8-24-2015

Fungal Communities in Oral Health and Disease

Amanda K. Dupuy
University of Connecticut - Storrs, amanda.dupuy@uconn.edu

Fungal Communities in Oral Health and Disease

Amanda K. Dupuy, PhD

University of Connecticut, [2015]

With advances in, and cost reduction of next generation sequencing technologies, assessing the presence of microbes in host niches in healthy and diseased states has become a more feasible task. However, these studies are often only limited to bacterial characterization, ignoring other important community members such as fungi, viruses, and protists. The research presented here begins to fill this gap by creating a roadmap for fungal community analysis. With fungal outbreaks on the rise, it is essential that the fungal kingdom is included in future microbiome analyses to gain a more comprehensive understanding of commensal species, how they maintain a healthy niche, and their impact on acquired diseases.

One particularly at-risk group of immune compromised patients are those undergoing chemotherapy. Approximately 40% of such patients develop a debilitating side effect known as oral mucositis, which is complicated by severe pain, inability to eat and speak, and severe bacterial and fungal infections. The research in this dissertation focuses on three aims necessary for answering the question of how fungal genera are implicated in oral mucositis. First, we present a roadmap from sample processing to data analysis, describing challenges and solutions for characterizing fungal communities in any human-health related metagenomics study. Second, we address the healthy fungal mycobiome of saliva, providing evidence for new and existing members of the oral niche, while assessing the temporal variability in community composition in a healthy state. Third, we characterize the oral genera during chemotherapy in a longitudinal study of cancer patients, and document their changes during the progression and development of oral mucositis. Revealing and meeting the challenges associated with fungal metagenomic analysis by means of initial hand curation will pave way for development of new,

much needed library preparation and bioinformatics tools. But above all, pinpointing community trends for susceptible subjects will ultimately provide unprecedented insight for implementation of prophylactic measures in cancer patients.

Fungal Communities in Oral Health and Disease

Amanda K. Dupuy

B.S., University of Central Florida, **[2009]**

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

[2015]

APPROVAL PAGE

Doctor of Philosophy Dissertation

Fungal Communities in Oral Health and Disease

Presented by

Amanda K. Dupuy, B.S.

Major Advisor_____
Linda D. Strausbaugh

Associate Advisor_____
Patricia Diaz

Associate Advisor_____
Rachel O'Neill

Associate Advisor_____
Anna Dongari-Bagtzoglou

Associate Advisor_____
Michael O'Neill

University of Connecticut
[2015]

ACKNOWLEDGMENTS

Thank you first and foremost to my advisor, Linda, for giving me the opportunity to work with you in your lab. There aren't enough words to tell you how privileged I am to be your (last!) graduate student. I feel extremely grateful for the time and energy you put into me to help me become a great *woman* scientist. You taught me that asking for help is not a sign of weakness, that everything takes longer than you think it will, to let the data speak for itself, and that it is important to make time for fun in life. I will never forget our scientific musings about the complexities of fungi, lab meetings at Willington Pizza and the Clam Digger, and our "never-ending" laptop parades. From the very beginning you took a genuine interest in me, professionally and personally, and showed me what being a great mentor, teacher, and friend is all about. You trusted me with the perfect amount of independence to make my own decisions, never letting me feel like I was in it on my own or floundering too long without guidance. Your unwavering support, kindness, and generosity has given me strength through all the most challenging moments I've faced during my time in grad school. Thank you for your patience through the writing of my thesis and for your invaluable suggestions on taking it from good to great. Above all, thank you for believing in me

Thank you to the PSM students who worked closely with me on this project: Lynn, Jason, Marika, Liz, and Joe. Without any of you this thesis would not be possible. You each taught me lessons about what it takes to be a leader and about delegating tasks instead of trying to do them all myself. Thank you for letting me teach you about fungi, for being my cheerleaders, and for sticking through all of the hard work with samples and analysis.

Thank you to my former lab mates and colleagues who taught me the skills to work independently at the bench and with bioinformatics: Bo, Lu, Craig, Tom, and Ranyelle. From ordering primers, to 454 sequencing, to navigating through Linux and command line, I appreciate your direction, assistance, and perspective on doing careful science.

Thank you to my committee, Patricia, Rachel, Anna, Michael, and Jonathan for your open door policies, insights, reference letters, and continued support toward finishing my degree.

Thank you to all of my friends in Beach Hall and in MCB. Your camaraderie made Beach an inviting and awesome place to be. You made days filled with failed experiments bearable. Teaching and working with you was a pleasure. Special thanks to Judy for supporting me and giving me a place to stay during the writing of this thesis.

Thank you to the support staff who helped things flow smoothly for me and for making my concerns a priority: Jess, Anne, Lois, Stephanie, and Nick.

Thank you to my boyfriend, C.J., and to my dearest friends for emotional support, for helping take the edge off, and for giving me a life outside of grad school.

Last but not least thank you to my family. I may be far from home, but your excitement for me has made me want to be the best that I can. I am happy to make you proud.

Table of Contents

Contents

Tables and Figures

Chapter 1. Introduction

1.1 Overview

The ultimate purpose of this study is to provide a basis for understanding how fungi play a role in the development and progression of oral mucositis or other comorbidities arising as a result of chemotherapy. With the mouth being the main entryway into the body, characterizing the fungi here is especially important for understanding the relationship between fungal profiles and their implications in human health and disease.

1.2 Fungi as opportunistic pathogens

Fungi are arguably the most underappreciated and least understood organisms that inhabit planet Earth. They persist in nearly every ecological niche and are crucial in sustaining all other life forms by supplying essential nutrients through the decomposition of organic matter. With species estimates ranging from 1.5 to 5.1 million, fungi are among the most environmentally abundant and diverse eukaryotes[1,2].

Detrimental fungal infections of eukaryotes have occurred throughout history, causing ecosystem and economic turmoil with incidents such as the Irish potato famine of the twentieth century and the recent decline in over 40% of Central American amphibian species[3]. On a much longer time scale, fungi have been postulated to shape the deep evolutionary history of life. Although naturally increased body temperatures tend to prevent fungal colonization in endothermic animals, ectothermic animals must rely on increasing internal temperatures by adjusting their external surroundings. Prehistoric climate changes toward colder temperatures and decreased sunlight may have led to unhindered fungal infection and extinction of the dinosaurs, and ultimately to the rise of mammals[4]. Fungal adaptation to rising global temperatures will require new host defenses and give opportunity for bouts of mammalian extirpation[4].

Fungal threat to human survival is exacerbated by the marked increase of immunocompromised individuals over the last several decades due to the ability of opportunistic fungi to turn from commensal to pathogenic[5]. Investigations of fungal outbreaks at the Centers for Disease Control and Prevention have increased from 1-2 per year in 1990 to 3-6 per year in 2015[6].Most commonly, these outbreaks are caused by contamination in medications or environmental transmission. People with lowered defense systems are the most susceptible to such infections. Immunosuppression, whether induced or acquired, has led to changes in fungi from non-invasive commensals to dangerous pathogens. Emerging infectious diseases caused by invasive species of *Fusarium*, *Scedosporium*, *Trichoderma*, *Paecilomyces*, *Dactylaria*, *Wangiella*, *Cladophialophora*, *Rhizopus*, *Cunninghamella*, and *Mucor* have joined the ranks of those caused by the more well-known human pathogens of *Candida*, *Aspergillus,* and *Cryptococcus* species[7].

Several recent outbreaks have been due to medicinal contamination. A recent review of 21st century fungal outbreaks by the CDC in 2012 of *Fusarium* sp. and *Bipolaris* sp. were caused by contamination of medications used for visualizing vitreous tissues during virectomies (Brilliant Blue G and triamcinolone). As a result, these products caused 47 patients to acquire endophthalmitis, with the majority losing vision. Another 2012 outbreak, now the deadliest to date, was due to methylprednisolone acetate, used to treat joint swelling, arthritis, and severe allergic reactions. These injections affected 752 patients of the potentially 14,000 that were exposed across 23 states. Of those affected, there were onsets of meningitis, arachnoiditis, or spinal abscesses, and a total of 64 deaths[6].

Environmental exposure has also caused unexpected transmission of pathogenic fungi. In 2008 and 2009 the CDC also reported that a *Rhizopus* species was isolated from hospital linens in New Orleans. Of the five children infected, all of them died from cutaneous mucormycosis. A species of *Apophysomyces* was also identified in a patient injured by debris in a tornado. Whole

genome sequencing revealed that three isolates were the cause of the necrotizing cutaneous mucormycoses in 13 patients, giving further evidence to the cause being environmental in nature rather than as acquired as a single source infector[6].

