	54.1%	0	mock9	4066	8831	10401	12638	8751	22860	21680	20682
Gloeoporus pannocinctus	292	43.8%	0	mock10	2603	2922	3025	2567	9718	11150	11792	14265
Wolfiporia dilatohypha	293	54.6%	0	mock11	3957	94	110	109	8775	243	224	194
Schizopora sp.	293	48.1%	0	mock12	4037	6965	7030	6626	8676	12857	13947	14860
Fomitopsis ochracea	295	44.1%	0	mock13	3689	2913	2860	2651	9471	5522	5432	6883
Laetiporus cremeioporus	296	54.7%	0	mock14	3922	10279	11920	12440	8262	16454	16390	16798
Phanerochaete laevis	300	47.7%	1	mock15	3863	6970	7650	6876	9242	15667	15543	18168
Laetiporus cincinnatus	302	54.0%	0	mock16	3133	5699	7645	7505	7675	16819	16157	14608
Punctularia strigosozonata	303	53.1%	0	mock17	4019	8271	7688	8217	7669	10701	11572	11671
Phellinus cinereus	314	49.7%	0	mock18	3672	2937	2985	2597	9807	6314	5953	7496
Antrodiella semisupina	315	43.8%	1	mock19	3089	3047	3406	2741	9297	9356	8990	11593
Leptoporus mollis	315	45.4%	3	mock20	3551	4969	4320	4028	9047	8847	8747	9987
Leptoporus mollis 2	315	45.1%	1	mock21	3776	207	366	249	9250	405	302	414
Mortierellales sp.	353	45.0%	0	mock22	3264	4668	4311	3812	9151	10865	9728	13365
Laetiporus persicinus	379	51.2%	2	mock23	2147	2651	2385	2053	6486	488	421	521
Penicillium nothofagi	260	66.2%	1	mock24	3644	NA	NA	NA	8278	NA	NA	NA
Metapochonia suchlasporia	291	64.6%	1	mock25	1976	NA	NA	NA	2045	NA	NA	NA
Wolfiporia cocos	548	59.7%	0	mock26	7	NA	NA	NA	5979	NA	NA	NA

Figure 4. Read abundance is an unreliable proxy for actual abundance within a mixed community. Using an equimolar mixture of cloned ITS sequences in plasmid form (MockA, MockB1, MockB2) in comparison to equimolar mixture of individual PCR products (Stds) illustrates that the initial PCR reaction during library preparation heavily biases the read abundance obtained after sequencing on both the Ion Torrent PGM and Illumina MiSeq platforms. While read abundances are unreliable, all members of the mock community were recovered. MockA represents a 1:16,000 dilution and MockB1/MockB2 are replicates of a 1:32,000 dilution of the BioMock community. The Ion Torrent PGM platform has a length threshold of approximately 450 bp; therefore longer amplicons like *Wolfiporia cocos* ITS2 sequence very poorly.

In HTAS experiments, considerable effort is made to try to sequence to an equal depth for each sample. However, in practice this rarely works perfectly and thus a typical HTAS dataset has a 2-4X range in number of reads per sample. The depth of sequence range for the HTAS runs presented here is within a range of 2X for each run and the smallest number of reads per sample in any of our sequencing runs was nearly 60,000 (Supplemental Table S5). Unequal sequencing depth has been used as rationale for explaining the lack of correlation between read abundance and actual abundance. Therefore, random subsampling of reads in each sample prior to clustering (also called rarefying) has been widely used in the literature, despite a compelling statistical argument that this method is flawed (McMurdie & Holmes, 2014). Randomly subsampling reads for each sample using our BioMock community yielded nearly identical results (Supplemental Figure S2). Sequencing depth has been shown to be an important variable for HTAS experiments (Smith & Peay, 2014), therefore we typically employ a 5,000 reads per sample cutoff when processing environmental datasets.

A non-biological (synthetic) mock community to measure index-bleed among samples

Index-bleed is a phenomenon that has been described on Roche 454 platform (Carlsen *et al.*, 2012) as well as Illumina platforms (Kircher *et al.*, 2012; Wright & Vetsigian, 2016). A consensus on a mechanism of index-bleed during the sequencing run has yet to be reached. Index-bleed is a significant challenge to overcome as sample crossover has the potential to over-estimate diversity and lead to inaccurate representations of microbial communities, especially considering that read abundance is an unreliable proxy for biological abundance (Figure 4). Using our BioMock sequencing results, we also discovered this phenomenon on both Ion Torrent and Illumina platforms. We calculated the rate of index-bleed in our BioMock Ion Torrent sequencing run to be 0.033% and on Illumina MiSeq between 0.233% and 0.264%. We also confirmed that index-bleed was happening on the Illumina flow-cell by re-running a subset of samples that had shown high index-bleed on different flowcell that did not contain the BioMock (Supplemental Figure S3). One problem that we noticed in measuring index-bleed using a mock community of actual ITS sequences (BioMock) was that these same taxa in the mock community could be present in environmental samples, which would lead to inaccurate estimation of index-bleed. In our environmental data, it was likely that at least one of the BioMock members was present in several of the environmental samples, suggesting the calculated index-bleed could be over-estimated. To overcome this problem, we designed a non-biological (synthetic) mock community composed of ITS-like sequences that contained conserved priming sites (SSU and LSU regions), ITS1 region, 5.8S region, and an ITS2 region (Figure 6). We designed the ITS1 and ITS2 portions of the sequences to be non-biological; that is, no similar sequences are known to occur in nature (based on searches of known databases and based on the infinitesimally low probability that a randomly generated sequence would match something found in nature) and therefore these non-biological sequences can be used to accurately track index-bleed in HTAS studies. Using the summary statistics from the analysis of the UNITE reference database for guidance, we also varied the length, GC content, and homopolymer stretches to be representative of real ITS sequences.

