HHS Public Access

Author manuscript

J Immunol Methods. Author manuscript; available in PMC 2016 June 01.

Published in final edited form as:

J Immunol Methods. 2015 June ; 421: 112–121. doi:10.1016/j.jim.2015.04.004.

Mycobiome: Approaches to Analysis of Intestinal Fungi

Jie Tang[#1,2], **Iliyan D. Iliev**[#2,3,*], **Jordan Brown**[1], **David M. Underhill**[#2,3], and **Vincent A. Funari**[#1,2,*]

[1]Genomics Core, Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

[2]Department of Biomedical Sciences, Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

[3]F. Widjaja Foundation, Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

[#] These authors contributed equally to this work.

Abstract

Massively parallel sequencing (MPSS) of bacterial 16S rDNA has been widely used to characterize the microbial makeup of the human and mouse gastrointestinal tract. However, techniques for fungal microbiota (mycobiota) profiling remain relatively under-developed. Compared to 16S profiling, the size and sequence context of the fungal Internal Transcribed Spacer 1 (ITS1), the most common target for mycobiota profiling, are highly variable. Using representative gastrointestinal tract fungi to build a known "mock" library, we examine how this sequence variability affects data quality derived from Illumina Miseq and Ion Torrent PGM sequencing pipelines. Also, while analysis of bacterial 16S profiles is facilitated by the presence of high-quality well-accepted databases of bacterial 16S sequences, such an accepted database has not yet emerged to facilitate fungal ITS sequence characterization, and we observe that redundant and inconsistent ITS1 sequence representation in publically available fungal reference databases affect quantitation and annotation of species in the gut. To address this problem, we have constructed a manually curated reference database optimized for annotation of gastrointestinal fungi. This Targeted Host-associated Fungi (THF) database contains 1,817 ITS1 sequences representing sequence diversity in genera previously identified in human and mouse gut. We observe that this database consistently outperforms three common ITS database alternatives on comprehensiveness, taxonomy assignment accuracy and computational efficiency in analyzing sequencing data from the mouse gastrointestinal tract.

[*]Correspondence to: Iliyan D. Iliev, Department of Biomedical Sciences and F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA. Tel: (310)423-3504; Iliyan.Iliev@cshs.org.
Vincent A. Funari, Genomics Core, Cedars-Sinai Medical Center, 8700 Beverly Blvd, Los Angeles, CA 90048, USA. Tel: (310)423-2544; Vincent.Funari@cshs.org

Keywords

Fungal ITS; Internal transcribed spacer; Database; Microbiome; Gut; Amplicon sequencing

1. Introduction

Advances in high-throughput sequencing technologies have driven enormous progress recently in our understanding of the prevalence and diversity of the microbial communities associated with nearly all of our mucosal surfaces (Clemente et al., 2012; Methé et al., 2012; Knights et al., 2013). This “microbiota” influences processes ranging from digestion to behavior and is increasingly being appreciated as a key component of our physiology that directly influences health and disease. While most studies of the gastrointestinal microbiota focus on commensal and pathogenic bacteria, we and others have recently begun to explore the fungi that are also present in these communities (Scupham et al., 2006; Iliev et al., 2012; Dollive et al., 2013; Hoffmann et al., 2013).

There are two main approaches to microbial community profiling: metagenomics and targeted amplicon sequencing. Whole genome sequencing (WGS) of a community, also known as metagenomics, is a powerful, but expensive and computationally challenging approach (Morgan and Huttenhower, 2014). It is not suitable for evaluation of fungi in the microbiota. First, in communities dominated by bacteria (or other host or microbial DNA sources), extremely deep and costly sequencing is required to uncover fungi. Second, while many reference genomes are available for the types of bacteria typically found in the mammalian microbiome, there are significantly fewer complete fungal genomes yet available (Underhill and Iliev, 2014). In the targeted amplicon approach to evaluating bacterial microbiota, highly variable regions of the 16S rDNA are amplified by PCR and sequenced. Comparison of these sequences to well-established reference databases of 16S sequences allows identification of the bacteria from which they are derived. This approach allows for very deep analysis of a community and is highly cost-effective.

While the targeted amplicon approach as described only gives information about bacteria in a community, it can be modified to investigate fungi as well. The most common approach is to amplify the fungal “internal transcribed spacer” (ITS) regions. This is in part because of the historical broad depth of sequencing for this target available in public databases (Schoch et al., 2012): as of February 2015, there are 434K fungal ITS1 (or 182K ITS2) sequences in GenBank. There are two ITS regions, ITS1 is located between the 18S and 5.8S genes while ITS2 is located between 5.8S and 25S genes. Since these ITS regions are not part of the conserved transcribed regions of the structural ribosomal RNAs, they are highly divergent between fungi, often being sufficiently different to classify fungi to the species level. In eukaryotes and bacteria, the rDNA region typically exists in multiple copies per genome. In fungi, the locus is typically duplicated 100-200 times, thus caution must be used when trying to derive quantitative comparisons between various species in mixed populations by this approach.

Compared to bacterial 16S sequencing, fungal ITS sequencing presents the investigator with two new problems. First, unlike bacterial 16S amplicons, fungal ITS sequences from

different species can differ widely in size and sequence content (Abarenkov et al., 2010; Santamaria et al., 2012). ITS fragments isolated from mouse or human feces typically vary in length between 100-550 base pairs, and it is not yet clear how the variable lengths affect recovery of sequences through the various steps of sequencing on high-throughput platforms. Further, with the recent retirement of Roche 454 chemistry, the prior "gold standard" platform for microbiome analyses, the new platform of choice for mycobiome applications needs to be evaluated. The two most common third generation "benchtop" sequencing platform are MiSeq (Illumina) and Ion Torrent (Life Technologies). Both platforms' limitations have been well-characterized but not so well recognized. For instance, Illumina's sequencing-by-synthesis approach is more sensitive to sequence complexity while the Ion Torrent semiconductor sequencing method has a higher overall error rate in homopolymer regions (Loman et al., 2012). Given the greater complexity of ITS sequences in size and sequence compared to 16S, it is necessary to elucidate any 3rd generation sequencing platform specific bias that affects fungal profiling. In this manuscript, we offer a side-by-side comparison of how these two new technologies handle the variable lengths of fungal ITS sequences.

Second, unlike bacterial 16S sequences, there is no well-established database of ITS sequences. Publically available repositories of fungal sequences are replete with redundant sequences containing incomplete and/or incorrect taxonomic assignments. In fact, it has been estimated that as many as 20% of the fungal sequences in the International Nucleotide Sequence Databases (INSD) are incorrectly annotated at the species level (Nilsson et al., 2006). In addition, of the 45,979 entries in the consortium integrative fungal database UNITE (Abarenkov et al., 2010), 12,096 (26.31%) entries cannot be assigned to a family. Finally, fungal taxonomy is enormously complicated. It is very common for sexual (telomorph) and asexual (anamorph) forms of a single fungus to be classified as different taxa, sometimes having different accepted names at the phylum level. In this manuscript, we report producing a new custom, hand-curated database and evaluate the effectiveness of this database compared to several existing public options.