In nature, fungi exhibit a vast potential for mutualistic relationships due to their ability to adapt to new surroundings. However, other organisms must evolve to protect themselves from fungal infection, as fungi do not rely solely on survival of their hosts[3]. Fungi thrive in soil and on decaying matter. It is in these environments that they have evolved a great deal of their virulence factors against amoebae, bacteria, and other threatening microorganisms[8]. The notable human pathogen, *Cryptococcus neoformans,* provides a well-understood illustration of the linkages between adaptation, virulence, and unintentional consequences for vulnerable hosts. For example, nonvirulent strains of *Cryptococcus neoformans* are easily eliminated due to ingestion by other environmental occupants such as amoebae and nematodes. But with pathogenicity intact, the fungus is able to replicate inside of its captor and induce cytotoxicity due to its protective capsule. The capsule is made mostly of glucuronoxylomannan (GXM), which increases in size in the presence of phospholipids released by cells in the surrounding environment, granting protection from reactive oxygen species[9]. The capsule also provides a mechanism against dehydration for environments with low humidity. Coincidentally, the evolution of the capsule for protection in its natural environment gave *C. neoformans* the ability to cause disease in unintentional hosts, which use macrophages to employ defensive strategies similar to *C. neoformans's* natural predators. *C. neoformans* infections have been found in many mammalian species including domestic pets, dolphins, and sheep[10].

In the 1970s, *C. neoformans* infection incidences were approximately 1 in a million per year in the U.S. They have risen markedly since then because of the spread of HIV, and account for 13-44% of HIV patient deaths from sub-Saharan Africa, implicating *C. neoformans* as an opportunistic pathogen[11]. Immunocompetent hosts may also be infected with *C. neoformans*,

especially in tropical climates, but those that are immunocompromised have a much higher rate of infection. Infections are generally caused by inhalation of Cryptococcus spores and lead to three diseases including pulmonary cryptococcosis, cryptococcal meningoencephalitis, and cutaneous cryptococcosis. Cryptococcal meningoencephalitis, the most deadly cryptococcal disease, causes 10-30% of HIV related deaths worldwide. Spores remaining undetected in the lungs and are able to disseminate and target the central nervous system if gone untreated or the host's immune system becomes compromised. As the infection spreads to the brain, symptoms include headache, fever, vision or hearing loss, cognitive impairment, and hydrocephalus[11].

1.3 Fungi in oral mucositis

Understanding how fungi become pathogenic is especially important in populations that undergo induced immune deficiencies. Many cancer patients receiving radiation and chemotherapy treatments experience the debilitating side effect of oral mucositis. Oral mucositis (OM) is the manifestation of lesions in the mouth due to damage of mucosal epithelial cells[12]. The breakdown of this important immunological barrier oftentimes leads to severe pain, poor nutrition, and microbial infections[12]. The onset and progression of OM cannot currently be predicted, but it is clear that oral microbiota play a large role in sustaining oral health[13].

OM is commonly experienced by all groups of patients undergoing chemo- and radiation therapies. Nearly 100% of patients irradiated for head and neck cancers and over 50% of patients receiving chemotherapy for breast cancer have been diagnosed with this complication[12]**.** OM occurs in five complex stages after administration of chemotherapeutics, generally over the course of two weeks: 1) initiation of direct damage to mucosa due to therapeutics; 2) primary damage response by the immune system;3) amplification of the immune response; 4) ulceration; and 5) spontaneous healing[14].

Of those afflicted with OM, approximately 39% suffer from oral fungal infections during treatment, with *Candida* spp. as prime pathogens[15]. It is commonly assumed that all forms of candidiasis associated with oral mucositis are caused by *C. albicans,* but frequently, physicians do not perform tests for species level identification. Commonly, patients are encouraged to manage mucositis with increased oral hygiene (about once every four hours), and are often prescribed rinses for washing away food and bacteria, for pain, or for offering a protective coating over the mucosal lining. One such product, MuGard®, defended 57% of head and neck cancer patients, from experiencing ulcerative oral mucositis and showed mean weight-loss to be less than half of the control group[16]. MuGard® contains benzyl alcohol and benzalkonium chloride, which are known to be bacteriostatic agents, but also contains Carbomer Homopolymer A, providing a gel layer that protects from any outside aggressors. This suggests that environmental factors coupled with microbial communities influence nearly half of patients that develop OM. While these treatments are available, there are still a great deal of patients that continue to suffer from OM.

1.4 Fungi in next-generation genomics

Fungi have traditionally been classified using morphological and culture-based techniques. Although it is still the gold standard in the diagnosis of some infections, this approach is limited by the inability to culture many fungi, lengthy incubation times, and the microscopic similarities that many species share[17]. Providing a suitable treatment in the early stages of fungal infection is paramount in patient recovery and survival[17].

With the introduction of DNA sequence technologies, fungi can now be identified with higher sensitivity using molecular markers. In 2011, a multinational and multilaboratory initiative was undertaken to determine a DNA barcode region for accurate identification of fungal species[18]. The largest subunit of RNA Polymerase II (RPB1), and three regions of the nuclear rDNA cistron: 28S large subunit (LSU), 18S small subunit (SSU), and internal transcribed spacers

(ITS) were evaluated as promising candidates based largely on their representation in reference sequence databases. Of these, the ITS region demonstrated both a high rate of success in PCR amplification and sequencing as well as sufficient inter- and intra-species variability. This led to the proposal of ITS as the universal barcode for identification of fungi[18]. The general presence of the ITS regions is conserved to maintain spacing in the fungal rDNA operon, but their specific nucleotide sequences are independent of ribosomal function. The flanking, sequence-conserved 18S, 5.8S, and 28S subunits are appropriate targets for primer design, allowing for universal amplification of ITS in clinical and environmental samples. However, mixtures of microorganisms are commonly encountered in real-world samples, thus requiring time-consuming cloning steps to accomplish traditional methods of dideoxy sequencing.

The next-generation sequencing era has eliminated lengthy culturing and cloning steps and allowed for the high-throughout, massively parallel acquisition of ITS sequences. The ability to sequence millions of ITS DNA fragments in a single experiment has revolutionized the exploration of fungal communities and microbial interactions. Many organizations, such as the Human Microbiome Project, have been established to catalog the diversity in microbiomes across different regions of the human body in states of health and disease, but their main focus has been on bacteria. Indeed, many sophisticated pipelines have been developed to evaluate the complexity and biodiversity of bacterial communities. Although bacteria are essential for sustaining equilibrium of the niches they inhabit, they do not act alone, and are also impacted by the fungi that surround them. Pioneering next-gen studies on fungi have revealed that the complexity of human-associated fungal communities, or “mycobiomes,” may even increase with the complexity of bacterial microbiomes[19]. Nevertheless, tools developed for prokaryotes are not easily translated to analysis of fungal ITS because they rely on global alignments of small subunits. While the variability in ITS permits identification of species, it confounds alignment across fungal genera[20].

The entire ITS region, spanning approximately 450-700 bp for most fungi and including the 5.8S rRNA gene was recommended as the universal fungal barcode, but its length complicates next-generation sequencing experiments and analysis[21]. The addition of sequencing adaptors and conserved flanking regions for primer binding can extend these lengths by several hundred base pairs. Many next-gen platforms sequence approximately 200 bp without the need for bidirectional sequencing and assembly. With less than 1% of the estimated 5.1 million fungi represented in ITS sequence databases[21], relying on sequence assembly to accurately assign and quantitatively portray fungal communities is non-ideal. Roche's 454 pyrosequencing with Titanium XLR70 chemistry offers mode read lengths of 450 bp, while the most recent Titanium XL+ reagents extend mode lengths to 700 bp. At the time this study began, 454 was the only method capable of obtaining long reads (>400 bp) without the need for sequence assembly. ITS1 and ITS2, sequenced independently, are sufficient for fungal identification[21] and reduce sequences to 454 manageable lengths. There is no consensus between groups on which ITS region should be used for fungal identification, with amplification biases apparent in both instances and preferential representation of non-fungal sequences using only ITS2[21,22].

At the time this project was started, there was only one other publication attempting to perform metagenomic analysis for oral samples[23]. Ghannoum and colleagues sequenced ITS1 amplicons from 20 healthy subjects and reported the core components as those representing 1% relative abundance within each subject at ≥ 20% frequency. The most frequent species were *Candida, Cladosporium, Saccharomycetales, Aspergillus, Fusarium,* and *Cryptococcus,* four of which are well known human pathogens. The average number of reads was 1,702 per sample, obtained with 454 sequencing. In some samples, as many as 60% of the sequenced fungi were considered to be "non-culturable". Other biocompartments have recently been assessed for fungi including gut and sputum[24,19], also using the ITS1 region as a target for sequencing.

Though sequencing-based approaches of fungal identification have made characterizing communities a possibility, they are not without limitations. Further complications arise due to ambiguous taxonomic classification. Fungal species undergo morphological changes between their sexual and asexual states, originally leading scientists to assign multiple names to a single species without knowledge that their DNA sequences were identical. In April 2011, the Amsterdam Declaration on Fungal Nomenclature was enacted to transition mycologists toward a nomenclature system that allows only one name for one fungus[25]. This document declares that regardless of sexual state, a fungus should be considered by its first given name, with exceptions to those that are more widely recognized by younger synonyms. Not all situations are resolved with this document and it will not be until the 2017 Shenzhen Congress that other exceptions are addressed[26]. The work presented herein attempts to take these matters into consideration by employing a holistic approach to making fungal representation relevant to the biomedical community. It is the first to quantify the effects of collapsing synonymous genera into a single category.