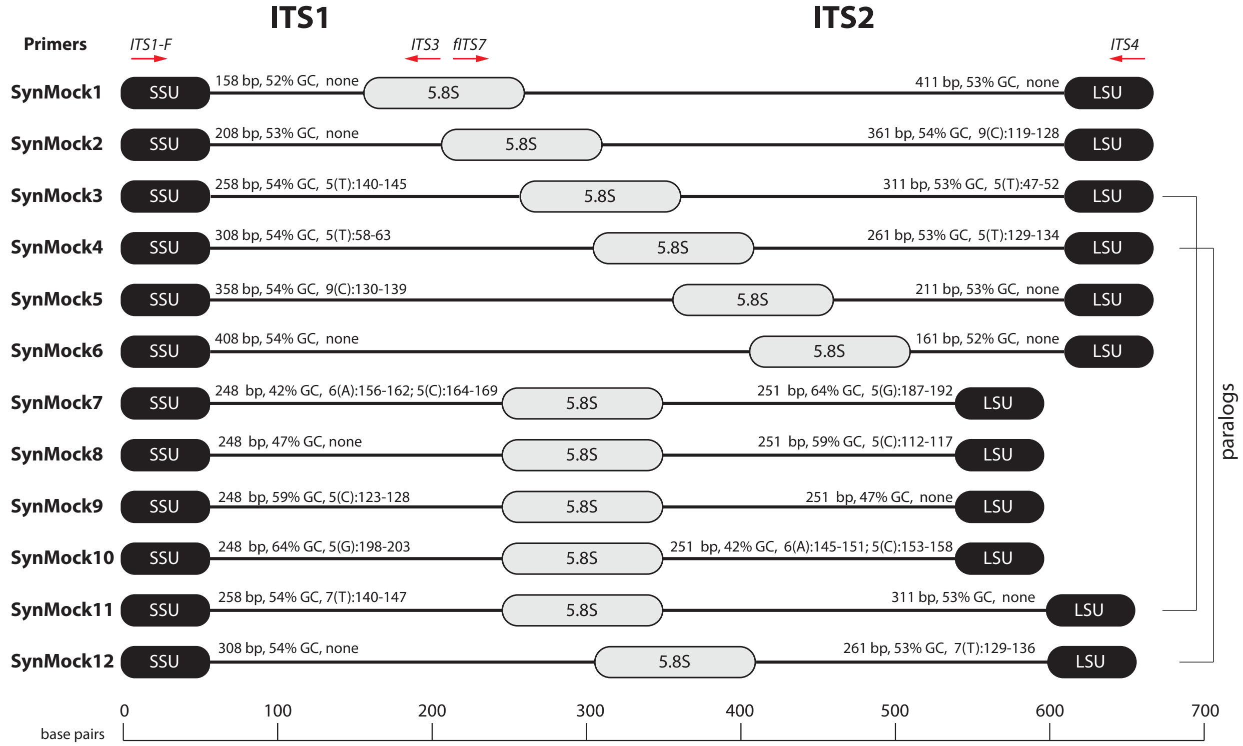
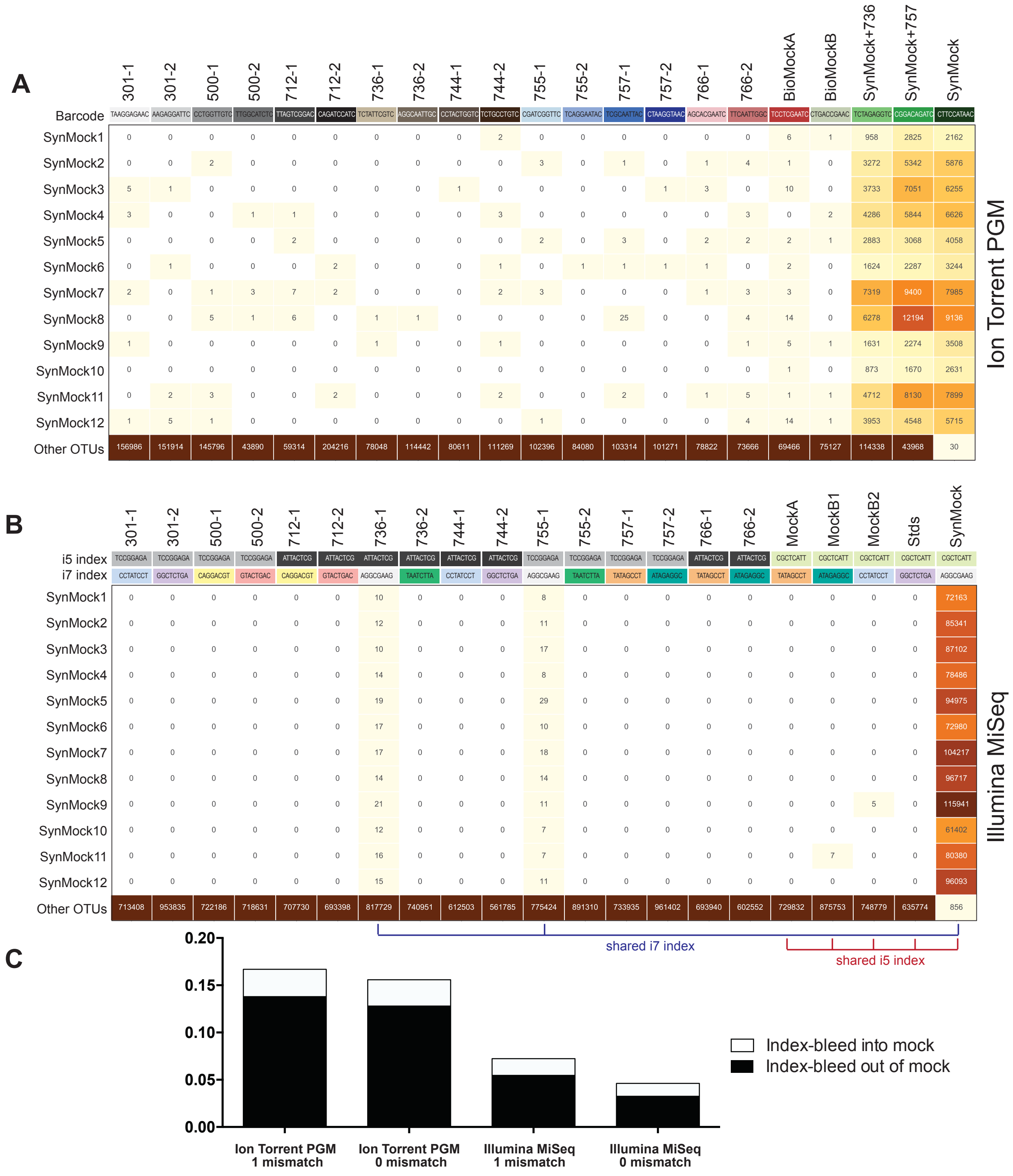


Figure 5. Schematic drawing of the 12-member non-biological synthetic mock community (SynMock). Conserved priming sites for either ITS1 or ITS2 amplicons are retained for versatility. The length distribution, GC content, and homopolymer stretches are representative of curated public databases, however, the sequences are non-biological and thus not found in nature.

The SynMock was tested as a spike-in control on both the Ion Torrent and Illumina MiSeq platforms. The raw data were processed using AMPtk and clustered using UPARSE. These data illustrate that the synthetic sequences are able to be processed simultaneously with real ITS sequences and provide a way to track the level of index-bleed between multiplexed samples (Figure 6). The increased benefit of being able to track the SynMock sequences as they “bleed” out of the sample allows for a more accurate measurement of index-bleed. Using default Illumina de-multiplexing (allowing 1 mismatch in the index sequence), index-bleed using the SynMock community was 0.072% (Figure 6C). To determine if allowing mismatches in the index reads was increasing index-bleed, we reprocessed the data with 0 mismatches and found that index-bleed was reduced to 0.046%. While index-bleed was reduced by nearly half, the tradeoff was that 0 mismatches resulted in approximately 10% fewer reads. For most datasets, a loss of 10% of the sequencing reads should not be problematic, especially if the benefit is to reduce index-bleed in the data. We noted that in our Illumina dual-indexing library prep that there was increased index-bleed on samples that had a shared reverse index (i7), suggesting that errors are increased at later stages of an Illumina sequencing run (Figure 6B). A similar pattern was observed with Ion Torrent PGM data, although not as drastic. Allowing 1 mismatch in the barcode resulted in 0.167% index-bleed while allowing 0 mismatches in the barcode resulted in 0.156% index-bleed (Figure 6C). While these data would suggest that index-bleed is perhaps higher in Ion Torrent PGM datasets, we have subsequently used the SynMock on more than 10 different HTAS Ion Torrent PGM experiments and have since seen much lower levels of index-bleed, occasionally approaching 0% index-bleed.

Many environmental samples can contain hundreds of taxa and thus a legitimate concern is that the 12 member SynMock community does not represent a realistic community in terms of diversity in a sample. To test if the SynMock was able to be recovered in a more complex community, we mixed SynMock together with two environmental samples that had more than 200 OTUs in previous sequencing runs. These mixed samples show that SynMock could be recovered from a complex community and the sequences behave like real ITS sequences (Figure 6A). While many studies have set a read count threshold to filter “noisy” data from OTU tables, this threshold has been typically selected arbitrarily, i.e. OTUs with read counts less than 100 or less than 10% of the total, etc. Use of the SynMock spike-in control allowed for data driven thresholds to be measured and moreover for the ability to filter the OTU table based on the calculated index-bleed. The AMPtk filter command calculates index-bleed by mapping the OTUs to the mock community and then provides a way to filter the OTU table based on this calculated value. AMPtk filters across each OTU in the table such that difficult to sequence or “low abundance” OTUs are not indiscriminately dropped. Taken together, these data illustrate the utility of a non-biological mock community in parameterizing data processing steps and importantly providing a method in AMPtk to reduce index-bleed from HTAS datasets. AMPtk provides an easy to use method to accurately process variable length amplicons, cluster them into OTUs or denoise sequences, generate an OTU table, filter the OTU table for index-bleed, and assign taxonomy.