2. Material and Methods

2.1. DNA isolation

Fecal, skin, oral, and lung samples were obtained from 6-10 weeks old female C57BL/6J mice (Jackson Laboratories). A representative fecal sample was selected from a pool of fecal samples characterized in Iliev, et.al. 2012 for further analysis. DNA was isolated as described below. Samples were homogenized and incubated with 200U of Lyticase (Sigma) for 30 min. Sample pellets were resuspend in 800 ul of DNA stool stabilizer (Stra Tec) and transferred to tubes containing 1:3 ratio of 0.1mm and 0.5mm beads. The tubes were placed on Omni Bead Ruptor 12 Homogenizer (Omni International) and subjected to two 1 minute cycles of beating. Samples were then heated at 95°C for 10 minutes, placed on ice for 1 minute, and centrifuged at 12,000 g for 1 minute. The supernatant were then processed using QIAamp DNA Mini Kit (Qiagen) according to the manufacturer's instructions. Individual strains including, *Candida tropicalis* (ATCC 750) and *Saccharomyces cerevisiae* (ATCC 4007164) were obtained from the American Type Culture Collection, while

Saccharomycopsis fibuligera was isolated from murine feces and identified by rDNA sequencing (Iliev, et.al. 2012). DNA from individual strains was isolated using the method described above.

2.2. Sample Quantity Assessment

Four steps (sample quality control, PCR, library preparation, and library quality control) are needed for optimizing ITS amplification and recovery (**Supplementary Fig. 1A**). To assess quantitatively fungal, bacterial and host DNA, DNA concentration in samples were first quantified with the Qubit Fluorometer 2.0 dsDNA High Sensitivity Assay (Life Technologies). Microbiota and host genomic content were quantified by further assayed using 2 μL of each sample using Pan-*Aspergillus*/*Candida* hydrolysis probes (Qiagen, cat# BPCL00359A) for fungi, Pan-Bacteria (cat# BPCL00360A) for bacteria, and mouse genomic HBB1 hydrolysis probes (cat# BPCL00541A) for host. Each reaction used 1 μL of hydrolysis probes/primers and 10 μL of Microbial qPCR Mastermix supplied with the assay kits. Microbial DNA-free water (Qiagen) was added to each reaction for a final volume of 20 μL. The qPCR cycling conditions were set according to the manufacturer's instructions.

2.3. ITS PCR

Fungal ITS1 amplicons were typically generated in 20 uL PCR reactions using 1 ng of each sample with 35 cycles using Phusion DNA Polymerase (New England BioLabs) at an annealing temperature of 56.1°C using the primers ITS1F (CTTGGTCATTTAGAGGAAGTAA) and ITS2 (GCTGCGTTCTTCATCGATGC)(Gardes and Bruns, 1993). While this generally yields sufficient amplification of ITS1 targets (and did so for samples used in this study), we sometimes identified samples in which fungal content is underrepresented in 1 ng of DNA (**Supplementary Fig. 1B**). In such cases, the cycles of amplification and amount of sample added for each ITS amplicon PCR must be adjusted according to the cycle threshold (Ct) of the microbiota/host qPCR in order to normalize all sample reactions to the same amount of fungal template. Resultant ITS amplicons were purified using Agencourt AmPure Magnetic Beads (Beckman Coulter), resuspended in 20 μL of nuclease-free water, and quantified using a Qubit fluorometer. Amplicons were further qualified using the DNA 1000 assay on the Agilent Bioanalyzer (Agilent Technologies).

2.4. Generation of controlled mock microbial communities

To evaluate the two different sequencing platforms mock microbial communities were constructed from common model microbes. First, a mock bacterial community was constructed with amplicons that have similar size from the 16S V1-V3 variable regions of three bacterial strains (*Streptococcus agalactiae*, *Listeria monocytogenes*, and *Staphylococcus aureus)*. Amplicons from the three strains were combined in equimolar ratios. Similarly, a mock fungal community was constructed by mixing known concentrations of PCR amplified ITS1 regions of different sizes from three fungal strains, that are commonly found in gut, in the following ratio: 33% *Candida tropicalis*, 33% *Saccharomycopsis fibuligera*, and 33% *Saccharomyces cerevisiae*. Adapter ligation and

multiplexing, nick repair, and library enrichment were performed according to manufacturer's protocol (See Methods 2.6).

2.5. Fragmentation of mock fungal community amplicons for Ion Torrent Personal Genome Machine (PGM) sequencing

We built sequencing libraries from enzymatically fragmented amplicons in order to evaluate the size-dependent enrichment of some species on the PGM. From each of the three fungal strains , 100 ng of ITS amplicon was sheared for 20 minutes using Ion Shear reagents (Ion Xpress Plus Fragment Library preparation kit,Life Technologies). Adapter ligation and multi-plexing, nick repair, and library enrichment were performed according to manufacturer's protocol (See Methods 2.6).

2.6. Library preparation

Illumina paired-end adapters with unique indexes were ligated to 100 ng of ITS1 amplicons using the TruSeq DNA Nano Sample Preparation Kit (Illumina). Library enrichment was performed with 10 cycles of PCR and purified using Agencourt Ampure Magnetic Beads (Beckman). Successful ligation of Illumina adapters results in 120 base pair shifts in the size of ITS amplicons and are confirmed using Agilent DNA 1000 Bioanalyzer assays (**Supplementary Fig. 1C**). Qubit fluorometric assay (Life Technologies) is used for final quantitation and libraries were then pooled at equimolar concentrations.

Afterwards, 100 ng of pooled ITS amplicons were used to generate Ion Torrent sequencing libraries using the Ion Xpress Library Kit (Life Technologies). Adapters and primers were diluted 1:10 to accommodate for the low input into the library preparation. Libraries were barcoded using Ion Xpress Barcode Adapter Kit (Life Technologies) to allow for multiplexed sequencing. Libraries were quantified and qualified in the same manner as the Illumina libraries.

2.7. Sequencing

Illumina sequencing was performed on the Illumina MiSeq platform (Illumina) which uses a paired-end sequencing mode, giving the possibility to obtain relatively long overlapping reads (250 base pairs) with less than 1% error rates. Fungal ITS1 libraries were clonally amplified directly onto adapters on Single-End (SE) flow cells and sequenced with standard Illumina sequencing primers according to manufacturer instructions. Raw data processing and run de-multiplexing was performed using on-instrument analytics as per manufacture recommendations.

For Ion Torrent sequencing, libraries were conjugated to Ion Sphere Particles® (ISPs) using the Ion Torrent OneTouch® system (Life Technologies). Each template strand is then amplified, creating a clonal community of template molecules on each ISP. Enrichment of template-positive ISPs was performed using the Ion Torrent Enrichment System, according to manufacturer instructions. Sequencing was performed on an Ion Torrent PGM (Life Technologies) using 400-base pair sequencing chemistry and PGM 314 v2 chips according to manufacturer instructions. Base-calling and run de-multiplexing was performed using on-instrument analytics.

2.8. Data Processing

For Illumina MiSeq sequencing reads, raw FASTQ data were filtered to enrich for high quality reads including removing the adapter sequence by cutadapt v1.4.1 (Martin, 2011), removing any reads that do not contain the proximal primer sequence or any reads containing a single N (unknown base). Sequence reads are then quality trimmed using a custom script (split_libraries_fastq.py) from QIIME v1.6 (Caporaso et al., 2010), by truncating reads not having an average quality score of 20 (Q20) over a 3 base pair sliding window and then removing trimmed reads having less than 75% of their original length. For Ion Torrent semiconductor sequencing, raw FASTQ reads not containing the designed proximal ITS primers (>2nt mismatches) or any reads containing a single N were precluded from further analysis.

2.9. Calculation of relative abundance in mock community controls

Three technical replicates of the mock control library were sequenced independently on each platform (Ion Torrent and MiSeq) in different sequencing runs as described above. The high quality MiSeq and Ion Torrent reads were aligned with BLAST v2.2.22 to ITS1-containing reference sequences (GenBank accession numbers EU288196, JX094776 and U10409) for the three strains. Relative abundances for each fungal strain in one of the three technical replicates were calculated before and after library prep using Agilent 2100 expert software version b.02.08.sI648(sr2) (Santa Clara, Ca) as well as after sequencing by counting the reads aligned to each fungal strain. Relative bias was calculated by subtracting the observed value from the expected abundance and plotted in Figure 2C. The total bias for each platform is reported as the sum of the library prep and sequencing bias in Fig 2D. A two-tailed T-Test was used to determine significant differences between PGM and MiSeq platforms.