1.5 Specific aims

This study provides a roadmap for analyzing fungal communities by first characterizing oral fungi in health, and then temporally after administration of oncologic treatments. The findings here will offer possibilities for development of preventative OM strategies and shed light on profiles susceptible to infection in all cases of immunosuppression. The project has three specific aims:

1. **To develop a method for obtaining fungal ITS1 sequences from human saliva and to empirically curate taxonomic results.** This aim has been accomplished by optimizing methods for breaking fungal cells, validating taxonomic legitimacy using e-

value thresholds, addressing dual and synonymous nomenclature of fungi, and evaluating non-genera classifications for re-assignment to the genus level.

2. **To refine the concept of the core mycobiome in the healthy mouth.** This aim has been accomplished by characterizing the core oral fungi in healthy saliva from 6 donors using a 454 deep sequencing approach. Genera at a frequency of at least half of the participants were considered to be core oral members. Validation of empirically determined parameters from Aim 1 were also employed here and 24 additional healthy subjects were mapped to the core and used to determine minimum sequencing depths necessary for capturing 95% of sample richness. Genus level mycoprofile frequencies were determined and variation in two week differences were explored.

3. **To characterize the fungal communities in patients undergoing treatment for cancer.** This aim has been accomplished by 454 sequencing and pipeline application from aim 1 of 21 healthy subjects assessed for baseline time points two weeks apart and 17 cancer chemotherapy patients at 4 time points of Day 0, Day 2, Day 9, and Day 14, with minimum sequencing requirements met for every collection. Comparisons were made to determine differences in mycoprofile abundances between healthy and cancer cohorts.

Chapter 2. Methods and standard operating procedures

2.1 Introduction

This study employs metagenomic techniques to explore the fungal composition of the human salivary microbiome. While many of these approaches are used in similar metagenomic studies, our pipeline has been optimized to yield amplifiable, fungal genomic DNA for sequencing Internal Transcribed Spacer 1, while reducing pre-sequencing artifacts. The subsequent computational pipeline has proven to reduce non-informative post-sequencing artifacts and to return legitimate identifications for fungal genera. Using controls to evaluate the breadth of identifiable genera showed the need to take precaution in accepting automated taxonomic assignments and in considering the complications of fungal nomenclature when reporting community proportions. The methods in this thesis are presented as detailed standard operating procedures as performed by Dupuy, et al.[27] and are applied to all clinical samples in a parallel fashion.

Precautions were taken to keep samples free of contamination for each protocol by using personal protective equipment (gloves, lab coat at extraction step) and sterile filter tips for pipetting. Optimal standard operating procedures were developed by comparing different extraction methods and PCR protocols (methods and results not shown) for a series of control subjects, reagent blanks, and *C. albicans* samples.

2.2 Ethics statement

Development of methods for this research project required healthy human volunteers and was performed according to a protocol (number X13-030) approved by the Institutional Review Board (IRB) of the University of Connecticut. The Institutional Review Board has determined that this study meets the criteria for Waiver of Informed Consent stated in 45 CFR 46.116(d).Conduct of research with respect to cohorts 1-3 was performed according to a

protocol (number UCHC11-037S-2) approved by the IRB of the University of Connecticut and UConn Health Center (UCHC). The UCHC was assigned as the IRB of record and agreed to inform the UConn Storrs IRB of all instances of unanticipated problems, should they occur.

2.3 Library preparation and sequencing workflow

2.3.1 Sample collection

Saliva Collection for Healthy Pilot Samples (Samples 50-52, 54-57)- location UConn Storrs:
Volunteers were instructed to refrain from eating and drinking non-water beverages for at least one hour before donating saliva samples. While medical records and health statuses were not formally measured, all six subjects were in their twenties and reported to be systemically healthy, non-smokers, and had no known oral conditions. Subjects expectorated about 3 mL of saliva into 50 mL Falcon tubes. Saliva was resuspended gently with a pipette and duplicate 1.5 mL aliquots were centrifuged at 3,300×g for 10 minutes. Supernatants were carefully removed to leave 200–300 μL and a pellet in each tube; in the case of large stringy pellets, as much supernatant as possible was removed without interfering with visible pellet material. Pellets from duplicate tubes were combined, re-pelleted, and supernatants removed to leave 200–300 μL that was extracted immediately or stored at −80°C.

Saliva Collection for Oral Mucositis Study (Arms 1-3)- location UCHC:
Subjects were restricted from oral hygiene on the date of collection, must have gone at least two days without antibiotics and antimicrobial rinses, must have refrained from smoking and gum chewing for at least four hours prior, and must have refrained from eating at least one hour prior to collection. Unstimulated saliva was collected for a total of five minutes in a 15 mL falcon tube placed in ice using a sterile plastic funnel. Tube content was aliquotted into as many 1.5 mL tubes as required and was centrifuged for 10 minutes at 3,300×g. Supernatants were removed without disturbing pellets or “jelly” like glycoprotein clumps. Pellets were frozen at −80°C. Pellets were thawed and resuspended in 200 µL TE buffer (20 mM Tris pH 7.0, 2mM EDTA). For 16S

bacterial amplification, 50 µL was removed. The remaining sample volume was re-frozen and delivered to UConn Storrs on dry ice for fungal amplification.

2.3.2 Fungal extraction using modified MP Biomedical FastDNA™ Spin Kit

Note: During this protocol it is important to open tubes only in the biosafety cabinet and to keep any exposed skin covered when handling. If gloves become contaminated at any point, replace 2nd glove layer or use 70% EtOH to clean. It is not recommended to exceed 8 sample extractions at a time. An extraction negative should also be included as a control totaling 9 parallel extraction processes.

1. With gloved hands, for each sample, use lysing matrix B and pour out B beads until total mass of tube + beads = 1.8 g. Add 1 gram of yttria stabilized zirconium beads to create B&Y custom lysing matrix. (Total weight = 2.8 g)
2. Place lysing matrix tubes on their sides in the small UV instrument for 15 minutes, ensuring that beads are distributed evenly across side of tubes.
3. When tubes are prepared, put on a fresh pair of gloves and lab coat.
4. Tape sleeves of lab coat to first pair of gloves so that skin remains covered.
5. Clean biosafety cabinet and pipettes with 10% bleach, followed by 70% ethanol.
6. Turn UV on in the hood for 15 minutes with pipettes inside.
7. Place a tube rack inside incubator and pre-heat to 55 °C.
8. After UV cycle completes, turn on biosafety cabinet, and put on a second pair of gloves over first pair of gloves.
9. Add 800 µL of CLS-Y solution to the lysing matrix.
10. Short spin sample tubes for 15 seconds.
11. Resuspend sample pellet and add entire sample volume to the lysing matrix tube.
12. Homogenize samples in FastPrep® 24 at speed 5 for 30 seconds.
    Note 1: Ensure that lysing matrix tubes are labeled on their sides, not tops, as FP®24 will rub off labels on caps of tubes.
    Note 2: Tighten top of FP24 until 3 clicks are heard.
13. Store samples on ice during 5 minute cool down of FP24.
14. Repeat steps 12-13 2x, omitting ice incubation step after 3rd homogenization.
15. Spin samples for 10 minutes at 14,000 RCF.
16. Carefully transfer as much supernatant as possible to a 2 mL tube, ensuring that no beads or pelleted cellular material is removed in the process.
17. Vortex Binding Matrix until it is resuspended.
18. Add 700 µL of well-mixed Binding Matrix to 2 mL tube with supernatant. Re-vortex stock Binding Matrix if it begins to settle before distributed to all samples.
19. Invert 2 mL tubes ~20 times by hand to mix.
20. Rotate in Labquake® for 5 minutes.
21. Pipette up and down 2-3 times before transferring 700 µL to spin filter in catch tube.
    Note: Make sure to collect any sample that remains in the lid of the 2 mL tube.
22. Spin at 14,000 RCF for 1 minute.
23. Discard supernatant from catch tube and replace spin filter.
    Note: Make sure to check supernatant for any Binding Matrix beads that may have leaked through filter. If spin filter is “leaky”:
    i. Resuspend liquid and beads in catch tube and return it to spin filter.
    ii. Resuspend entire contents of spin filter.
    iii. Transfer bead/liquid mixture into a new spin filter.

iv. Repeat steps 21-22.

24. Transfer remaining resuspended sample (650 µL max) to spin filter.
25. Spin at 14,000 RCF for 1 minute.
26. Discard supernatant from catch tube and replace spin filter.
27. Repeat steps 24-26 if any sample is left.
28. *Thoroughly* resuspend the pellet with 500 µL SEWS-M to wash (gently, so as to not shear DNA).
29. Spin for 1 minute at 14,000 RCF.
30. Discard supernatant in catch tube and replace spin filter.
31. Repeat steps 28-30.
32. Spin for 2 minutes at 14,000 RCF without any new addition of solutions.
33. Transfer spin filter to a clean catch tube.
34. *Thoroughly* resuspend the pellet with 100 µL DES to elute DNA.
35. Place tubes into pre-heated 55 °C incubator for 5 minutes (make sure lids are closed).
36. Spin for 1 minute at 14,000 RCF.
37. Discard spin filters and transfer eluate from catch tube to a clean 1.5 mL tube.
    Note: 1.5 mL tubes should be labelled with the sample identification number, the date extracted, and "B&Y Ex".
38. Make sure no crystals form in DNA after ~10 minutes.
    Note: If crystals have formed, spin tube at max speed for 10 minutes and transfer supernatant to a clean 1.5 mL tube.
39. Store eluted DNA in 4 °C (or -20 °C long-term).
40. Clean up space as described in steps 6-7 and turn off biosafety cabinet.