	301-1	301-2	500-1	500-2	712-1	712-2	736-1	736-2	744-1	744-2	755-1	755-2	757-1	757-2	766-1	766-2	BioMockA	BioMockB	SynMock+736	SynMock+757	SynMock
Barcode	TAAGGAGAAC	AAGAGGATTC	CCTGGTTGTC	TTGGCATCTC	TTAGTCGGAC	CAGATCCATC	TCTATTCGTC	AGGCAATTGC	CCTACTGGTC	TCTGCCTGTC	CGATCGGTTC	TCAGGAATAC	TCGCAATTAC	CTAAGGTAAC	AGCACGAATC	TTCAATTGGC	TCCTCGAATC	CTGACCGAAC	TCTAGAGGTC	CGGACAGATC	CTTCCATAAC
SynMock1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	6	1	958	2825	2162
SynMock2	0	0	2	0	0	0	0	0	0	0	3	0	1	0	1	4	1	0	3272	5342	5876
SynMock3	5	1	0	0	0	0	0	0	1	0	0	0	0	1	3	0	10	0	3733	7051	6255
SynMock4	3	0	0	1	1	0	0	0	0	3	0	0	0	0	0	3	0	2	4286	5844	6626
SynMock5	0	0	0	0	2	0	0	0	0	0	2	0	3	0	2	2	2	1	2883	3068	4058
SynMock6	0	1	0	0	0	2	0	0	0	1	0	1	1	1	1	0	2	0	1624	2287	3244
SynMock7	2	0	1	3	7	2	0	0	0	2	3	0	0	0	1	3	3	0	7319	9400	7985
SynMock8	0	0	5	1	6	0	1	1	0	0	0	0	25	0	0	4	14	0	6278	12194	9136
SynMock9	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	5	1	1631	2274	3508
SynMock10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	873	1670	2631
SynMock11	0	2	3	0	0	2	0	0	0	2	0	0	2	0	1	5	1	1	4712	8130	7899
SynMock12	1	5	1	0	0	0	0	0	0	0	1	0	0	0	0	4	14	1	3953	4548	5715
Other OTUs	156986	151914	145796	43890	59314	204216	78048	114442	80611	111269	102396	84080	103314	101271	78822	73666	69466	75127	114338	43968	30

	301-1	301-2	500-1	500-2	712-1	712-2	736-1	736-2	744-1	744-2	755-1	755-2	757-1	757-2	766-1	766-2	MockA	MockB1	MockB2	Stds	SynMock
i5 index	TCCGGAGA	TCCGGAGA	TCCGGAGA	TCCGGAGA	ATTACTCG	ATTACTCG	ATTACTCG	ATTACTCG	ATTACTCG	ATTACTCG	TCCGGAGA	TCCGGAGA	TCCGGAGA	TCCGGAGA	ATTACTCG	ATTACTCG	CGCTCATT	CGCTCATT	CGCTCATT	CGCTCATT	CGCTCATT
i7 index	CCTATCCT	GGCTCTGA	CAGGACGT	GTACTGAC	CAGGACGT	GTACTGAC	AGGCGAAG	TAATCTTA	CCTATCCT	GGCTCTGA	AGGCGAAG	TAATCTTA	TATAGCCT	ATAGAGGC	TATAGCCT	ATAGAGGC	TATAGCCT	ATAGAGGC	CCTATCCT	GGCTCTGA	AGGCGAAG
SynMock1	0	0	0	0	0	0	10	0	0	0	8	0	0	0	0	0	0	0	0	0	72163
SynMock2	0	0	0	0	0	0	12	0	0	0	11	0	0	0	0	0	0	0	0	0	85341
SynMock3	0	0	0	0	0	0	10	0	0	0	17	0	0	0	0	0	0	0	0	0	87102
SynMock4	0	0	0	0	0	0	14	0	0	0	8	0	0	0	0	0	0	0	0	0	78486
SynMock5	0	0	0	0	0	0	19	0	0	0	29	0	0	0	0	0	0	0	0	0	94975
SynMock6	0	0	0	0	0	0	17	0	0	0	10	0	0	0	0	0	0	0	0	0	72980
SynMock7	0	0	0	0	0	0	17	0	0	0	18	0	0	0	0	0	0	0	0	0	104217
SynMock8	0	0	0	0	0	0	14	0	0	0	14	0	0	0	0	0	0	0	0	0	96717
SynMock9	0	0	0	0	0	0	21	0	0	0	11	0	0	0	0	0	0	0	5	0	115941
SynMock10	0	0	0	0	0	0	12	0	0	0	7	0	0	0	0	0	0	0	0	0	61402
SynMock11	0	0	0	0	0	0	16	0	0	0	7	0	0	0	0	0	0	7	0	0	80380
SynMock12	0	0	0	0	0	0	15	0	0	0	11	0	0	0	0	0	0	0	0	0	96093
Other OTUs	713408	953835	722186	718631	707730	693398	817729	740951	612503	561785	775424	891310	733935	961402	693940	602552	729832	875753	748779	635774	856

Figure 6. Index-bleed or sample mis-assignment occurs on both Ion Torrent PGM and Illumina Miseq. (A) Read counts from the SynMock community run on the Ion Torrent PGM platform. SynMock reads can be found in environmental samples and reads from the environmental samples are found in the SynMock sample. The data were processed allowing 0 mismatches in the barcode sequence and there is no clear pattern to index-bleed on the Ion Torrent PGM platform. (B) Data processed on the Illumina MiSeq (2x300) allowing 0 mismatches in the index reads show index-bleed in and out of the SynMock sample. Samples that share an index (i5 or i7) show an increase in index-bleed. (C) Index-bleed between samples can be tracked using the

SynMock spike-in control, where AMPtk will measure both index-bleed into the SynMock as well as index-bleed into other samples. These calculated values are then used by AMPtk to filter an OTU table to remove read counts that fall below the index-bleed threshold. Index-bleed is reduced if 0 mismatches are allowed in the barcode/index sequence, however, this is still not sufficient to reduce all index-bleed.

Discussion

HTAS studies have the goal of measuring environmental diversity; however, there are technical limitations that need to be understood in order to reach justifiable conclusions. Mock communities and negative controls have been shown to have great utility for HTAS studies, and expanding upon this concept, we present a non-biological synthetic mock community of ITS-like sequences for use as a technical spike-in control for fungal biodiversity studies. Additionally, we describe AMPtk, a software tool kit for analyzing variable length amplicons such as the fungal ITS1 or ITS2 molecular barcodes. These two tools can be coupled together to validate data processing pipelines and reduce index-bleed from OTU tables prior to downstream community ecology analyses. The concept of a non-biological synthetic spike-in control can be expanded to many different genes and organisms, including 16S for microbiome studies.

The ITS region is widely used as a molecular barcode in fungal biodiversity studies as it is easy to amplify and public reference databases are robust. However, HTAS with the ITS region presents some unique challenges due to variability in sequence characteristics such as length and copy number. Most HTAS software development and optimization has been focused on the 16S molecular barcode, a region that is near uniform in length across prokaryotic taxa. Thus, there is a need for a software solution that can more accurately account for variable length amplicons. We developed a single-copy mock community based on cloned ITS sequences as a tool to validate and compare different NGS platforms and data processing pipelines. Using an artificial single-copy mock community of cloned ITS sequences in plasmids (BioMock), we determined that the core clustering/denoising algorithms work for variable length amplicons; however, pre-processing techniques widely used for uniform length amplicons introduce significant error into the pipelines. Simplifying the pre-processing of sequencing reads (i.e., identifying unique sequence barcodes, forward/reverse primers, and trimming reads to a uniform length without data loss) resulted in large improvement in downstream OTU clustering. The pre-processing of reads prior to quality filtering is critical for variable length amplicons and is implemented in AMPtk.