2.10. Database Performance

2.10.1 Findley, RTL and UNITE database—The 2,593 reference sequences from the Findley database (Findley et al., 2013) were downloaded from http://www.mothur.org/w/images/2/20/Findley_ITS_database.zip. The 23,456 accession numbers from the RTL database (Findley et al., 2013; Schoch et al., 2014) and the matching sequence for each accession numbers were downloaded from GenBank using a custom python script. The most recent release (QIIME release version 6) of UNITE database (Abarenkov et al., 2010) was downloaded from http://unite.ut.ee/.

2.10.2 Targeted Host-associated Fungi (THF) Database—We generated a list of genera that could be identified upon aligning ITS sequences from mouse and human samples with UNITE sequences. We reasoned that our custom-curated database should contain ITS sequences representing the diversity in these genera. We identified publications in the mycology literature reporting phylogenetic trees for these genera and reasoned that the ITS sequences/fungal strains demonstrated to be related in these trees should reliably represent the genera with minimal misidentifications. Where possible, we tried to use phylogenetic trees based on ITS sequences specifically, as this should best represent the relationships between the ITS sequences we generate. Where ITS phylogenies were not readily available,

we relied on phylogenies based on other genomic regions (e.g. D1/D2 Large Subunit rDNA) and manually identified ITS sequences from the reference strains used for inclusion. We therefore seeded our database with these ITS sequences and include reference to the publications supporting these name assignments. Where convenient based on the literature being used, we expanded the database to include sequence diversity from related genera. Where anamorphs and telomorphs could be identified, we selected one name to represent the organism, trying to choose the name that is more commonly used. The present version of this database (THF 1.0) includes 1,817 sequences linked to species names and to references supporting their phylogenetic relationships to similar fungi. As a result, THF database provides reliable taxonomy at every taxonomic level down from kingdom to species for every entry in the database. The THF database can be downloaded freely at: https://risccweb.csmc.edu/microbiome/thf/.

2.10.3 Fungal Database comparisons—We evaluated the UNITE, Findley, RTL and THF databases for species representation, complete ITS1 sequences, taxonomic assignment, and computational efficiency. In brief, the processed high-quality reads of one representative fecal sample (~373K reads) from our published dataset (Iliev, et.al. 2012) and newly sequenced three murine samples from mouth, lung, skin and feces on MiSeq platform (ca. 8.2M reads) were aligned to the respective reference database, using BLAST v2.2.22 in the QIIME v1.6 wrapper with an identity percentage ≥97% for operational taxonomic unit (OTU) picking on a high performance cluster consisting of 100 nodes with 1.8 GHz Dual Core Opteron processors and 4 Gigabytes of RAM per core. OTUs are compiled into genera using a custom Perl script.

We identified ITS1 and ITS2 regions in the four databases using an ITS sequence search tool, ITSx version 1.0.9 (Bengtsson-Palme et al., 2013) which is based on a hidden Markov model (HMM). A reference sequence with the presence of both 18S end and 5.8S start was considered to have complete ITS1 region; similarly, the presence of both 5.8 end and 28S start was required to identify complete ITS2 region. The reference sequences were grouped further based on the completeness of ITS1 and ITS2 regions.

3. Results

3.1. Sequencing approaches

The types of fragments amplified during PCR of bacterial 16S regions and fungal ITS regions are substantially different from each other. 16S amplicons recovered after PCR from a variety of mouse fecal DNA samples are relatively uniform in size (≈369 bases, **Fig. 1A**). In contrast, the abundances and sizes of ITS amplicons from these same samples are highly diverse, ranging from ≈170 to ≈550 bases (**Fig. 1B**). These patterns reveal substantial mouse-to-mouse diversity in fungal content, but also illustrate the technical challenge encountered when trying to sequence each fungal species in a complex mixture with equal efficiency.

To investigate the potential for bias in sequencing library preparation methods and sequencing technologies when characterizing DNA sequences of different lengths, we generated two “model” libraries made up of either 16S amplicons from three different

bacteria or ITS amplicons from three different fungi. Equimolar concentrations of 16S amplicons from *Streptococcus agalactiae* (366 base pairs), *Listeria monocytog*enes (369 base pairs), and *Staphylococcus aureus* (360 base pairs) were mixed (**Fig. 1C**). Similarly, equimolar concentrations of ITS amplicons from *Candida tropicalis* (250 base pairs), *Saccharomycopsis fibuligera* (320 base pairs) and *Saccharomyces cerevisiae* (445 base pairs) were mixed (**Fig. 1D**). These fungi, all common in mouse fecal samples, replicate the diversity of fungal ITS sequence lengths, and the model libraries allow us to quantify biases inherent in the sequencing analyses.

When the bacterial mock 16S amplicon community was made into libraries and sequenced using Illimina MiSeq protocols, we found that the semi-quantitative detection of the different component amplicons was highly reproducible, and 16S sequences from each of the model bacteria were detected with approximately the expected ratios (**Fig. 2A, B**). This is consistent with MiSeq being selected by many investigators as a robust sequencing platform for targeted 16S microbial community profiling (Caporaso et al., 2012).

We next directly compared two current sequencing platforms (Illumina MiSeq and Ion Torrent PGM) with the model fungal ITS library to determine how they would handle the variably-sized component fragments. When analyzed on the Illumina MiSeq, we observed modest sequence recovery bias in favor of smaller fragments (**Fig. 2C, D**)($p<0.029$). Further analysis of the process revealed that the majority of the bias present comes from the sequencing platform and that the Illumina library preparation method itself contributed minimally (**Fig. 2C**). In contrast, when analyzed on the Ion Torrent PGM, we observed very strong bias towards smaller fragments, nearly loosing detection of the 450 base pair *S. cerevisiae* sequences altogether (**Fig. 2C, D**) ($p<0.0051$). Further analysis of the process revealed that both the library preparation method and the sequencing platform contributed to this bias (**Fig. 2C**).

We reasoned that if the poor performance of the Ion Torrent PGM was truly due to being given ITS fragments of differing sizes, that a method that normalized the sizes of the fragments would reduce the observed bias. We therefore performed shearing of equimolar pools of mock community amplicons (**Fig. 3A**) before library preparation and sequencing and found that the profile of fungal species in the mock community is in good agreement (50%, 31% and 19%) with expected abundances and independent of original amplicon sizes (**Fig. 3B**). Thus, while it may be possible to develop protocols that facilitate use of the PGM, additional steps and validation will be required compared to Illumina MiSeq.

3.2. Database approaches

Once fungal ITS sequences have been generated, the next challenge faced by investigators is how to identify the fungi from which the sequences originate. In contrast to bacterial 16S analysis where well-established, commonly-accepted databases of sequences are available, fungal ITS analysis is comparably undeveloped. As noted above, errors in public sequence annotation, the evolving nature of our understanding of the relationships between different fungi, and heterogeneity in fungal taxonomy make databases difficult to work with. The UNITE database (http://unite.ut.ee/repository.php) is probably the most commonly-used

fungal ITS database. It mirrors the International Nucleotide Sequence Database Collaboration and attempts to cover all public fungal ITS sequences available (currently 45,976 sequences in its “non-redundant” version). Findley et al. created a custom database to support their mycobiota studies at the NIH National Human Genome Research Institute (Findley et al., 2013). In this database, the investigators applied a bioinformatic approach to mining all of the ITS sequences available in Genbank and distilling it to a nonredundant set (23,456 sequences). Some level of manual curation of taxonomic discrepancies and anamorphs/telomorph naming was applied. Taking a different approach, Schoch et al. have sought to bring together a large consortium of mycologists to generate a definitive fungal rDNA database called the RefSeq Targeted Loci (RTL) database. In this database, every sequence included (currently 2,593) has been individually scrutinized for sequence length, quality, and taxonomic identification by scientists with expertise in each type of fungus included.