2.3.3 Determination of DNA concentration via the NanoDrop 2000

1. Clean biosafety cabinet and pipettes inside it with 10% bleach, followed by 70% ethanol.
2. Turn UV on in the hood for 15 minutes.
3. Check the sample tubes for crystal formation. If crystals have formed:
    a. Spin samples for 10 minutes at 14,000 RCF.
    b. Remove supernatant and place into a new 1.5 mL tube.
4. If no crystals have formed, vortex samples briefly (1-2 seconds).
5. Short spin the sample tubes to bring gDNA to bottom.
6. Remove 2 µL of extracted gDNA and place into a labelled PCR tube for each sample.
7. Add 2 µL of DES to a PCR tube for blanking the instrument.
8. Clean biosafety cabinet as in steps 1-2 and turn off.
9. Open NanoDrop 2000 program.
10. Create new file or open previous file.
11. Clean the NanoDrop plate and arm with a Kimwipe® pre-wetted with dH_2O and then dry.
12. Add 1 µL of DES to the NanoDrop plate.
13. Lower the arm of NanoDrop onto the plate and hit the blank button on the program.
    Note: Before measuring, ensure that no bubble has formed on top of the plate. If a bubble has formed, use the arm to try and pop the bubble on the plate.
14. Clean the NanoDrop plate and arm with a dry Kimwipe®.
15. Add 1 µL of the gDNA sample onto the NanoDrop plate.
16. Name run to match your sample.
17. Hit the measure button on the program.
18. Repeat steps 13-16 for all samples.
19. Clean the NanoDrop plate and arm with a Kimwipe®.
20. Print results.

2.3.4 PCR protocol for universal fungal ITS1 amplification

The fusion primers used in this protocol were designed for use with the 454 GS FLX Titanium, for sequencing with Lib-A DNA Capture Beads "B" in the forward direction (from 18S to 5.8S). From 5' to 3', fusion primer design entails a 454 adaptor sequence (A/B), a Multiplex Identifier (MID), and the template specific sequence. Standard 454 adaptor sequences were:
Adaptor A 5'-CGTATCGCCTCCCTCGCGCCATCAG-3' and
Adaptor B 5'-CTATGCGCCTTGCCAGCCCGCTCAG-3'.The template specific sequences were: ITS1 Forward Primer (ITS1F) 5'-CTTGGTCATTTAGAGGAAGTAA-3'[28] and ITS1 Reverse Primer- (ITS2) 5'-GCTGCGTTCTTCATCGATGC-3'[29]. MID sequences were selected from 454's Technical Bulletin No. 004-2009 and are listed in table 2.1. The entire primer formats specific to the "B" sequencing direction were: forward primer 5'-454 Lib-A Adaptor B, MID, ITS1F-3' and reverse primer 5'-454 Lib-A Adaptor A, MID, ITS2-3'.

ID	MID Sequence- 5'→3'
MID1	ACGAGTGCGT
MID2	ACGCTCGACA
MID3	AGACGCACTC
MID4	AGCACTGTAG
MID5	ATCAGACACG
MID6	ATATCGCGAG
MID7	CGTGTCTCTA
MID8	CTCGCGTGTC
MID10	TCTCTATGCG
MID11	TGATACGTCT

Table 2.1: MID sequences used for multiplexing of all samples

1. Determine the amount of gDNA needed for each PCR reaction:

   DNA volume = 125 ng / NanoDrop concentration (in ng/μL)

   Hyclone PCR water volume = 25 μL reaction volume – 9.625 μL total PCR components volume (includes 5 μL buffer, 1 μL each primer, 2.5 μL dNTPs, 0.125 μL polymerase)

2. Determine MID assignment for each sample, spreading samples evenly between MIDs 1-8, 10, 11, and with controls assigned to MID 1.
3. Cover PCR hood area in new bench coat.
4. Wipe down pipettes with 10% bleach, followed by 70% ethanol.
5. Expose hood, pipettes, a PCR rack, PCR water, and OneTaq® Buffer to UV for 15 minutes.
6. Turn on hood and label triplicate PCR tubes for each sample and each extraction negative to be amplified, and a single tube for both a positive control and a reagent blank.
7. Thaw MID-tagged Lib-A ITS1 primer aliquots, dNTPs, and positive *Candida albicans* control DNA. Briefly vortex and centrifuge these tubes alongside samples to be amplified.
8. To each PCR tube, add the calculated volume of PCR water. (15.375 μL for reagent blank and 14.875 μL for positive control).
9. Add 5 μL of 5X OneTaq® Buffer (final concentration: 1X).
10. Add 2.5 μL 2.0 mM (pre-mixed) each dNTP (final concentration: 0.2 mM).
11. Add 1.0 μL 5 μM (Lib-A ITS1-F + adaptor B + MID-tagged) primer (final concentration: 0.2 μM).
12. Add 1.0 μL 5 μM (Lib-A ITS2 + adaptor A + MID-tagged) primer (final concentration: 0.2 μM).
13. Add the calculated volume of template gDNA for each sample. (0.5 μL of 2 ng/μL *Candida albicans* gDNA).
14. Add 0.125 μL 5 U/μL New England BioLabs OneTaq® Hotstart Polymerase (final concentration: 0.025 U/μL).
15. Briefly vortex and spin each PCR tube.
16. Run *pre-heated* PCR with the following conditions:
    a. 94 °C for 30 seconds (1x)
    b. 94 °C for 30 seconds, 50 °C for 60 seconds, 68 °C for 60 seconds (35x)
    c. 68 °C for 5 minutes (1x)
    d. 10 °C hold
17. Clean PCR hood and pipettes as in step 4. Turn off hood and begin UV for 15 minutes.
18. Store tubes up to 3 days at 4 °C.

1% Agarose and TBE Gel Procedure

Part 1: Preparation of Sample

1. Add 1.25 μL of 5X Cresol Red into new labelled PCR tubes.
2. Briefly vortex and spin down all PCR products to be run on gel.
3. Add 5 μL of each PCR product into the appropriate tube containing Cresol Red.
4. Briefly vortex and spin down samples.

Part 2: Preparation of a 1% agarose, 1X Tris-borate-EDTA gel with ethidium bromide

Gels reagents (TBE, agarose, ethidium bromide) can be proportionally scaled up to create a larger gel for running more samples.

1. Tare a quarter sheet of weighing paper.
2. Weigh 1.0 g of agarose and add to a pre-cleaned 500 mL flask.
3. Add 100 mL of 1X TBE.
   Note: Rinse the neck of the flask with the TBE solution while adding to ensure that any agarose powder trapped on the neck is brought down.
4. Swirl gently, but thoroughly, in both directions.
5. Place the flask into a microwave and heat until solution until the solution begins to bubble/foam.
   Note: Room lights can be turned off to aid in seeing bubbles.
6. Using a hot pad, remove the flask and swirl gently one way for 5 seconds, and then swirl gently the other way for 5 seconds.
7. Repeat steps 6-8 2X until bubbles are big (not foamy).
8. Check the solution to make sure it is completely clear of any remaining agarose powder.
9. Let solution cool for 20 minutes.
10. Carefully add 1 µL of ethidium bromide to the flask and dispose of tip in ethidium bromide waste.
    Note: Ethidium bromide is a DNA intercalating agent and may be carcinogenic.
11. Gently, but thoroughly, swirl the solution to ensure the ethidium bromide is dispersed evenly.
12. Pour the agarose solution into a 100 mL tray that has been tightly secured to a level gel mold with the appropriate number of gels combs placed for sample loading and ladders.
13. Quickly rinse the 100 mL flask a few times with de-ionized water to clean and leave to dry upside down.
14. Check to see if bubbles are in the gel. If bubbles are present:
    a. Take a P10 pipette tip and turn it upside down so the large opening is facing the gel.
    b. Insert the pipette tip until the bubble in the gel is sucked up by the tip.
    c. Remove and discard the pipette tip in ethidium bromide waste.
15. Let the gel solidify for approximately 1 hour.

Part 3: Running samples on the gel

1. Carefully loosen the handle on the gel mold and remove the gel plate.
2. Fill a BioRad Sub-Cell® GT (that is attached to a Bio-Rad PowerPac™ Basic ¾ full with 1X TBE.
3. Place gel plate on the center of the rack.
4. Fill with 1X TBE until gel wells are completely covered. Remove lane combs carefully.
5. For each row of lanes that is present, pipette 1.5 µL of prepared 1 Kb+ ladder into the 1st lane (i.e. 2 row of lanes=1st lane in each row has Kb+ ladder).
6. Pipette the entire PCR product and Cresol Red mixture into a lane, making sure not to poke the bottom of the gel.
7. Repeat step 6 for all samples.
8. Place BioRad SubCell® GT lid onto the SubCell GT, matching colors.
9. Power on the BioRad PowerPac™ Basic.
10. Set voltage to 100V.
11. Set mA to 400.