Proper pre-processing of variable length amplicons improves clustering results substantially. However, the BioMock results illustrated that read abundances obtained from

HTAS are not a reliable proxy for inferring biological relative abundance. These data do support use of presence/absence (binary) metrics as we were able to recover all members of our mock community, even when they were spiked into a diverse environmental sample. We identified the initial PCR reaction (library construction) as the major source of read number bias, a conclusion consistent with the literature (Polz & Cavanaugh, 1998; Wu *et al.*, 2010; Jusino *et al.*, 2017). To reduce PCR artifacts for any assay it is generally accepted that one should use the fewest cycles possible, the most concentrated DNA possible, and it has been suggested to use a proofreading polymerase (Oliver *et al.*, 2015). We have tested DNA concentration and PCR cycle numbers for HTAS library generation and subsequent sequencing on the Ion Torrent PGM platform, and our results were consistent with these general guidelines (Supplemental Figure S1). However, following these guidelines is not sufficient to reduce the bias in read abundance from a mixed community from PCR. The Ion Torrent PGM platform currently has an amplicon size limit of ~ 450 bp, and thus some very large ITS sequences are difficult to sequence. However, there are only a small number of known ITS1 or ITS2 sequences that are longer than 450 bp (Table 1) and therefore either platform, Ion Torrent or MiSeq, provided similar results under the conditions tested.

Index-bleed has recently been acknowledged by Illumina (https://tinyurl.com/illumina-hopping), although they limit their acknowledgement to a new flow cell on the HiSeq and NovaSeq platforms. Several studies have shown that older instruments/flowcells have also shown index-bleed, albeit at a much lower rate (Kircher *et al.*, 2012; Wright & Vetsigian, 2016) and index-bleed has been identified on Roche 454 (Carlsen *et al.*, 2012). Here we report a low rate of index-bleed on both Ion Torrent and Illumina MiSeq platforms. While the effective rate of index-bleed is low (< 0.2%), coupled with the fact that read number is not a reliable proxy of community abundance, index-bleed in datasets being analyzed by presence-absence metrics is a problematic scenario. To identify and combat index-bleed, we created a non-biological synthetic mock community (SynMock) of ITS-like sequences that behave like real ITS sequences during the HTAS workflow. Because the SynMock sequences are not known to occur in nature, they can be effectively used to measure index-bleed in a sequencing run. A similar approach was recently described for 16S amplicons using synthesized oligonucleotides (Kim *et al.*, 2017). We propose that HTAS studies of fungal ITS communities should employ SynMock or a similar non-biological mock community as a technical control. Additional controls such as a biological mock community of mixed fruiting bodies, spores, hyphae, etc. of taxa of interest are also useful if the experiment is designed to identify the prevalence of particular taxa.

The bioinformatics pipeline presented here, AMPtk, was developed to specifically address the quality issues that we have identified by using spike-in mock communities and to provide the scientific community with a necessary tool to study fungal community diversity. AMPtk is a flexible solution that can be used to study other regions used in HTAS, such as mitochondrial cytochrome oxidase 1 (mtCO1) of insects and the large subunit (LSU) of the rRNA array. The goal of AMPtk is to reduce data processing to a few simple steps and to improve the output of HTAS studies. Due to the inherent properties of HTAS and the ITS molecular barcode, we take the position that studies of this nature should be used as a preliminary survey of which taxa present in an ecosystem and that inferring relative abundance from read numbers should be considered cautiously. To understand relative abundance of particular taxa in a community, additional independent assays such as taxa specific qPCR or digital PCR are warranted.

Acknowledgements

We sincerely thank Rita Rentmeester for assisting with the growth of some the cultures used to create the biological mock community. Funding was provided by the US Forest Service, Northern Research Station.

Data availability

Raw sequencing reads and data processing scripts are available at the Open Science Framework at https://osf.io/4xd9r/. Data will be deposited in NCBI SRA prior to publication.

Author Contributions

All authors conceived and designed the experiments. MAJ and MTB conducted laboratory experiments. JMP analyzed sequence data and wrote AMPtk. JMP wrote the paper with input from all authors.

Competing financial interests

The authors declare no competing financial interests.

Literature Cited

Abarenkov K, Henrik Nilsson R, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, et al. 2010. The UNITE database for molecular identification of fungi--recent updates and future perspectives. *The New phytologist* **186**(2): 281-285.

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**(2): R18.

Amend AS, Seifert KA, Bruns TD. 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology* **19**(24): 5555-5565.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**(5): 335-336.

Carlsen T, Aas AB, Lindner D, Vrålstad T, Schumacher T, Kauserud H. 2012. Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology* **5**(6): 747-749.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)* **25**(11): 1422-1423.

De Filippis F, Laiola M, Blaiotta G, Ercolini D. 2017. Different amplicon targets for sequencing-based studies of fungal diversity. *Applied and Environmental Microbiology*.

Degnan PH, Ochman H. 2012. Illumina-based analysis of microbial community diversity. *The ISME journal* **6**(1): 183-194.

Edgar R. 2016. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*: 074161.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**(19): 2460-2461.

Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**(10): 996-998.

Edgar RC, Flyvbjerg H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*.

Ganley ARD, Kobayashi T. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research* **17**(2): 184-191.

Gardes M, Bruns T. 1993. ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. *Molecular Ecology* **2**: 113-118.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**(3): 90-95.

Ihrmark K, Bödeker IT, Cruz-Martinez K, Friberg H, Kubartova A, Schenck J, Strid Y, Stenlid J, Brandström-Durling M, Clemmensen KE. 2012. New primers to amplify the fungal ITS2 region–evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology* **82**(3): 666-677.

Jusino M, Banik M, Palmer J, Wray A, Xiao L, Pelton E, Barber J, Kawahara A, Gratton C, Peery M, et al. 2017. An improved method for utilizing high-throughput amplicon sequencing to determine the diets of insectivorous animals. *PeerJ Preprints* **5**: e3184v3181.

Kebschull JM, Zador AM. 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research* **43**(21): e143.

Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, Lauder A, Sherrill-Mix S, Chehoud C, Kelsen J, et al. 2017. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* **5**(1): 52.

Kircher M, Heyn P, Kelso J. 2011. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics* **12**: 382.

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research* **40**(1): e3.

Lindner DL, Banik MT. 2008. Molecular phylogeny of Laetiporus and other brown rot polypore genera in North America. *Mycologia* **100**(3): 417-430.

Lindner DL, Banik MT. 2011. Intragenomic variation in the ITS rDNA region obscures phylogenetic relationships and inflates estimates of operational taxonomic units in genus *Laetiporus*. *Mycologia* **103**(4): 731-740.

McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, et al. 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* **1**(1): 7.

McKinney W Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*. 51-56.

McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One* **8**(4): e61217.

McMurdie PJ, Holmes S. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology* **10**(4): e1003531.

Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. 2009. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics (Oxford, England)* **25**(19): 2607-2608.

Nguyen NH, Smith D, Peay K, Kennedy P. 2015. Parsing ecological signal from noise in next generation amplicon sequencing. *The New phytologist* **205**(4): 1389-1393.

Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L. 2016. FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild - ScienceDirect. *Fungal Ecology*.

Oliver AK, Brown SP, Callaham MA, Jumpponen A. 2015. Polymerase matters: non-proofreading enzymes inflate fungal community richness estimates by up to 15 %. *Fungal Ecology*.

Philippe E, Franck L, Jan P. 2015. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*: gkv107.

Pinto AJ, Raskin L. 2012. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PloS One* **7**(8): e43093.

Polz MF, Cavanaugh CM. 1998. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology* **64**(10): 3724-3730.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**(Database issue): D590-596.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Roper M, Ellison C, Taylor JW, Glass NL. 2011. Nuclear and genome dynamics in multinucleate ascomycete fungi. *Current Biology* **21**(18): R786-793.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**(23): 7537-7541.

Schnell IB, Bohmann K, Gilbert MTP. 2015. Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources* **15**(6): 1289-1303.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Bolchacova E, Voigt K, Crous PW. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences* **109**(16): 6241-6246.