We analyzed a single representative mouse fecal ITS sequence library (372,546 reads) with these three different currently available public databases, and the results illustrate the challenges (**Fig. 4**). The distribution of fungi detected can depend on the database used. While *Candida* and *Saccharomyces* are detected by all three (albeit to slightly different degrees), other organisms were detected very differently. The abundance of *Candida* and *Saccharomyces* in this sample was confirmed by direct quantitative PCR (Iliev et al., 2012). UNITE identified a significant number of sequences belonging to *Cryptococcus* not observed in the other databases. Interestingly, the Findley et al. database identified a significantly larger fraction of the reads as *Chaetomium*. Many of these reads were annotated as *Humicola* in other databases, upon closer inspection this may be because Findley lacks reference sequences for *Humicola*. UNITE and Findley identified a significant number of reads as “unclassified”, offering no clues as the genera most closely related to the sequences. UNITE identified both *Fusarium* and *Gibberella*, which are anamorph/telomorph names of the same organisms. RTL specifically named all identified fungal genera and no sequences were left “unclassified”. RTL also identified a significant number of sequences as belonging to *Cetraspora*, however upon closer inspection these reads were actually annotated as *Saccharomyces* in the other databases. We discovered the single *Saccharomyces* ITS1 reference sequence in RTL was actually an incomplete ITS1 reference sequence unable to align significantly to every *Saccharomyces* read. In addition, no *Plectosphaerella* was identified using the RTL database; upon closer inspection RTL lacks references for *Plectosphaerella*. Finally, we identified *Gliomastix* reads in all the databases except Findley then discovered there was no *Gliomastix* reference in the Findley database. To address these database challenges, we generated our own database tailored to specifically encompass the sequence diversity found in genera known to be found in gastrointestinal samples and giving them names that represent the relationships between organisms in the gut. We manually identified published sequence-based phylogenetic trees documenting the sequence diversity in genera that we selected based on ongoing analysis of many mouse and human gastrointestinal samples. Anamorph/telomorph naming ambiguities were also manually edited out to restrict sequences to one common name. Using this Targeted Host-associated Fungi (THF) database, we analyzed the same sample used in comparing the other databases (**Fig. 5A**). The database identifies a larger fraction of the sequences as

Saccharomyces. There is enormous sequence diversity and naming ambiguity in the genus *Saccharomyces*, and we believe that these sequences have been more accurately placed with this curated approach. Other genera are specifically named, and no sequences are left "unclassified" or confused in anamorphs/telomorph pairs. Literature references in the database allow deeper interrogation of sequence-based phylogenetic relationships as investigators require it.

While database specificity is a critical component of any analysis as discussed above, such analyses are also sensitive to mapping efficiencies; a database that identifies sequences confidently but that identifies far fewer sequences than an alternative database is of limited use. Databases which are not complete will affect the relative quantitation of fungi in a sample. We therefore compared the abilities of the three public databases and the custom THF database to map and name sequences from a variety of samples to the genus level (**Fig. 5B, Supplementary Figure 2**). THF confidently names as many or more sequences than the other three databases in fungal ITS sequencing datasets from mouse feces as well as several other anatomical locations (**Fig. 5B, Supplementary Figure 2**).

The completeness of ITS1 sequences in databases is crucial for correct mapping and relative quantification of fungal sequences. We compared the sensitivity for the rDNA transcribed strand ("forward orientation", starting with ITS1F primer) and its complementary strand ("reverse orientation", starting with ITS2 primer), respectively. We observed no significant differences in the sequence length distribution profiles of the four databases. However, there is a consistent mapping bias in favor of the transcribed strand in all the four databases (**Fig. 6A**) meaning that the currently-available ITS reference sequences are biased towards reads originating from the 18S region. At the four mouse body sites examined above, the Findley et al. database introduced the lowest mapping bias on the directionality of sequencing, although the results were very similar to UNITE and THF. In contrast, the RTL database has the largest bias in the majority of body sites with up to 3.3 fold mapping preferences towards the forward orientation.

To further characterize the bias in directionality among the four databases, we summarized the completeness of the ITS regions in the reference sequences in each database (**Fig. 6B**). Interestingly, we observed no significant differences in proportions (45.82%, 41.61% and 38.88%) of complete ITS1 references in the RTL, Findley and THF databases, which suggests that the greater mapping bias observed with the RTL database is unlikely due to the majority of ITS1 sequences in RTL being incomplete. Instead this data and our previous observations (e.g. incomplete S*accharomyces cerevisiae*), suggest that RTL may have incomplete ITS1 references for species that are more commonly found in gut.

A final consideration in fungal ITS database usage is the computational burden demanded by the different databases. The reference database sequence alignments are computationally intensive, and the running times are directly proportional to database size. We compared the analysis time required by the four databases for a full MiSeq dataset (8.2M reads in total). The UNITE database (45,976 entries) demanded 18.25 hours of processing time on our computer cluster composed of 200+ processors and 800 GB of RAM. The Findley et al. database is similarly large (23,456 sequences) and required 12.2 hours. The RTL and THF

databases are smaller (2,593 and 1,817 sequences respectively) and thus were processed much faster (4.1 and 3.6 hours respectively). Thus the smaller, more focused THF database may enable laboratories without easy access to large computer clusters to process large fungal ITS sequencing datasets.

4. Discussion

We describe here a combined sequencing and analysis approach for analyzing intestinal mycobiomes. With the recent retirement of Roche 454 chemistry, newer sequencing platforms will need to be utilized for fungal ITS sequencing-based community analysis. We evaluated the strengths and weaknesses of the Illumina MiSeq and Ion Torrent PGM platforms for this type of analysis. While Next Generation sequencing error profiles are well described (Loman et al., 2012), we observed a strongly distorted fungal community profile using Ion Torrent's PGM that could be the result of the failure to amplify longer amplicons on ISP's® which are limited to fragments <50-400bp. Highly inefficient clonal amplification on an ISP® could result in a sequences with such poor fidelity, they would be unreadable. In the future, isothermal amplification for the ion semiconductor sequencing platforms such as the PGM should minimize bias in clonal amplification, enabling fragments longer than 400 base pairs to be sequenced efficiently or using unbiased approaches that amplify whole genomes before ITS PCR (e.g. whole genome linear amplification) may further decrease any size-dependent amplification bias. Although, MiSeq introduces a small bias towards identification of fungi with smaller ITS1 regions, we anticipate that continuing improved chemistries will further reduce this bias. Importantly, the small bias is very reproducible, allowing common microbial comparative analysis biomarker approaches like LEfSe (Segata et al., 2011). Finally, the MiSeq platform enables identification of even the largest ITS1 fragments. Therefore, we have settled for now on the MiSeq platform. In the future, the development of a rich mock fungal community as a standard might help assess additional bias (e.g. PCR primer efficiency and sequence content) that is not covered here and may help standardize the field.