12. Set time for 60 minutes.
13. Begin by hitting the start (running man) button.
14. Once on, ensure that bubbles begin to form at the black end of the SubCell® GT.
15. Let gel run for 60 minutes, periodically checking to ensure that the DNA/ladder from one lane does not run into the second row of lanes.
16. Turn off BioRad PowerPac™ Basic.
17. Remove the gel, holding your fingers on the bottom of the gel to ensure that the gel does not slide off the gel plate and pour off excess buffer. Dry the bottom of the plate with Kimwipes®.
18. Open the tray of the BioRad Gel Doc™ and place gel with tray.
19. Close the tray and turn on the Trans UV light.
20. Log onto computer using Bo's username and password= P@ssW0rd.
21. Open BioRad Quantity One program.
22. Hit GelDoc EQ button, live focus, and auto expose.
23. Adjust the exposure manually if necessary.
24. Freeze and save the gel image.
25. Click the file button, export to tiff, click export gain, and click save.
26. Click the print button, print under "Mitsubishi P95".
27. Exit program and turn off the Trans UV light.
28. Remove gel and place in designated waste bucket.
29. Rinse off gel plate with water.
30. Clean BioRad Gel Doc™ with 70% ethanol.
31. Evaluate for presence of bands (~300 bp for positive control) for each sample and for contamination in reagent blank and extraction negatives.
32. For replicate samples that performed equivalently, combine into a new 1.5 mL tube.
33. Store samples at 4 °C.

2.3.5 AMPure® XP size selection of fungal ITS1 amplicons

1. Vortex Agencourt® AMPure® XP reagent bottle to resuspend magnetic particles that may have settled.
2. Briefly vortex and spin samples. Measure the volume of each sample being tested via pipette.
3. Add 1.8 µL Agencourt® AMPure® XP reagent beads for each 1 µL PCR product for the first cleanup.
   Note: After the first cleanup and Pippin Prep procedure, the ratio is reduced to 1.5 : 1 Agencourt® AMPure® XP reagent beads to sample volume or 1.3 : 1 and 1.0 : 1 for further cleanups.
4. Mix thoroughly by pipetting 10x. The color of the mixture should be homogenous after mixing.
5. Incubate samples for 10 minutes at room temperature to bind sample DNA to the magnetic beads.
6. Place the sample tubes on a magnetic particle concentrator (MPC) for 5 minutes to separate the beads from the solution. Ensure that the solution has cleared before proceeding.
   Note: The hinges of the tube should be directly next to the magnetic part of the stand.
7. With the tubes still on the MPC, aspirate the solution from the tubes and discard the solution.

Note 1: Do not disturb the beads. If necessary, leave a few µl of supernatant rather than disturbing the beads.

Note 2: If the beads are disturbed, mix the supernatant thoroughly with the beads, and allow them to sit on the MPC for another five minutes before trying again.

8. Remove tubes from the MPC and add 200 µL of 70% ethanol.
9. Mix solution well via pipetting.
10. Incubate for 30 seconds at room temperature and place tubes back onto the MPC.
11. After the solution has cleared, aspirate the solution from the tube and discard.
12. Repeat steps 8-11 a second time.
13. Quick spin all tubes to bring down any remaining ethanol from the sample tubes.
14. Place the tubes back onto the MPC and allow the beads to travel up to the side of the tube.
15. Remove any ethanol from the tube and discard.

    Note: Be sure to remove all ethanol, as it is a known PCR inhibitor. A P10 may be necessary to remove the small volume of ethanol in this step.
16. Open the tops of the tubes and allow the beads to dry for 5 minutes at room temperature.

    Note: Watch the beads during this step. If any of the beads begin to crack, immediately proceed to the next step. Over drying beads significantly decreases elution efficiency; however, leftover ethanol may inhibit PCR.
17. Remove tubes from the MPC and add 30 µL Qiagen elution buffer (EB) to each tube.
18. Mix by pipetting 10x.
19. Replace tubes back onto the MPC and allow the samples to incubate for 5 minutes at room temperature.
20. SLOWLY, and CAREFULLY, transfer as much of the 30 µL eluate to a new 1.5 mL tube.

    Note 1: Do not suck up any of the beads. Use a P200 pipette to remove most of the solution. Then tilt the Magnetic Stand 96 at an angle and remove as much of the solution as possible using a P10 pipette.

    Note 2: If the beads are disturbed in the solution, resuspend the solution with the beads, and allow the sample to pellet to the MPC before trying again.
21. Store samples at 4 °C.

2.3.6 Sage Science Pippin Prep™ size selection

1. Briefly vortex and spin each sample, Loading Solution, and DNA Marker B. Measure the volume of each 1.8 AMPure® size selected sample via pipette.
2. Bring volume of each sample up to 30 µL with Qiagen Elution Buffer (EB).
3. Slowly pipette 10 µL of Loading Solution to each sample tube.
4. Briefly vortex and spin down all sample tubes.
5. Turn on Sage Science Pippin Prep™ and monitor, and allow the PX00476 software to automatically turn on.
6. Remove the 2% agarose Pippin Prep™ gel cassette from its foil bag and inspect the buffer levels in all of the lanes. If the levels are not even, add Electrophoresis buffer until level.
7. Inspect the gel cassette for delamination in the lanes.

    Note: If a lane has delamination, that lane cannot be loaded with ladder. Sample solutions may be loaded onto this lane if necessary.
8. Remove any bubbles from the elution chamber (dark blue center) by tilting the cassette elution side up until the bubble moves away from the area.
9. Open the Pippin Prep™ and place the gel cassette onto the optical nest.

10. Pull off the adhesive tape from the gel cassette.
11. Remove buffer from the elution module.
12. Add 40 µL of fresh Electrophoresis Buffer to each elution module.
    Note: Make sure you don't have any bubbles. Pop the bubbles with the pipette tip, or remove the buffer and try inserting the buffer again.
13. Seal elution wells with adhesive tape strips, ensuring that the tape does not cover the optical module.
14. Fill sample wells with Electrophoresis Buffer until solution is "convex" in each well.
15. Close the Pippin Prep™ and perform the Continuity Test by clicking "TEST" in the main tab.
16. If a lane fails the elution criteria, use that lane for ladder. If it passes, it can be used for sample. If a lane fails the separation criteria, neither sample nor ladder should be used in that well.
17. On the Pippin Prep™ software go to the Protocol editor and click new.
18. Set as "2% marker B No Overflow Detection".
19. Considering steps 7 and 16 choose a lane for reference (marker B). Set the rest of the samples lanes as "range".
20. Set base pair start to 160 bp.
21. Set base pair end to 1000 bp.
22. Fill in the sample ID template with the sample ID or ladder.
23. Click Save-as, and name the file.
24. Open the Pippin Prep™ and refill sample wells with Electrophoresis Buffer if they are no longer convex.
25. Remove 40 µL of Electrophoresis Buffer from sample wells.
26. Add 40 µL of marker B or sample into the appropriate sample wells, ensuring that the pipette tip does not touch the bottom of the well or the sides of the well.
27. Add Electrophoresis Buffer in a drop wise manner until the sample well is convex, making sure to never touch the sample in each lane to avoid contamination.
28. Close the Pippin Prep™ lid and hit start.
29. Once the test has finished (2.5 hours), remove the ~40-60 µL of sample volume from each elution module and place it into a new labelled 1.5 mL tube.
30. Shut down program and turn off monitor.
31. Pour liquid from gel cassette into ethidium bromide liquid waste and dispose of the cassette in ethidium bromide solid waste.
32. Store samples at 4 °C.

2.3.7 Bioanalyzer dimer removal verification and AMPure® XP size selections

Follow the standard operating procedure for AMPure® XP Size Selection of Fungal ITS1 Amplicons procedure following the note for 1.5 µL beads: 1 µL PCR product at step 3. Then, dilute the amplicon appropriately to the range of detection for the Bioanalyzer in elution buffer. Dilutions can be estimated using the agarose gel result to get within 15-200 pg/µL.

Prepare the Bioanalyzer DNA High Sensitivity Gel-Dye Mix

If solution is already prepared skip this section

1. Allow the High Sensitivity DNA dye concentrate (blue cap) and High Sensitivity DNA gel matrix (red cap) to equilibrate to room temperature for 30 minutes in the dark before use.
2. Add 15 µL of High Sensitivity DNA dye concentrate (blue cap) to High Sensitivity DNA gel matrix (red cap).
3. Vortex solution well and spin down briefly.

4. Centrifuge solution of at 2240 RCF +/- 20% for 10 minutes.
5. Wrap tube with aluminum foil to protect the solution from light.
   Note: Once the Gel-Dye Mix is prepared, it is good for 6 weeks.

Clean the Agilent 2100 Bioanalyzer:

Note: This process should only be performed after all runs are completed for the day. Washing electrodes before or between runs introduces the risk of improper drying and consequently, poor results.

1. Add 350 µL of deionized water to the "Washing-DNA Only" Bioanalyzer chip.
2. Dab the top of the plate with a Kimwipe® to remove any excess DI water.
3. Place the chip onto the Agilent 2100 Bioanalyzer and close the lid.
4. Leave the chip in the instrument for 60 seconds, and then remove the chip.
5. Pipette water out from the chip.
6. Repeat steps 1-5 three times.