Simon UK, Weiss M. 2008. Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Molecular Biology and Evolution* **25**(11): 2251-2254.

Smith DP, Peay KG. 2014. Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PloS One* **9**(2): e90234.

Šošic M, Šikic M. 2017. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance. *Bioinformatics (Oxford, England)* **33**(9): 1394-1395.

Taylor DL, Walters WA, Lennon NJ, Bochicchio J, Krohn A, Caporaso JG, Pennanen T. 2016. Accurate estimation of fungal diversity and abundance through improved lineage-specific primers optimized for Illumina amplicon sequencing. *Applied and Environmental Microbiology* **82**(24): 7217-7226.

Tonge DP, Pashley CH, Gant TW. 2014. Amplicon-based metagenomic analysis of mixed fungal samples using proton release amplicon sequencing. *PloS One* **9**(4): e93849.

van der Walt Sf, Colbert SC, Varoquaux Gl. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **13**(2): 22-30.

Vesty A, Biswas K, Taylor MW, Gear K, Douglas RG. 2017. Evaluating the impact of DNA extraction method on the representation of human oral bacterial and fungal communities. *PloS One* **12**(1): e0169877.

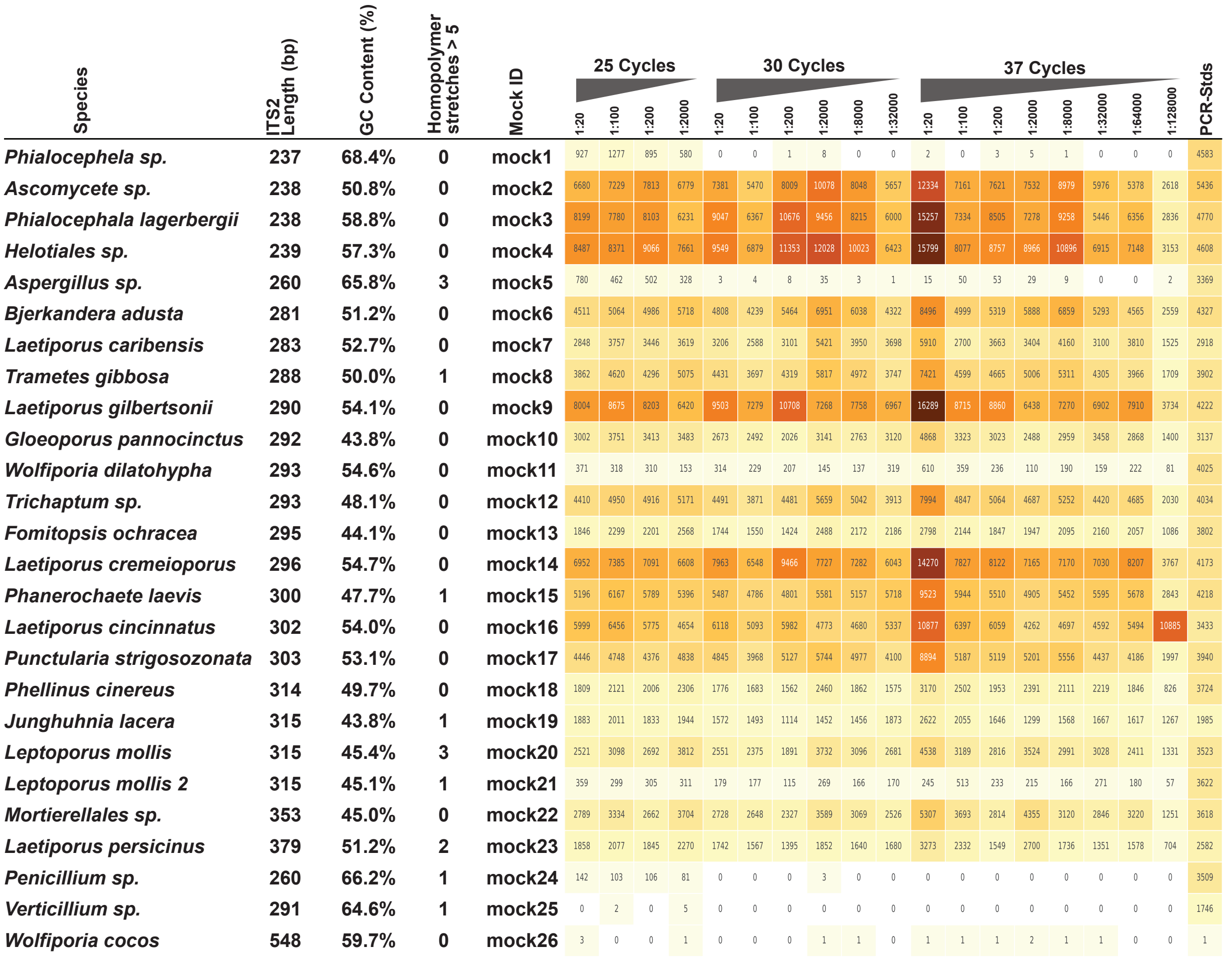
Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**(16): 5261-5267.

White T, Bruns TD, Lee S, Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods and applications*: 315 - 322.

Wright ES, Vetsigian KH. 2016. Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics* **17**(1): 876.

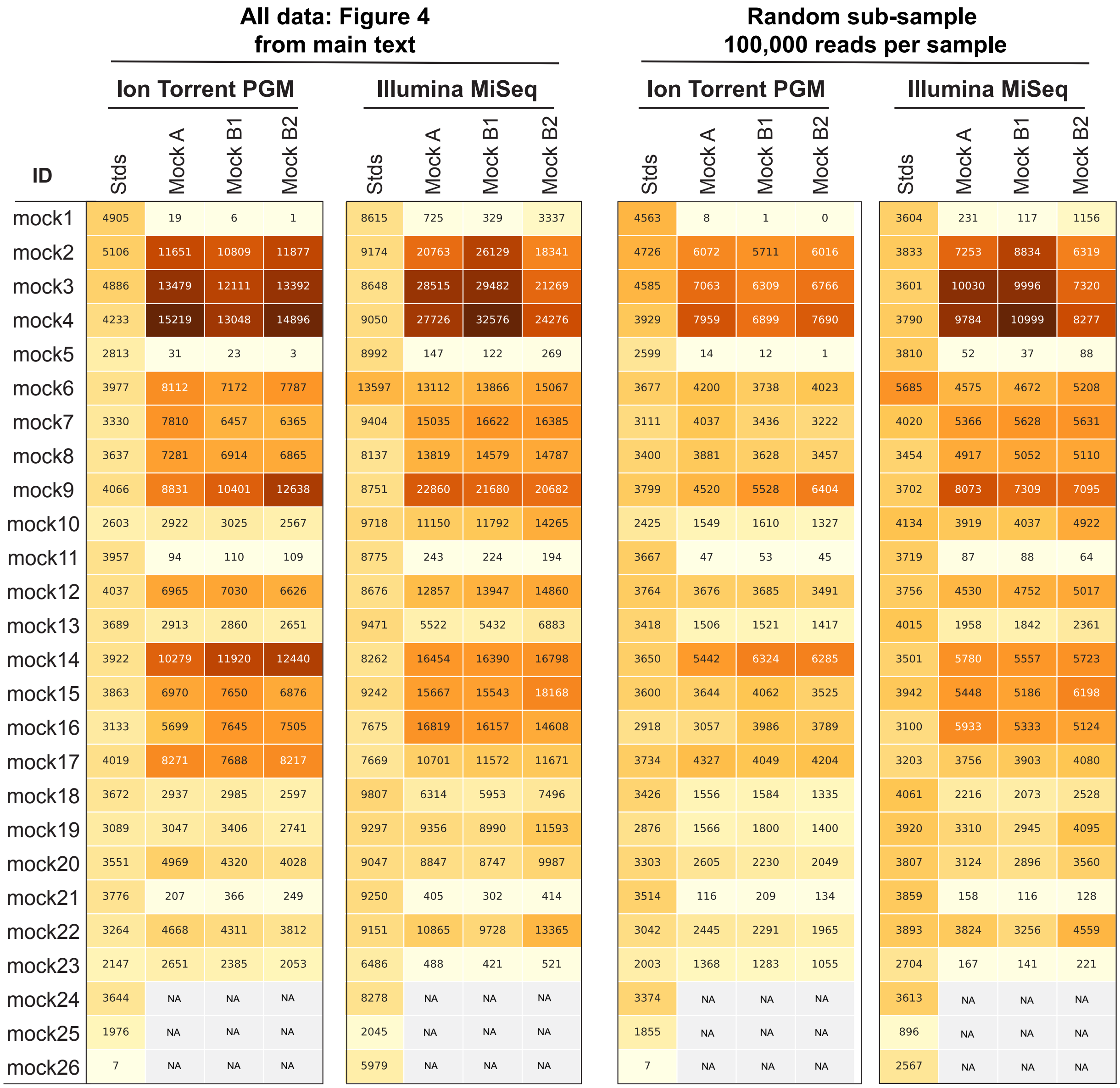
Wu J-Y, Jiang X-T, Jiang Y-X, Lu S-Y, Zou F, Zhou H-W. 2010. Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *BMC Microbiology* **10**: 255.