Given that as many as 20% of ITS sequences in GenBank are thought to be incorrectly annotated (Nilsson et al., 2006) and that many sequences are simply annotated as “fungi”, databases that are based on GenBank are prone to incorrect taxonomic assignment. Incorrect taxonomic assignments can cause sequences from very similar or identical organisms to be identified as very different, and this effect can be compounded by division of like sequences between sexual and asexual names of an organism. This may lead to decreased significance in identification of clinically important fungal species or decreased significance in identification of clinically important changes in fungal species. Thus, we see a need for highly-referenced, well-curated database tools for grouping host-associated fungi by ITS sequencing. The THF database generated here performs at least as well as existing databases (at least in the analysis of the mouse gastrointestinal tract samples it has been thus far tailored to), is thoroughly referenced, is directly linked to taxonomical information, and is hand curated. It also enables faster computation than alternatives. We expect that ongoing modification and expansion of the tool will improve its utility and function. Of course, THF is not designed to replace existing fungal discovery databases. Like all analysis tools for

high-throughput ITS sequencing, findings derived from THF also need to be confirmed by follow-up with alternative tools and approaches.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by grants from the US National Institutes of Health (DK098310 to I.D.I. and DK093426 to D.M.U.), as well as the Crohn's and Colitis Foundation of America.

References

Abarenkov K, Henrik Nilsson R, Larsson KH, Alexander IJ, Eberhardt U, Erland S, Hoiland K, Kjoller R, Larsson E, Pennanen T, Sen R, Taylor AF, Tedersoo L, Ursing BM, Vralstad T, Liimatainen K, Peintner U, Koljalg U. The UNITE database for molecular identification of fungi--recent updates and future perspectives. The New phytologist. 2010; 186:281–5. [PubMed: 20409185]

Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, De Wit P, Sánchez-García M, Ebersberger I, de Sousa F, Amend A, Jumpponen A, Unterseher M, Kristiansson E, Abarenkov K, Bertrand YJK, Sanli K, Eriksson KM, Vik U, Veldre V, Nilsson RH. Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Meth. Ecol. Evol. 2013; 4:914–919.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010; 7:335–6. [PubMed: 20383131]

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME journal. 2012; 6:1621–4. [PubMed: 22402401]

Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012; 148:1258–70. [PubMed: 22424233]

Dollive S, Chen YY, Grunberg S, Bittinger K, Hoffmann C, Vandivier L, Cuff C, Lewis JD, Wu GD, Bushman FD. Fungi of the murine gut: episodic variation and proliferation during antibiotic treatment. PLoS One. 2013; 8:e71806. [PubMed: 23977147]

Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos E, Park M, Program NIHISCCS, Kong HH, Segre JA. Topographic diversity of fungal and bacterial communities in human skin. Nature. 2013; 498:367–70. [PubMed: 23698366]

Gardes M, Bruns TD. ITS primers with enhanced specificity for basidiomycetes--application to the identification of mycorrhizae and rusts. Mol Ecol. 1993; 2:113–8. [PubMed: 8180733]

Hoffmann C, Dollive S, Grunberg S, Chen J, Li H, Wu GD, Lewis JD, Bushman FD. Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. PLoS One. 2013; 8:e66019. [PubMed: 23799070]

Iliev ID, Funari VA, Taylor KD, Nguyen Q, Reyes CN, Strom SP, Brown J, Becker CA, Fleshner PR, Dubinsky M, Rotter JI, Wang HL, McGovern DP, Brown GD, Underhill DM. Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. Science. 2012; 336:1314–7. [PubMed: 22674328]

Knights D, Lassen KG, Xavier RJ. Advances in inflammatory bowel disease pathogenesis: linking host genetics and the microbiome. Gut. 2013; 62:1505–10. [PubMed: 24037875]

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology. 2012; 30:434–9.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011:17.

Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, Gevers D, Petrosino JF, Abubucker S, Badger JH, Chinwalla AT, Earl AM, FitzGerald MG, Fulton RS, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayyaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi JV, Brooks P, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PS, Chen I-MA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Dunne WM Jr, Durkin AS, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney L, Foster L, Francesco VD, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, et al. A framework for human microbiome research. Nature. 2012; 486:215–21. [PubMed: 22699610]

Morgan XC, Huttenhower C. Meta'omic analytic techniques for studying the intestinal microbiome. Gastroenterology. 2014; 146:1437–1448. e1. [PubMed: 24486053]

Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson KH, Koljalg U. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PLoS One. 2006; 1:e59. [PubMed: 17183689]

Santamaria M, Fosso B, Consiglio A, De Caro G, Grillo G, Licciulli F, Liuni S, Marzano M, Alonso-Alemany D, Valiente G, Pesole G. Reference databases for taxonomic assignment in metagenomics. Briefings in bioinformatics. 2012; 13:682–95. [PubMed: 22786784]

Schoch CL, Robbertse B, Robert V, Vu D, Cardinali G, Irinyi L, Meyer W, Nilsson RH, Hughes K, Miller AN, Kirk PM, Abarenkov K, Aime MC, Ariyawansa HA, Bidartondo M, Boekhout T, Buyck B, Cai Q, Chen J, Crespo A, Crous PW, Damm U, De Beer ZW, Dentinger BT, Divakar PK, Duenas M, Feau N, Fliegerova K, Garcia MA, Ge ZW, Griffith GW, Groenewald JZ, Groenewald M, Grube M, Gryzenhout M, Gueidan C, Guo L, Hambleton S, Hamelin R, Hansen K, Hofstetter V, Hong SB, Houbraken J, Hyde KD, Inderbitzin P, Johnston PR, Karunarathna SC, Koljalg U, Kovacs GM, Kraichak E, Krizsan K, Kurtzman CP, Larsson KH, Leavitt S, Letcher PM, Liimatainen K, Liu JK, Lodge DJ, Luangsa-ard JJ, Lumbsch HT, Maharachchikumbura SS, Manamgoda D, Martin MP, Minnis AM, Moncalvo JM, Mule G, Nakasone KK, Niskanen T, Olariaga I, Papp T, Petkovits T, Pino-Bodas R, Powell MJ, Raja HA, Redecker D, Sarmiento-Ramirez JM, Seifert KA, Shrestha B, Stenroos S, Stielow B, Suh SO, Tanaka K, Tedersoo L, Telleria MT, Udayanga D, Untereiner WA, Dieguez Uribeondo J, Subbarao KV, Vagvolgyi C, Visagie C, Voigt K, Walker DM, Weir BS, Weiss M, Wijayawardene NN, Wingfield MJ, Xu JP, Yang ZL, Zhang N, Zhuang WY, et al. Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. Database : the journal of biological databases and curation. 2014 2014.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding C, Fungal Barcoding Consortium Author, L. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:6241–6. [PubMed: 22454494]

Scupham AJ, Presley LL, Wei B, Bent E, Griffith N, McPherson M, Zhu F, Oluwadara O, Rao N, Braun J, Borneman J. Abundant and diverse fungal microbiota in the murine intestine. Appl. Environ. Microbiol. 2006; 72:793–801. [PubMed: 16391120]

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biol. 2011; 12:R60. [PubMed: 21702898]

Underhill DM, Iliev ID. The mycobiota: interactions between commensal fungi and the host immune system. Nature reviews. Immunology. 2014; 14:405–16. [PubMed: 24854590]

Highlights

- Evaluation of two current sequencing technologies for assessing fungal ITS1 sequences of different lengths.
- Development of a custom, hand-curated database and evaluation of its effectiveness compared to several existing public databases.