Running a sample on the Agilent 2100 Bioanalyzer

1. Open 2100 ExpertBioA program.
2. Select DE13804730.
3. Select Assays.
4. Select ds DNA.
5. Select High Sensitivity assay.
6. Put a new High Sensitivity DNA chip on the chip priming station.
   Note: The next steps need to be done within 5 minutes.
7. Pipette 9 µL of gel-dye mix into the well marked with circled "G".
   Note: The solution is very viscous. To pipette the solution, push the pipette aspirator button down past the second stop. Remove the solution. Then pipette the solution into the well, pushing only to the first stop.
8. Align plunger at 1 mL position. Push syringe down until it is held by the metal clip.
9. Let sit for <u>exactly</u> 60 seconds.
10. Release the plunger and wait for 5 seconds. Pull the plunger back into the 1 mL position.
11. Open the chip priming station and add 9 µL of gel-dye mix to the remaining 3 wells in the far right hand column, excluding the previously filled well.
12. Remove 70 µL of marker (green cap) and place into a new 1.5 mL microcentrifuge tube.
13. Pipette 5 µL of marker into all sample wells (even if unused) and the ladder well.
14. Add 1 µL of High Sensitivity DNA ladder (yellow cap) into the well marked with a pictogram of a ladder.
15. Pipette 1 µL of sample into each of the remaining 11 sample wells.
    Note: If a well is not being used for sample, add 1 µL of marker.
16. In the 2100 ExpertBioA program, label the wells with their corresponding sample ID.
17. Place the Bioanalyzer chip onto the IKA MS3 vortex and push in the chip to start vortexing for 1 minute at 2400 rpm.
18. After the vortex is finished, remove the Bioanalyzer chip and place it into the Agilent 2100 Bioanalyzer.
19. Hit start on Expert BioA program.
    a. Note: Be cautious around the Agilent 2100 Bioanalyzer when it is running. Vibrations around the machine will adversely affect results.
20. When the chip is finished running, remove and discard the chip.
21. Perform Bioanalyzer cleaning procedure as outlined.

Bioanalyzer 2100 Expert Analysis

1. Navigate to appropriate run to be analyzed.
2. Click on the “Global” tab on the right hand side of the screen, then use the dropdown menu to select “Advanced”. Use this option.
3. Under ladder setpoints, check the box to perform Baseline Correction.
4. Under sample setpoints, check the box to perform Baseline Correction.
5. Click “Apply to All” at the bottom of the Global settings.
6. Double click on single sample to view it individually and select the “Region Table” tab.
7. Ensure that the baseline is flat and the upper marker is a single complete peak, (which returns to baseline and is higher than the lower marker). Ensure that the amplicon is not perturbed by artifacts. If electropherogram fails to meet these requirements retry sample in another Bioanalyzer run.
8. Click on the “range” tab at the bottom of the screen for each individual sample and record the concentration in pg/µL for the 200 bp-1 kb automated smear analysis. (If necessary, adjust the smear to include all DNA).

    Note: If primer artifact peaks appear at 160 bp or less, additional AMPure XP washes are necessary (1.3 : 1 or 1.0 : 1 volume ratios). Samples that require additional cleanups will need to be repeated on another Bioanalyzer run.

2.3.8 Dilution of samples to 1.00×10^8 molecules/µL and pooling

1. Use concentration derived from the Bioanalyzer chip to determine the volume of DNA and volume of Qiagen elution buffer (EB) necessary for diluting the sample down to 1.00×10^8 molecules/µL.

    Bioanalyzer concentration of sample (molecules/µL) =

    $$\frac{\text{Bioanalyzer concentration (pg/µL)} * 6.022 \times 10^{23} \text{ (Avogadro's number molecules/mole)}}{6.56 \times 10^{14} \text{ (Avg ng of bp/mole)} * 326 \text{ bp (estimated ITS1 amplicon length based on } \textit{C. albicans})}$$

    Bioanalyzer concentration of sample (molecules/µL) x sample volume (µL) =

    1.00×10^8 molecules/µL x 20* µL

    *Choose an appropriate volume in order to make the calculated volume reasonable for pipetting.

2. Vortex and spin samples briefly.
3. To a new 1.5 mL tube, add the exact volumes of DNA and EB to each tube.
4. Vortex and spin down all of the tubes.
5. Determine which samples to combine into one pool, while taking care not to use two of the same MIDs in one pool.
6. Add 5µL of each sample to its corresponding pool tube.
7. Vortex and quick spin the pool tubes.
8. Run the pooled samples on a High Sensitivity Bioanalyzer chip to ensure that the final concentration of the sample is 1.00×10^8 molecules/µL.
9. Store samples at 4 °C.

2.3.9 emPCR and sequencing

1. Dilute 1x10^8 molecule/µL pools to 4x10^6 molecules/µL in a new tube with EB.
2. Follow 454 emPCR Amplification Method Manual – Lib-A MV or SV using Capture beads "B" and medium or small volume emulsion oil for 4 or 8 region gaskets, respectively. Protocol found at: http://454.com/downloads/my454/documentation/gs-flx/method-manuals/GSFLX_GSFLXPlus_emPCRAmplificationMethodManual_Lib-A_SV_Jun2013.pdf
3. Follow 454 Sequencing Method Manual Protocol found at: http://454.com/downloads/my454/documentation/gs-flx/method-manuals/GS-FLX-Titanium-Sequencing-Method-Manual-%28Nov2010%29.pdf

2.4 Fungal ITS1 data analysis

Required downloads: QIIME, bioperl, DeconSeq, Virtual Box

2.4.1 Transfer raw data and create sff files

1. Move R Signal processing folder into the dated sequencing folder that was transferred to admin rig. Delete duplicate files from folder.
2. Create a folder for each region under signalprocessing/sff
3. Use terminal to parse raw region sff files by MID to generate sff files for each sample. sfffile -s <regionname>01.sff
4. Drag newly generated sff files into appropriate pool folder.
5. Repeat for other regions.
6. Generate a text file containing raw sequence outputs for each MID.
7. Copy entire dated sequencing run folder to DROBO or other external storage device.

2.4.2 Transfer data for analysis

1. On data analysis computer, open WinSCP and log on to CAGT computer using ssh.
2. Copy .fna (fasta) files from external storage to CAGT.
3. Open virtual box and start up QIIME machine.
4. Click on the terminal and ssh into CAGT.
5. cd into folder you just created with fna files inside.

2.4.3 Run dimer removal program:

1. perl ../DimerRemovalUseThisOne/08302013dimer_remove.pl 1.TCA.454reads.fna ../DimerRemovalUseThisOne/08302013dimer.fn [enter]
2. Repeat for remaining regions and grep to get sequence counts in output files. Example: grep ">" 1.TCA.454reads.fna_good.fn | wc -l [enter]

2.4.4 Run DeconSeq program

1. type “screen” then:
deconseq -f /home/amanda/[name of folder containing fasta file]/1.TCA.454Reads.fna_good.fn -c 90 -i 94 -dbs arch,bact,bacthmp,dr,hs_alt_celera,hs_alt_CRA,hs_alt_HuRef,hs_alt_KoRef,hsref,hs_unique,mm,senterica,vir
2. CTRL + A + D to detach
3. type screen again and run command on each "good" fasta file.
4. type more <run#>_a[tab] for each of the four run #s and note which region belongs to which run#.
5. type screen -list to note the screen numbers for each run: To resume a screen type screen -r <screen#>.pts-<#>.cagtflx. When process is finished type: exit (to erase screen).

2.4.5 Run QIIME split_libraries.py command

1. Create a tab separated mapping file for each region with this format:

#SampleID	BarcodeSequence	LinkerPrimerSequence	ReversePrimer	Treatment	SequencingDate	Description
#mapping file for the QIIME analysis package.						
2-019-1	ACGAGTGCGT	CTTGGTCATTTAGAGGAAGTAA	GCTGCGTTCTTCATCGATGC	Healthy	20140203	MID1
2-001-1	ACGCTCGACA	CTTGGTCATTTAGAGGAAGTAA	GCTGCGTTCTTCATCGATGC	Healthy	20140203	MID2
2-001-2	AGACGCACTC	CTTGGTCATTTAGAGGAAGTAA	GCTGCGTTCTTCATCGATGC	Healthy	20140203	MID3
2-001-3	AGCACTGTAG	CTTGGTCATTTAGAGGAAGTAA	GCTGCGTTCTTCATCGATGC	Healthy	20140203	MID4
2-001-4	ATCAGACACG	CTTGGTCATTTAGAGGAAGTAA	GCTGCGTTCTTCATCGATGC	Healthy	20140203	MID5
2-008-1	ATATCGCGAG	CTTGGTCATTTAGAGGAAGTAA	GCTGCGTTCTTCATCGATGC	Healthy	20140203	MID6
2-005-2	CGTGTCTCTA	CTTGGTCATTTAGAGGAAGTAA	GCTGCGTTCTTCATCGATGC	Healthy	20140203	MID7
2-014-3	CTCGCGTGTC	CTTGGTCATTTAGAGGAAGTAA	GCTGCGTTCTTCATCGATGC	Healthy	20140203	MID8
2-020-2	TGATACGTCT	CTTGGTCATTTAGAGGAAGTAA	GCTGCGTTCTTCATCGATGC	Healthy	20140203	MID11

2. Save file as MappingFile<PoolName#>.txt
3. Use WinSCP to copy clean and contam files to Analysis computer. Rename them so they reflect Poolname.
3. Logout of ssh and Split libraries in terminal of analysis computer QIIME:
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/OMFPools1-4$ split_libraries.py -m MappingOMFPool1.txt -f [filtered pool 1 file name]_clean.fa -l 100 -L 10000 -t -a 1 -H 10 -M 2 -e 2 -b 10 -z truncate_only --reverse_primer_mismatches 6
4. Make new folder for each Pool and transfer seqs.fna, log, and histogram files into appropriate folder so that next command doesn't overwrite these files.
5. Repeat for other pools and move files to folders as they complete.
6. Rename sequence files to contain pool name.