Supplemental Figures



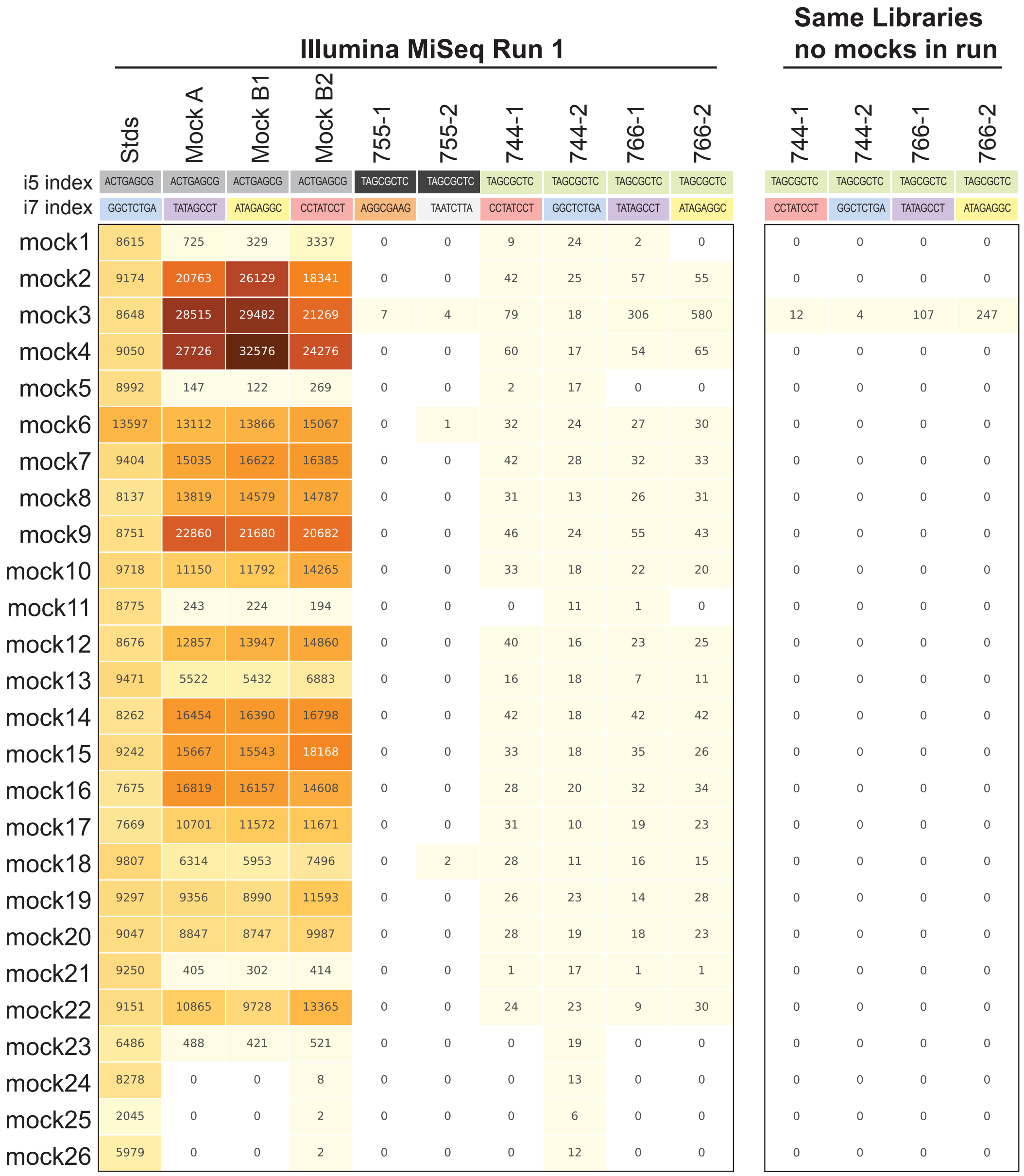
Species	ITS2 Length (bp)	GC Content (%)	Homopolymer stretches > 5	Mock ID	25 Cycles				30 Cycles						37 Cycles								PCR-Stds
					1:20	1:100	1:200	1:2000	1:20	1:100	1:200	1:2000	1:8000	1:32000	1:20	1:100	1:200	1:2000	1:8000	1:32000	1:64000	1:128000	
Phialocephela sp.	237	68.4%	0	mock1	927	1277	895	580	0	0	1	8	0	0	2	0	3	5	1	0	0	0	4583
Ascomycete sp.	238	50.8%	0	mock2	6680	7229	7813	6779	7381	5470	8009	10078	8048	5657	12334	7161	7621	7532	8979	5976	5378	2618	5436
Phialocephala lagerbergii	238	58.8%	0	mock3	8199	7780	8103	6231	9047	6367	10676	9456	8215	6000	15257	7334	8505	7278	9258	5446	6356	2836	4770
Helotiales sp.	239	57.3%	0	mock4	8487	8371	9066	7661	9549	6879	11353	12028	10023	6423	15799	8077	8757	8966	10896	6915	7148	3153	4608
Aspergillus sp.	260	65.8%	3	mock5	780	462	502	328	3	4	8	35	3	1	15	50	53	29	9	0	0	2	3369
Bjerkandera adusta	281	51.2%	0	mock6	4511	5064	4986	5718	4808	4239	5464	6951	6038	4322	8496	4999	5319	5888	6859	5293	4565	2559	4327
Laetiporus caribensis	283	52.7%	0	mock7	2848	3757	3446	3619	3206	2588	3101	5421	3950	3698	5910	2700	3663	3404	4160	3100	3810	1525	2918
Trametes gibbosa	288	50.0%	1	mock8	3862	4620	4296	5075	4431	3697	4319	5817	4972	3747	7421	4599	4665	5006	5311	4305	3966	1709	3902
Laetiporus gilbertsonii	290	54.1%	0	mock9	8004	8675	8203	6420	9503	7279	10708	7268	7758	6967	16289	8715	8860	6438	7270	6902	7910	3734	4222
Gloeoporus pannocinctus	292	43.8%	0	mock10	3002	3751	3413	3483	2673	2492	2026	3141	2763	3120	4868	3323	3023	2488	2959	3458	2868	1400	3137
Wolfiporia dilatohypha	293	54.6%	0	mock11	371	318	310	153	314	229	207	145	137	319	610	359	236	110	190	159	222	81	4025
Trichaptum sp.	293	48.1%	0	mock12	4410	4950	4916	5171	4491	3871	4481	5659	5042	3913	7994	4847	5064	4687	5252	4420	4685	2030	4034
Fomitopsis ochracea	295	44.1%	0	mock13	1846	2299	2201	2568	1744	1550	1424	2488	2172	2186	2798	2144	1847	1947	2095	2160	2057	1086	3802
Laetiporus cremeioporus	296	54.7%	0	mock14	6952	7385	7091	6608	7963	6548	9466	7727	7282	6043	14270	7827	8122	7165	7170	7030	8207	3767	4173
Phanerochaete laevis	300	47.7%	1	mock15	5196	6167	5789	5396	5487	4786	4801	5581	5157	5718	9523	5944	5510	4905	5452	5595	5678	2843	4218
Laetiporus cincinnatus	302	54.0%	0	mock16	5999	6456	5775	4654	6118	5093	5982	4773	4680	5337	10877	6397	6059	4262	4697	4592	5494	10885	3433
Punctularia strigosozonata	303	53.1%	0	mock17	4446	4748	4376	4838	4845	3968	5127	5744	4977	4100	8894	5187	5119	5201	5556	4437	4186	1997	3940
Phellinus cinereus	314	49.7%	0	mock18	1809	2121	2006	2306	1776	1683	1562	2460	1862	1575	3170	2502	1953	2391	2111	2219	1846	826	3724
Junghuhnia lacera	315	43.8%	1	mock19	1883	2011	1833	1944	1572	1493	1114	1452	1456	1873	2622	2055	1646	1299	1568	1667	1617	1267	1985
Leptoporus mollis	315	45.4%	3	mock20	2521	3098	2692	3812	2551	2375	1891	3732	3096	2681	4538	3189	2816	3524	2991	3028	2411	1331	3523
Leptoporus mollis 2	315	45.1%	1	mock21	359	299	305	311	179	177	115	269	166	170	245	513	233	215	166	271	180	57	3622
Mortierellales sp.	353	45.0%	0	mock22	2789	3334	2662	3704	2728	2648	2327	3589	3069	2526	5307	3693	2814	4355	3120	2846	3220	1251	3618
Laetiporus persicinus	379	51.2%	2	mock23	1858	2077	1845	2270	1742	1567	1395	1852	1640	1680	3273	2332	1549	2700	1736	1351	1578	704	2582
Penicillium sp.	260	66.2%	1	mock24	142	103	106	81	0	0	0	3	0	0	0	0	0	0	0	0	0	0	3509
Verticillium sp.	291	64.6%	1	mock25	0	2	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1746
Wolfiporia cocos	548	59.7%	0	mock26	3	0	0	1	0	0	0	1	1	0	1	1	1	2	1	1	0	0	1