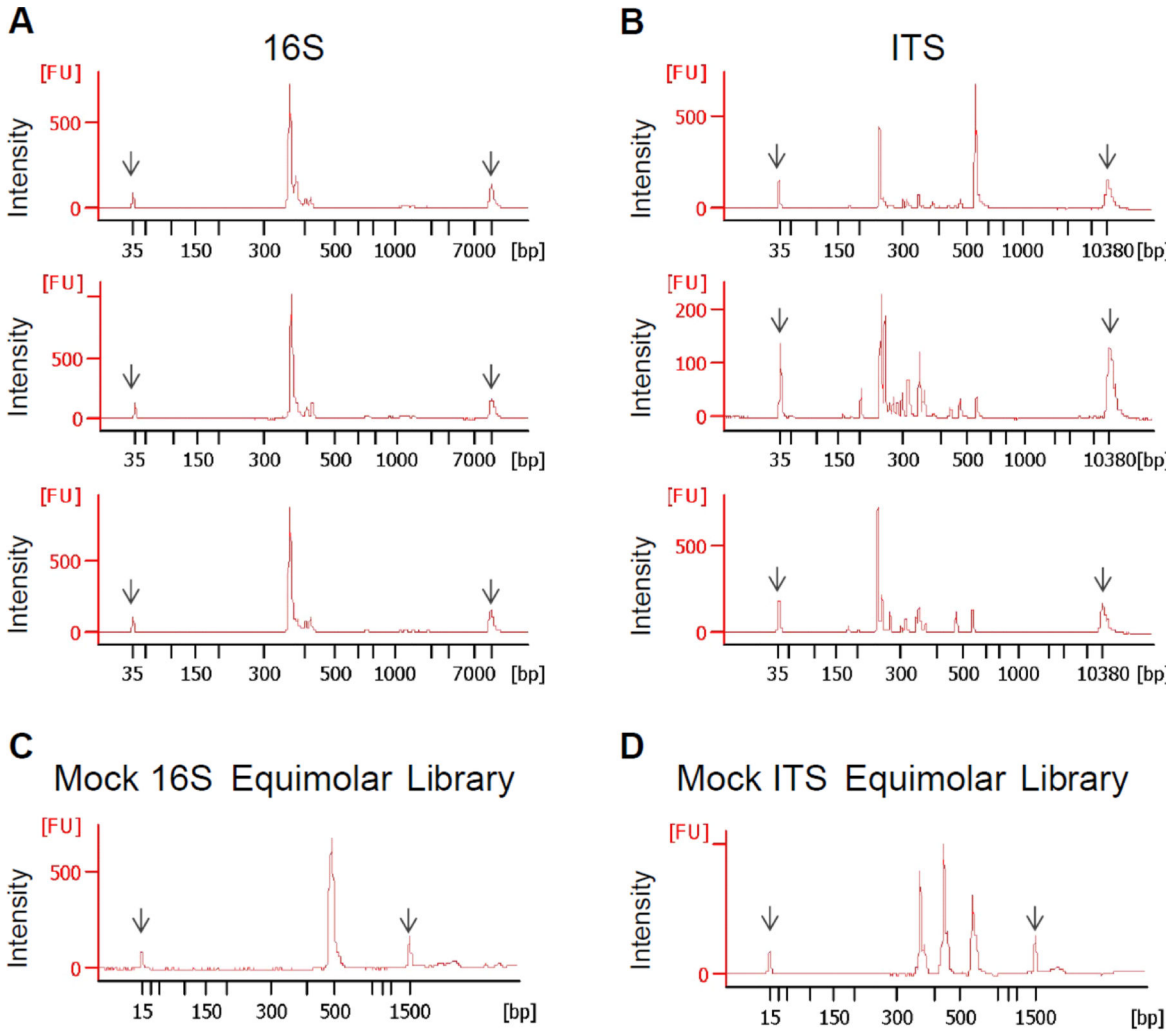

Figure 1. Fungal ITS amplicons from different species vary in size and content while bacterial 16S amplicons are similar

DNA was isolated from mouse fecal samples and from respective bacterial or fungal strains. 16S and ITS regions were amplified by PCR, and amplicon profiles were analyzed using an Agilent Bioanalyzer. A) Representative Agilent profiles of bacterial 16S V1-V3 amplicons from three different fecal samples. B) Representative Agilent profiles of ITS amplicons produced from the three different fecal samples C) Agilent profile of a bacterial 16S "mock community" library produced by pooling equimolar concentrations of 16S V1-V3 amplicons derived from *Streptococcus agalactiae*, *Listeria monocytogenes*, and *Staphylococcus aureus*. D) Agilent profile of ITS "mock community" produced by pooling equimolar concentrations of ITS1 amplicons from *Candida tropicalis* (250 base pairs), *Saccharomycopsis fibuligera* (350 base pairs), and *Saccharomyces cerevisiae* (450 base pairs).

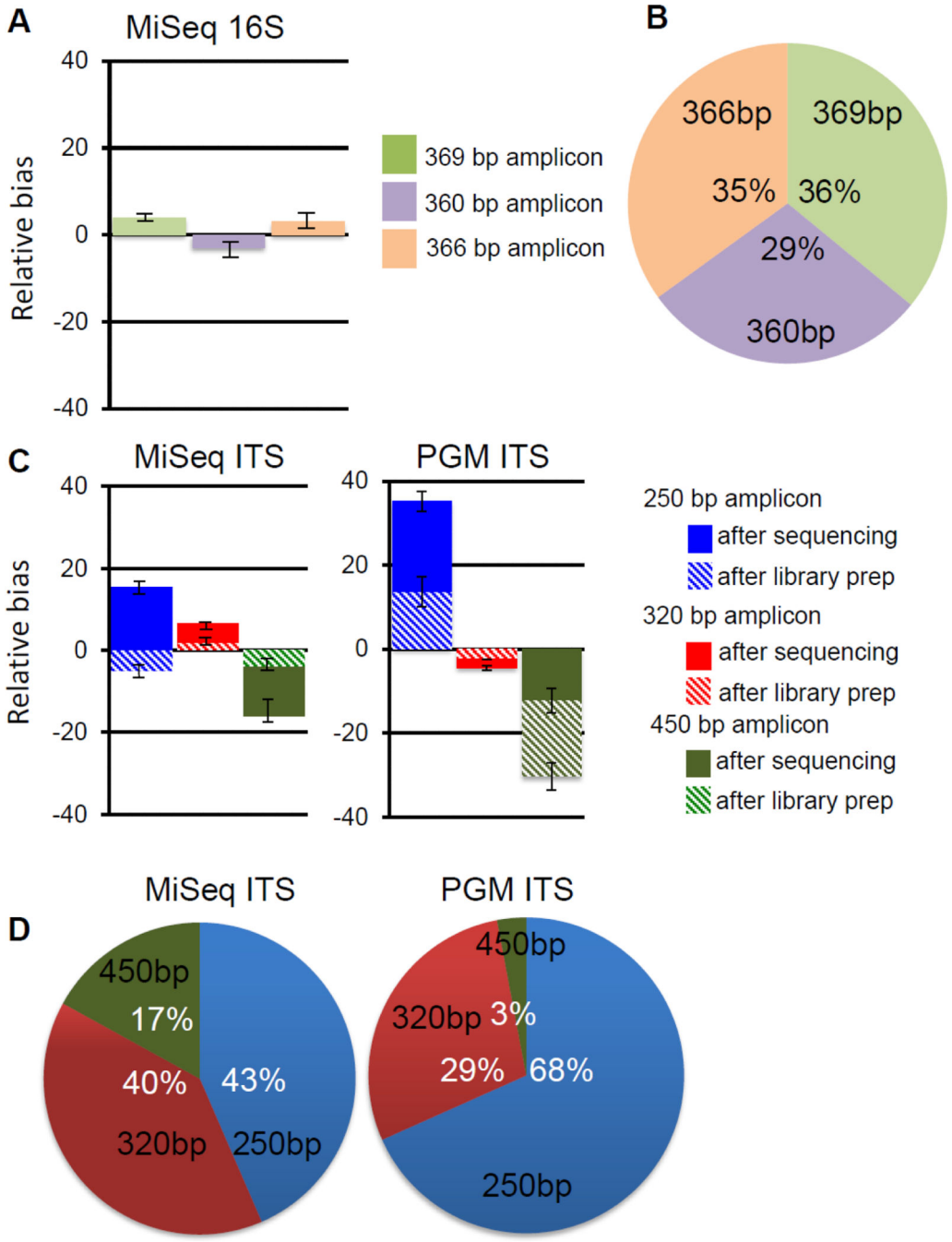

Figure 2. Sequencing platform and corresponding library preparation biases

The model 16S community was processed into libraries and sequenced on the Illumina MiSeq platform. The observed versus predicted sequence biases for each of the 3 sequences was calculated (A), and the distribution is shown (B). The model fungal ITS1 community was processed into libraries and sequenced on the Illumina MiSeq platform and the Ion Torrent PGM platform. The observed versus predicted sequence biases for each of the 3 sequences was calculated (C), and the distribution is shown (D).