2.4.6 Remove short sequences with Galaxy

1. Go to windows explorer and navigate to: ftp://usegalaxy.org
2. Login with galaxy id and credentials.
3. Copy QIIME pass fna files to galaxy ftp.
4. In internet browser go to https://usegalaxy.org and login.
5. Click Get Data> Upload File and check boxes in FTP section once files are finished uploading. Click Execute.
6. Once files turn green click FASTA manipulation > Filter sequences by length. Select file, Minimal length = 100, maximum length = 0 [execute]
7. Click pencil to rename output files to <Poolname>100plus
8. Click Name of file > Disk image to save.

9. Remove brackets ([ ]) from saved files.

2.4.7 Submit to FMP for taxonomic identification

1. Open FMPShare Folder on analysis computer.
2. Copy and paste 100plus files into FMPShare.
3. Navigate to FMP Portal: https://biotech.inbre.alaska.edu/portal/ and log in with credentials.
4. Once files finish uploading click Applications > Microbial Pipeline > Search Fungal ITS.
5. Type email, name job, and Select from Secure File Share. Browse for file. Choose curated_its in dropdown and keep Blast in search dropdown. Click Run.
6. Save blastall.summary file.

2.4.8 Removal of sequences by e-value filter

1. Use excel to open blastall.summary file.
2. Find and replace “e-“ with “1E-“ using case sensitivity.
3. Sort e-value column from high to low.
4. Paste data for “good” e-values ≤ $1x10^{-42}$ in a new file.
5. Search for each sample name and create a new tab delimited text file for the good e-values, leaving the first two rows without data.

2.4.9 Run taxa counting program

1. Open QIIME virtual box and transfer .txt files for each sample using transferring procedure in 2.4.2.
2. Run perl gi_sgp.pl [samplefilename].txt for each file to generate a list of genera and their counts.

2.4.10 Standard operating procedure for combining genera

Finally, genera were collapsed by hand curation using biblioinformatic guidelines suggested by Hawksworth[30]. In addition to the recommended citations in Google, Google Scholar, and Bibliography of Systemic Mycology (BSM), NIH PubMed citations were added in deference to the biomedical orientation of this research. Google searches were qualified with “fungus” when the genus names mapped to objects other than fungi (as an example, valsa refers to a waltz as well as a fungal genus). A holistic approach was used for conjoining genera. Synonyms were identified using Uniprot, BSM, and original literature. Because sexual and asexual pairs have largely been identified by binary names, we first compared species alternatives by citation numbers, weighing PubMed searches more heavily based on the health-related aspects of this research. When alternative species names had similar citation numbers, citation searches for

genera were considered. Genera were not entirely conjoined unless all of the species identified in our study had synonyms in the alternative genus. To retain access to information inherent in the dual nomenclature system[31], we continued to list other genera that were collapsed into the first listed priority genus (as an example: Aspergillus/Emericella/Eurotium). When alternative genera had citation numbers that were too close for comfortably naming one as a priority designation, the original name identified by the NCBI BLAST searches was retained (as examples, Sporidiobolus and Sporobolomyces).

2.4.11 Analysis and diversity measurements

Collapsed genera were placed in excel tables to show relative abundance of each for every sample. Determination of capturing 95% of sample richness was measured using rarefaction curves in iterations of 10 with an implementation of Rscript: alpha_rare.R in VAMPS[32]. Phylogenetic trees, Morisita-Horn distance calculations, heat maps, genus level mycoprofile pie charts and bar graphs were generated using VAMPS community visualization tools with a 0.1% relative abundance cut off.

Chapter 3. Considerations of protocol refinement and limitations for fungal metagenomics

3.1 Introduction

Serving as the lead example for this study was the only other publication on the oral mycobiome using a non-culture-based metagenomic approach: “Characterization of the oral fungal microbiome (mycobiome) in healthy individuals” by Ghannoum et al[23]. However, when replicating Ghannoum and colleagues’ protocols, several complicating factors were encountered that began to confound our results. From DNA lysis, to sequence curation, to assigning taxonomy, care had to be taken to ensure productive and reliable results. This chapter details the measures taken to refine methods based on Ghannoum’s protocol, to challenge the current standard for sequence assignments, to validate our approach, and to provide a roadmap for tackling fungal mycobiome samples.

3.2 Evaluation of fungal lysis using common DNA isolation kits

In order to represent all fungi present in a sample from their DNA, sufficient measures must be taken to adequately lyse cells. Fungi are encased in a resilient cell wall of chitin and in some cases an additional polysaccharide capsule, which may even enlarge and enhance virulence when they become threatened by a host, complicating extraction techniques[33,34]. Methods described for fungal DNA isolation have historically included CTAB (cetyltrimethylammonium bromide), which solubilizes cell walls, while denaturing proteins[35] or freezing with liquid nitrogen before grinding with a mortar and pestle[36]. Modifications are generally made to extraction methods depending on the resilience of fungal spores or unique cell wall of the organisms being studied. While the CTAB method provides the benefits of sufficiently removing carbohydrates during DNA isolation, it is considered insufficient for cracking most spore types as well as some mycelial fungi[37].

Often times a bead beating step is used to break these barriers that prevent access to gDNA. The characteristics of the bead (smooth vs. rough and varying density based on chemical composition), and its relative size to that of the sample and its cells, will often determine its effectiveness in lysing a sample. For instance MP Biomedicals currently offers 16 different bead compositions of “lysing matrices”, which are specifically designed to break cell types and extract specific cellular components. Ghannoum and colleagues used MP Biomedicals’s FastDNA Spin Kit, which employs a bead beading technique in its protocol. However, no clear indication was made as to which lysing matrix was used. It was presumed that Ghannoum et al. used Lysing Matrix A, as it is standardly supplied with the kit by default. A technical representative from the MP Biomedicals company recommended a more robust bead that might release DNA from fungal cells better than their own matrices commercially available in 2011. The bead suggested was made of yttria stabilized zirconia which is an extremely high density material. At the time, MP Biomedicals did not provide this bead, so it was specially ordered from an outside milling company for testing on equimolar aliquots of cultured *C albicans* cells. A sample pack of several other types of Lysing Matrices were also ordered for comparison purposes: Matrix A- garnet flakes and one ¼ inch ceramic sphere (good for all sample types except soil), Matrix B- 0.1 mm silica spheres (good for isolating bacterial DNA), Matrix E- 1.4 mm ceramic spheres, 0.1 mm silica spheres, and one 4 mm glass bead (good for mixed tissue samples), Matrix Y- 0.5 mm yttria stabilized zirconia spheres (good for lysing fungal tissue and spores), and Matrix B&Y- 0.1 mm silica spheres and 0.5 mm yttria stabilized zirconia spheres (combination of Matrix B and Matrix Y). After DNA extraction with respective matrices, ITS1 PCR was performed as in section 2.3.4, but with BioRad iTaq™ DNA Polymerase. Surprisingly the matrix presumably used by Ghannoum and colleagues (A) produced no detectable amplification, while matrices B, E, Y, and B&Y showed clear amplification in an agarose gel. In all instances of sufficient gDNA extracted for PCR amplification, there was absence of a dimer band < 100 bp, but when the same gDNA was diluted there was presence of the dimer band. This begs the question of why

lysing matrix A ITS1 product has no dimer band. It is possible that the band for this matrix is present, but is below the limit of detection and remains unseen in an agarose gel. By combining matrices B and Y, the brightest amplicon signal was obtained, implying that more gDNA was intact and available for amplification than for other matrices. Further tests comparing matrices B, E, and B&Y with amplification of ITS1 from saliva also showed evidence that the B&Y matrix would give consistent results for relevant samples (data not shown), and that this combination was the best option for the purpose of lysing fungal cells from saliva.