Figure S1. Read abundances do not correlate with actual abundances even when DNA concentration is high and PCR cycles are low. Creating libraries of the equimolar BioMock community by varying PCR cycles and DNA concentrations for sequencing on the Ion Torrent PGM did little to change read abundances. However, these data are consistent with traditional recommendations to use as few PCR cycles as possible during library prep.



	All data: Figure 4 from main text								Random sub-sample 100,000 reads per sample							
	Ion Torrent PGM				Illumina MiSeq				Ion Torrent PGM				Illumina MiSeq			
ID	Stds	Mock A	Mock B1	Mock B2	Stds	Mock A	Mock B1	Mock B2	Stds	Mock A	Mock B1	Mock B2	Stds	Mock A	Mock B1	Mock B2
mock1	4905	19	6	1	8615	725	329	3337	4563	8	1	0	3604	231	117	1156
mock2	5106	11651	10809	11877	9174	20763	26129	18341	4726	6072	5711	6016	3833	7253	8834	6319
mock3	4886	13479	12111	13392	8648	28515	29482	21269	4585	7063	6309	6766	3601	10030	9996	7320
mock4	4233	15219	13048	14896	9050	27726	32576	24276	3929	7959	6899	7690	3790	9784	10999	8277
mock5	2813	31	23	3	8992	147	122	269	2599	14	12	1	3810	52	37	88
mock6	3977	8112	7172	7787	13597	13112	13866	15067	3677	4200	3738	4023	5685	4575	4672	5208
mock7	3330	7810	6457	6365	9404	15035	16622	16385	3111	4037	3436	3222	4020	5366	5628	5631
mock8	3637	7281	6914	6865	8137	13819	14579	14787	3400	3881	3628	3457	3454	4917	5052	5110
mock9	4066	8831	10401	12638	8751	22860	21680	20682	3799	4520	5528	6404	3702	8073	7309	7095
mock10	2603	2922	3025	2567	9718	11150	11792	14265	2425	1549	1610	1327	4134	3919	4037	4922
mock11	3957	94	110	109	8775	243	224	194	3667	47	53	45	3719	87	88	64
mock12	4037	6965	7030	6626	8676	12857	13947	14860	3764	3676	3685	3491	3756	4530	4752	5017
mock13	3689	2913	2860	2651	9471	5522	5432	6883	3418	1506	1521	1417	4015	1958	1842	2361
mock14	3922	10279	11920	12440	8262	16454	16390	16798	3650	5442	6324	6285	3501	5780	5557	5723
mock15	3863	6970	7650	6876	9242	15667	15543	18168	3600	3644	4062	3525	3942	5448	5186	6198
mock16	3133	5699	7645	7505	7675	16819	16157	14608	2918	3057	3986	3789	3100	5933	5333	5124
mock17	4019	8271	7688	8217	7669	10701	11572	11671	3734	4327	4049	4204	3203	3756	3903	4080
mock18	3672	2937	2985	2597	9807	6314	5953	7496	3426	1556	1584	1335	4061	2216	2073	2528
mock19	3089	3047	3406	2741	9297	9356	8990	11593	2876	1566	1800	1400	3920	3310	2945	4095
mock20	3551	4969	4320	4028	9047	8847	8747	9987	3303	2605	2230	2049	3807	3124	2896	3560
mock21	3776	207	366	249	9250	405	302	414	3514	116	209	134	3859	158	116	128
mock22	3264	4668	4311	3812	9151	10865	9728	13365	3042	2445	2291	1965	3893	3824	3256	4559
mock23	2147	2651	2385	2053	6486	488	421	521	2003	1368	1283	1055	2704	167	141	221
mock24	3644	NA	NA	NA	8278	NA	NA	NA	3374	NA	NA	NA	3613	NA	NA	NA
mock25	1976	NA	NA	NA	2045	NA	NA	NA	1855	NA	NA	NA	896	NA	NA	NA
mock26	7	NA	NA	NA	5979	NA	NA	NA	7	NA	NA	NA	2567	NA	NA	NA

Figure S2. Random subsampling reads for each sample does not improve accuracy of read abundances. Each sample was randomly sub-sampled to 100,000 reads using 'amptk sample' and then reads were mapped to the BioMock community. Chi-square test for each of these BioMock samples was significant (p < 0.001), indicating the read abundances are not equally distributed.