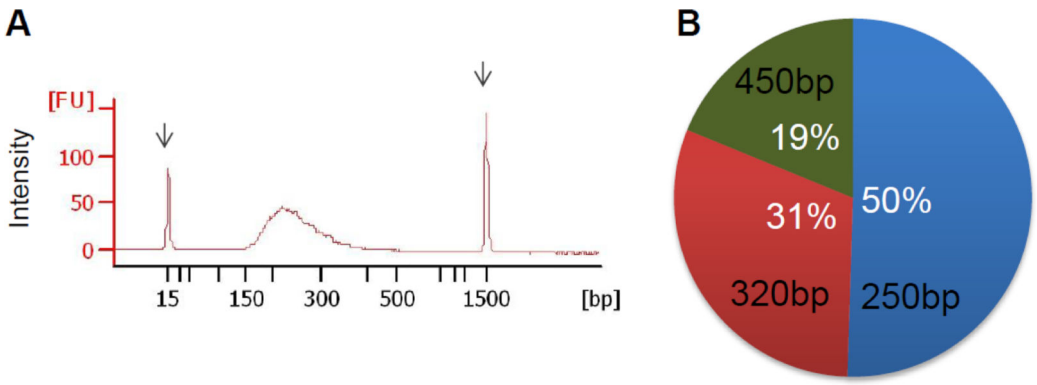

Figure 3. Fragmentation of ITS amplicons before sequencing decreases the bias during PGM sequencing

A) Agilent profile of an ITS mock community sequencing library after amplicon fragmentation. B) Fraction of reads mapping to reference sequences for all three fungal strains after sequencing of a fragmented library on the Ion Torrent PGM sequencing platform.

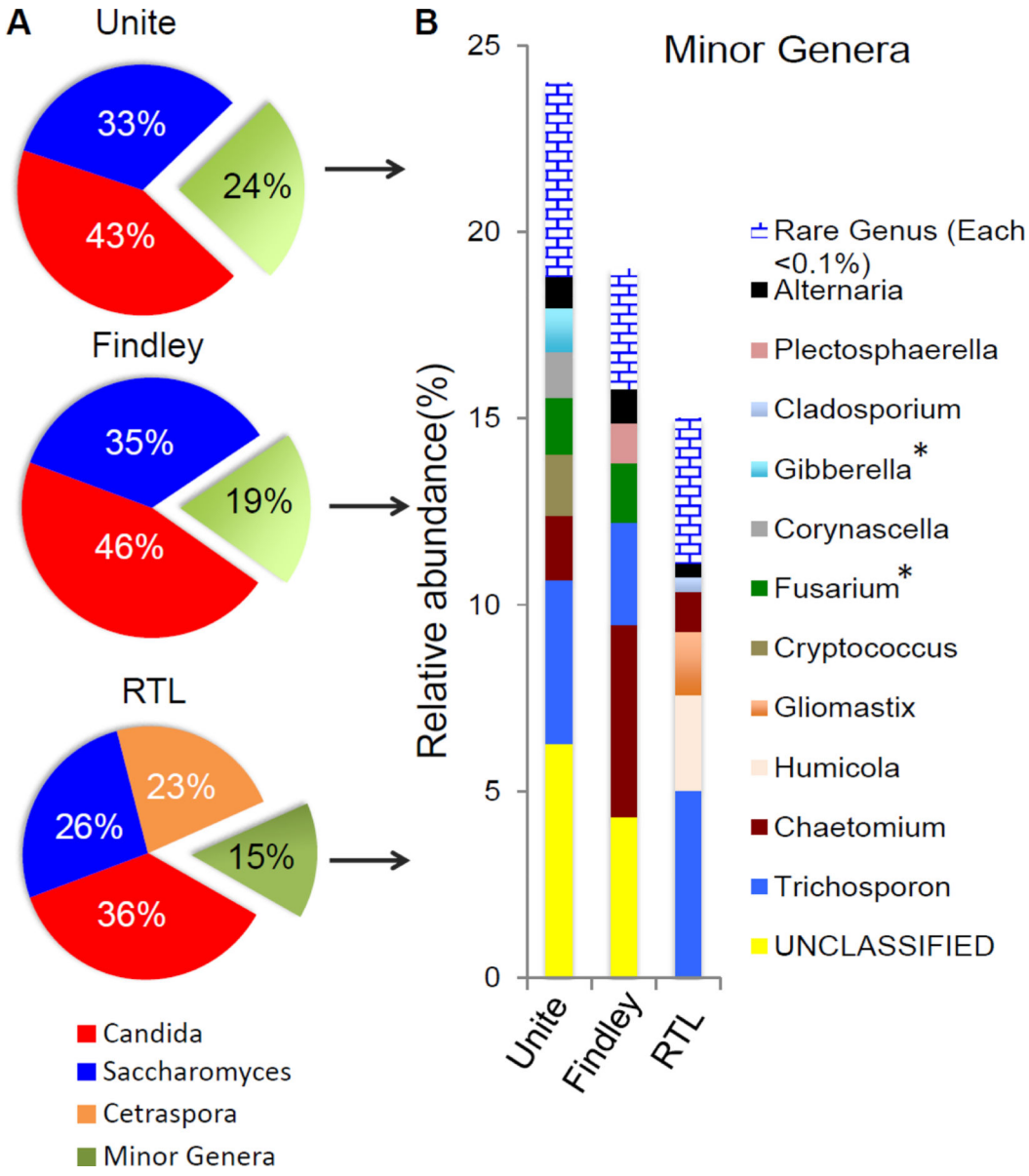

Figure 4. Representative fungal genus profiles in mouse fecal material using three different public fungal ITS databases

Major and minor fungal genera profiles identified in a representative murine fecal sample are shown. Major genera represent more than 20% of the identified sequences, while minor genera each represent less than 10% of the identified sequences. “Unclassified” genera were summed and illustrated (yellow) and very rare genera (each <0.1% of total reads) were summed and illustrated (blue brick). *Denotes genera related to each other by sexual morphotype (telomorph/anamorph).