	Illumina MiSeq Run 1										Same Libraries no mocks in run			
	Stds	Mock A	Mock B1	Mock B2	755-1	755-2	744-1	744-2	766-1	766-2	744-1	744-2	766-1	766-2
i5 index	ACTGAGCG	ACTGAGCG	ACTGAGCG	ACTGAGCG	TAGCGCTC	TAGCGCTC	TAGCGCTC	TAGCGCTC	TAGCGCTC	TAGCGCTC	TAGCGCTC	TAGCGCTC	TAGCGCTC	TAGCGCTC
i7 index	GGCTCTGA	TATAGCCT	ATAGAGGC	CCTATCCT	AGGCGAAG	TAATCTTA	CCTATCCT	GGCTCTGA	TATAGCCT	ATAGAGGC	CCTATCCT	GGCTCTGA	TATAGCCT	ATAGAGGC
mock1	8615	725	329	3337	0	0	9	24	2	0	0	0	0	0
mock2	9174	20763	26129	18341	0	0	42	25	57	55	0	0	0	0
mock3	8648	28515	29482	21269	7	4	79	18	306	580	12	4	107	247
mock4	9050	27726	32576	24276	0	0	60	17	54	65	0	0	0	0
mock5	8992	147	122	269	0	0	2	17	0	0	0	0	0	0
mock6	13597	13112	13866	15067	0	1	32	24	27	30	0	0	0	0
mock7	9404	15035	16622	16385	0	0	42	28	32	33	0	0	0	0
mock8	8137	13819	14579	14787	0	0	31	13	26	31	0	0	0	0
mock9	8751	22860	21680	20682	0	0	46	24	55	43	0	0	0	0
mock10	9718	11150	11792	14265	0	0	33	18	22	20	0	0	0	0
mock11	8775	243	224	194	0	0	0	11	1	0	0	0	0	0
mock12	8676	12857	13947	14860	0	0	40	16	23	25	0	0	0	0
mock13	9471	5522	5432	6883	0	0	16	18	7	11	0	0	0	0
mock14	8262	16454	16390	16798	0	0	42	18	42	42	0	0	0	0
mock15	9242	15667	15543	18168	0	0	33	18	35	26	0	0	0	0
mock16	7675	16819	16157	14608	0	0	28	20	32	34	0	0	0	0
mock17	7669	10701	11572	11671	0	0	31	10	19	23	0	0	0	0
mock18	9807	6314	5953	7496	0	2	28	11	16	15	0	0	0	0
mock19	9297	9356	8990	11593	0	0	26	23	14	28	0	0	0	0
mock20	9047	8847	8747	9987	0	0	28	19	18	23	0	0	0	0
mock21	9250	405	302	414	0	0	1	17	1	1	0	0	0	0
mock22	9151	10865	9728	13365	0	0	24	23	9	30	0	0	0	0
mock23	6486	488	421	521	0	0	0	19	0	0	0	0	0	0
mock24	8278	0	0	8	0	0	0	13	0	0	0	0	0	0
mock25	2045	0	0	2	0	0	0	6	0	0	0	0	0	0
mock26	5979	0	0	2	0	0	0	12	0	0	0	0	0	0

Figure S3. Index-bleed on Illumina MiSeq occurs during the sequencing run and is not a result of contamination. Sequencing the BioMock on Illumina MiSeq resulted in elevated levels of apparent index-bleed during our first run. To rule out that this was a result of contamination during library prep/cleanup, the same libraries were sequenced on a second run in the absence of any of the BioMock samples. The index-bleed discovered in the first run then disappeared, however, one of the BioMock members (mock3) was actually found in these environmental samples.

Supplemental Tables

Table S1. Cultures from the CFMR culture collection used to construct the BioMock community.

Species	Voucher ID	Mock ID	ITS2 Length	% GC	GenBank Accession
Phialocephala fusca	FP-170182	mock1	237	68.35%	KU668953
Ascomycete sp.	FP-170235	mock2	238	50.84%	KU668968
Phialocephala lagerbergii	FP-170134	mock3	238	58.82%	KU668951
Helotiales sp	RF10JR	mock4	239	57.32%	KU668958
Aspergillus candidus	RF1JR	mock5	260	65.77%	KU668969
Bjerkandera adusta	RF3JR	mock6	281	51.25%	KU668970
Laetiporus caribensis	GDL-1	mock7	283	52.65%	KU668960
Trametes gibbosa	RF5JR	mock8	288	50.00%	KU668971
Laetiporus gilbertsonii	OR-2	mock9	290	54.14%	KU668967
Gloeporus pannocinctus	MR5-1	mock10	292	43.84%	KU668965
Wolfiporia dilatohypha	FP-72162	mock11	293	54.61%	KU668959
Schizopora sp.	FP-170198	mock12	293	48.12%	KU668955
Fomitopsis ochracea	FP-170231	mock13	295	44.07%	KU668957
Laetiporus cermeioporus	L34-2	mock14	296	54.73%	KU668963
Phanerochaete laevis	RF9JR	mock15	300	47.67%	KU668973
Laetiporus cincinnatus	DA-37	mock16	302	53.97%	KU668950
Punctularia strigosozonata	RF7JR	mock17	303	53.14%	KU668972
Phellinus cinereus	IN4-1	mock18	314	49.68%	KU668962
Antrodiella semisupina	MR-3	mock19	315	43.81%	KU668966
Leptoporus mollis	TJV-93-174	mock20	315	45.40%	KU668975
Leptoporus mollis 2	RLG-7163	mock21	315	45.08%	KU668974
Mortierellales sp	FP-170186	mock22	353	45.04%	KU668954
Laetiporus persicinus	HHB-9564	mock23	379	51.19%	KU668961
Penicillium nothofagi	FP-170215	mock24	260	66.15%	KU668956
Metapochonia suchlasporia	FP-170177	mock25	291	64.60%	KU668952
Wolfiporia cocos	MD-275	mock26	548	59.67%	KU668964

Table S2. OTU clustering results using default QIIME pre-processing of reads.

Platform	Clustering method	Reads	Total OTUs	Mock OTUs (n = 12)	Error Rate (mismatches / total)
Ion Torrent PGM (400 bp)	UCLUST	2 562 316	97 175	1 347	3.760%
	USEARCH	2 562 316	9 812	560	4.237%
	SWARM	2 562 316	276 403	225	3.517%
	UPARSE	2 562 316	1 609	82	1.100%
Illumina Miseq (2 x 300)	UCLUST	15 696 636	122 802	528	0.131%
	USEARCH	15 696 636	9 785	545	4.694%
	SWARM	15 696 636	614 133	165	4.447%
	UPARSE	15 696 636	2 483	38	0.077%

Table S3. Expected errors quality trimming removes most errors from Ion Torrent PGM data using 12 member SynMock community.[1]

Method	Aligned reads	Subst. errors	Indel errors	UPARSE OTUs	OTUs (chimera filtered)
No Qual Filter	67 185	0.237%	0.342%	26	21
Cutadapt -q 25	73 092	0.152%	0.222%	28	26
Seqtk (Phred)	75 535	0.204%	0.314%	83	79
Sickle –q 25	71 221	0.098%	0.087%	31	30
Exp. Errors < 1	35 810	0.078%	0.100%	18	14

[1] Total of 78,525 reads from the SynMock Ion Torrent PGM run demuxed with AMPtk.

Table S4. Expected errors quality trimming removes most errors from Illumina MiSeq data using 12 member SynMock community.[1]

Method	Aligned reads	Subst. errors	Indel errors	UPARSE OTUs	OTUs (chimera filtered)
No Qual Filter	1 081 931	0.333%	0.006%	44	27
Cutadapt -q 25	1 148 274	0.253%	0.007%	361	337
Seqtk (Phred)	1 115 657	0.316%	0.007%	173	150
Sickle	1 153 190	0.166%	0.006%	304	285
Exp. Errors < 1	961 458	0.094%	0.006%	45	27

[1] Total of 1,167,662 reads from the SynMock Illumina MiSeq run demuxed with AMPtk.

Table S5. Summary statistics for the HTAS runs used in this study.

Run	Platform	Total Reads	Valid Reads	Num Samples	Range reads per sample	Total UPARSE OTUs	Mock Community	Mock Calculated Error Rate	Index-Bleed
Mock3	Ion Torrent PGM	4,332,502	3,029,824	19	107,416 - 217,372	1,010	BioMock	0.086%	0.033%
Mock4a	Illumina Miseq	5,668,955	5,661,700	20	237,035 - 334,455	1,778	BioMock	0.019%	0.264%
Mock4b	Illumina Miseq	659,738	658,730	4	145,405 - 191,095	477	None	NA	NA
Mock4c	Illumina Miseq	6,103,680	6,096,296	20	221,130 - 392,118	1,625	BioMock	0.020%	0.233%
Mock5	Ion Torrent PGM	4,341,392	2,602,544	21	59,394 - 254,269	927	SynMock	0.099%	0.156%
Mock6	Illumina Miseq	18,005,575	17,979,995	21	623,128 - 1,167,662	2,497	SynMock	0.082%	0.046%