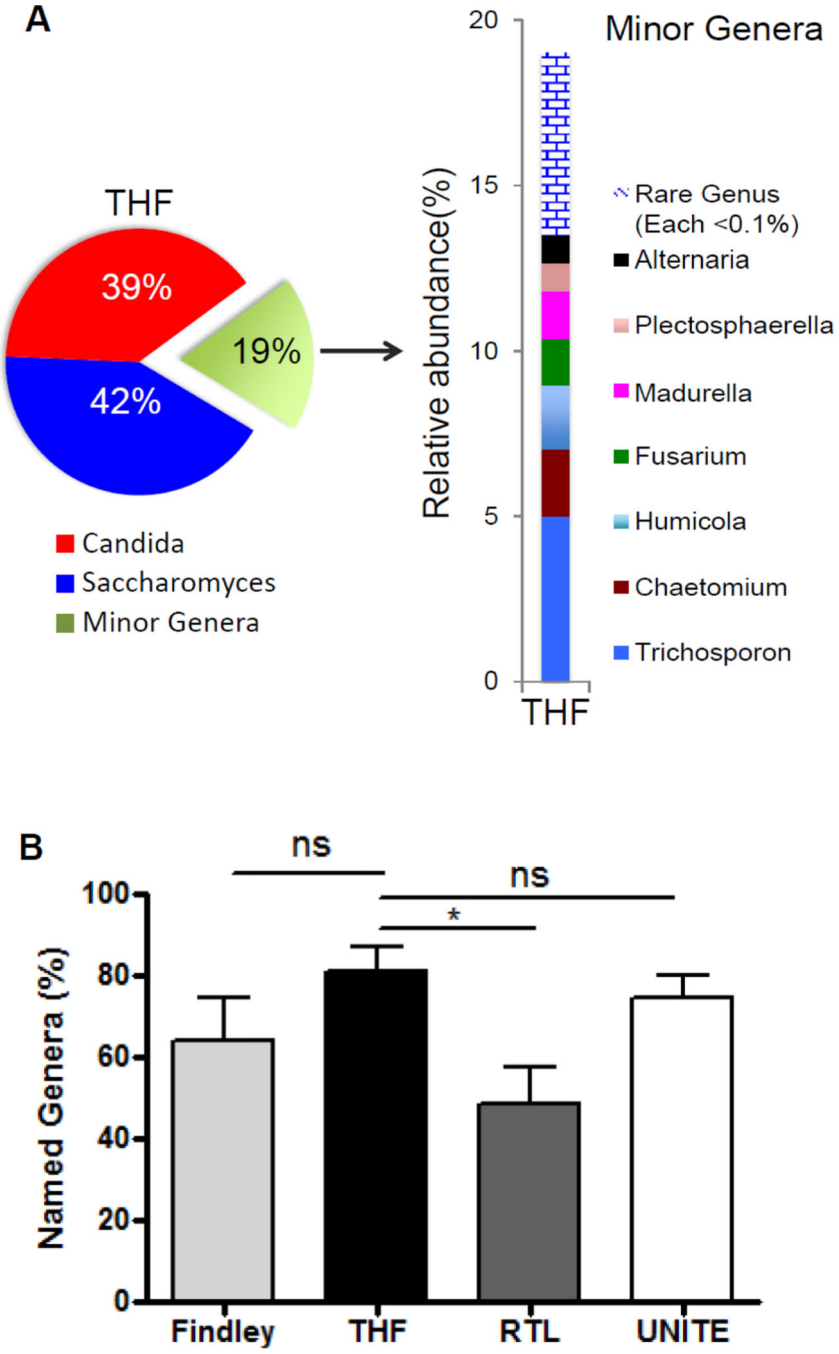

Figure 5. Fungal genus profile in mouse fecal material identified using the custom targeted host-associated fungi (THF) database

A) The genus profile generated by analyzing the same representative sample as in figure 4 with the custom THF database. As in figure 4, major genera represent more than 20% of the identified sequences, while each minor genera represent less than 10% of the identified sequences. Rare genera (each <0.1% of total reads) were summed and illustrated (blue brick). B) All four databases were used to analyze ITS amplicons recovered from three different murine fecal samples. The percentages of the sequences that could be identified to the genera level are illustrated. $P < 0.05$ by Student's t test; n.s., not significant.

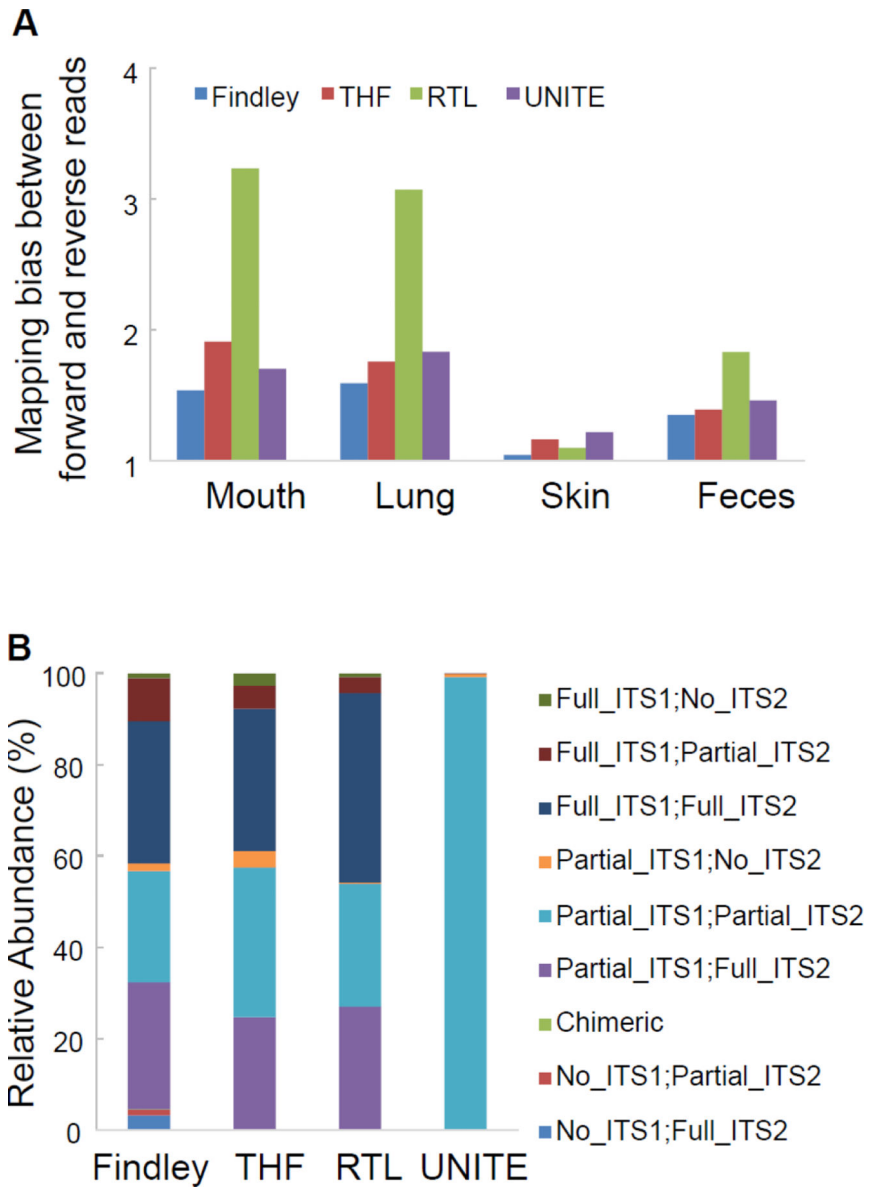

Figure 6. Evaluation of biases in the sequences comprising the four databases

A) Mapping efficacy among four databases using forward and reverse ITS amplicon reads. The mapping percentages of forward and reverse ITS reads were calculated for each database using a representative mouse sample from each of four different body sites as indicated. The ratio of sequences matching the forward strand versus the sequences matching the reverse strand was calculated and displayed. All libraries show some bias for the forward strand. B) ITS sequence completeness was evaluated in the four databases as described in Materials and Methods.