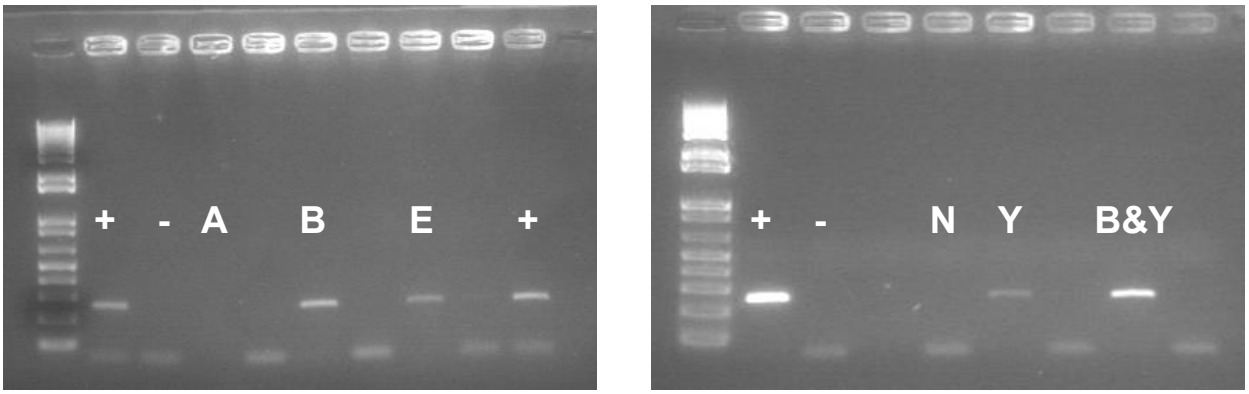


Figure 3.1: Gel images of lysing matrices compared for *C albicans* ITS1 amplification, with *C. albicans* dilutions to the right of each matrix depicted. + represents amplified *C. albicans* positive control DNA, - represents reagent blanks, A, B, E, Y, and B&Y refer to matrices

The FastDNA spin kit was not without its weaknesses. At the elution step of initial extractions there was a precipitate produced that contained small clear crystals. Amplifications performed with such crystals did not result in visual DNA bands on an agarose gel. In order to determine the cause of the crystal formation, the DNA isolation was performed without DNA and was started at different places in the protocol, so that one chemical was added at a time in a "backwards fashion". For example, elution from the column was performed first, and no noted crystals were formed. This showed that the elution buffer was not adversely reacting with the tubes or column to create precipitate. In new tubes, the protocol was started at the addition of

DNA binding matrix, and also at the addition of CLS-Y lysing solution. The protocol was followed through to completion for each test. Only one of the tests produced crystals; the component responsible for the precipitate was the CLS-Y lysing solution. There appeared to be a reaction between CLS-Y and the elution buffer after several minutes of incubation. To remedy this, an additional ethanol wash was added to the manufacturer's protocol, which resulted in more adequate removal of the CLS-Y before elution. Both ethanol washes in the modified protocol were also advised to be performed thoroughly to ensure such removal. If washed well, the eluate contained no crystals.

3.3 Comparison of additional extraction methods

While pure cultures of *C. albicans* perform exceedingly well in DNA isolation kits designed for yeast[38], it was important to remember that our samples contained many types of fungi, so it was imperative to test kits with whole saliva to evaluate productivity of sequencing and to test for inherent contaminants. Four extraction methods were compared using equal volume aliquots of the same saliva sample: 1) the FastDNA Spin Kit Lysing Matrix A, 2) the FastDNA Spin Kit with Lysing Matrix B&Y, 3) the UltraClean Soil (UCS) DNA Isolation Kit, and 4) the MasterPure Yeast (MPY) DNA Purification Kit, following the manufacturer's protocol or with modifications as described in section 3.2. The resulting extracts were amplified with ITS1 primers and sequenced using protocols detailed in chapter 2. After removing sequences composed of primers and filtering for sequences characterized by an acceptable amount of mutations in primers, homopolymers, "N" calls, and length, the FastDNA A matrix and UCS extraction methods produced the fewest number of sequences (Table 3.1). Though Ghannoum and colleagues do not report the 454 region size used, the relatively low number of total sequences obtained in their study is in concordance with low sequence counts obtained for FastDNA lysing Matrix A presented here. The MPY method showed relatively high numbers of reads in the reagent blank, but should be tested further as it performs well in other studies[38]. Our modified FastDNA

B&Y method yielded the greatest percentage of productive sequences and showed that contamination would not confound results, confirming results of comparisons between lysing matrix A and B&Y from section 3.2 and providing support for our modified method in all downstream experiments.

Table 3.1: Comparison of four extraction methods on whole saliva

(Region[a]) Method[b]	Raw sequence count	Sequences remaining after QIIME restrictions (% previous column)[c]	Sequences assigned to negative (% previous column)
(2) Fast DNA/matrix A	8,493	119 (1.40)	1(0.84)
(3) Ultra Clean Soil	13,434	38 (0.28)	2 (5.26)
(4) MasterPure Yeast	32,604	16363 (50.19)	524 (3.20)
(5) Fast DNA/matrix B:YSZ	41,448	30326 (73.17)	19 (0.06)

[a] Region on an 8 field picotiter plate
[b] FastDNA Spin Kit/Lysing Matrix A (MP Biomedicals); UltraClean Soil DNA Isolation Kit (MoBio); MasterPure Yeast DNA Purification Kit (Epicentre); FastDNA Spin Kit/Lysing Matrix B (MP Biomedicals) + 0.5 mm yttria stabilized zirconia beads (GlenMills);
[c] QIIME restrictions: Minimum length = 100 (after trimming forward primer and MID); maximum "N" = 1, maximum homocopolymer = 10; maximum forward primer mismatch = 2; maximum barcode mismatch = 2

3.4 Recognition and elimination of primer artifacts

The first attempt at sequencing the healthy mycobiome included 5 barcoded subjects over 1/8 454 Picotiter plate. This region produced a total of 74,693 raw reads. Many of the raw sequences were run through NCBI BLAST to confirm their identity as fungal DNA. Nearly all sequences were assigned to an "uncultured fungal clone", suggesting the possibility that several new taxa may have been sequenced. After a closer inspection, these sequences consisted solely of concatenated primers averaging ~65 bp, but could reach more than 200 bp in length. The presence of primer artifacts accounted for ~95% of sequences and was the primary cause for reduction of total counts to a mere 4,509 classifiable sequences. As the majority of these artifacts were relatively small compared to the amplicon size of interest (~200-600 bp after addition of fusion primers[39]), measures were implemented as part of the library preparation procedure to remove them prior to sequencing. Because not all primer artifacts were eliminated

through size selection, a custom perl script was created to remove them in the first step of the pipeline.

One additional library preparation step for the removal of small fragments was the Sage Science Pippin Prep™. A 2% Agarose Gel Cassette was chosen in order to collect fungal ITS1 fragments 160 bp or greater. This length was determined by adding an extra 60 bp for barcode and adapter sequences (trimmed before the 100 bp filtering step) to the minimum number of nucleotides required for ITS1 sequence length. The Pippin Prep™ employs pre-poured agarose gels and electrophoresis to collect custom sized fragments by changing electric current to direct DNA into an elution module. While this step greatly reduced the amount of primer artifacts, an additional library preparation step was sometimes needed if small fragments persisted. AMPure® XP beads were employed for their size selection properties, which work by preferential binding of DNA to solid-phase reversible immobilization paramagnetic beads in the presence of polyethylene glycol and salt. For the majority of samples these two additions were adequate for removing primer dimer. However, not all clinical samples are created equal, and some produced fragment profiles that were comprised almost completely of primer dimer. No matter how many additional AMPure® XP purifications were performed, they were inadequate to remove dimer yet retain sufficient amounts of ITS1 DNA for sequencing. Last, in order to

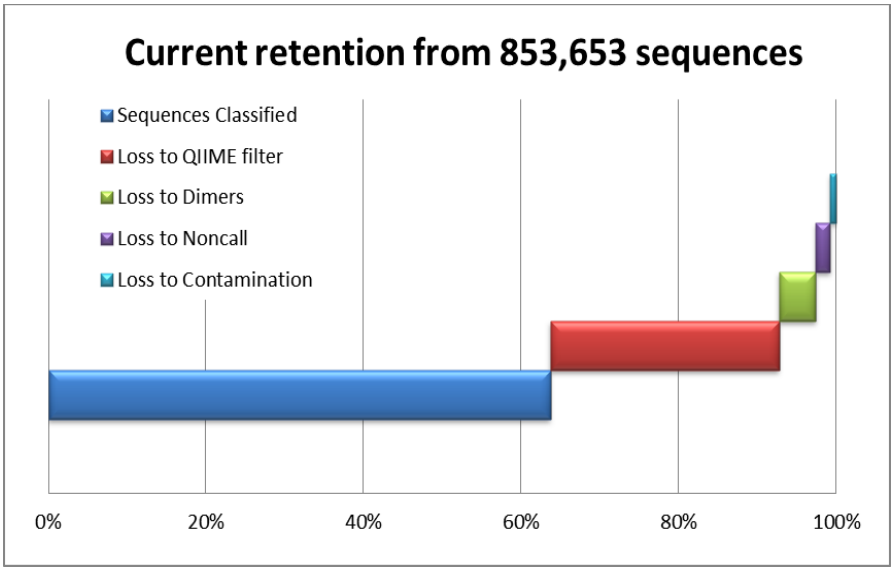


Figure 3.2: Bar graph of improved sequence retention applied to 6 healthy subjects with the improved library preparation modifications and bioinformatics pipeline

remove any remaining primer concatenations, especially those that were within the length of amplicon interest, a custom bioperl script was implemented. The bioperl script creates a "database" out of sequences collected from saliva samples and uses BLAST to query for the commonly found dimer, filtering sequences that it matches. Application of these modifications to a deep sequencing effort of 6 healthy subjects illustrated the improvements in loss due to dimer (Figure 3.2).

3.5 Using upstream analyticals to predict successful sequencing

One of the challenges encountered with samples in restricted availability and/or limited quantities is that they may be fully consumed in processing if not handled very carefully. Several assessments emerged as indicators of such samples and provided predictive values of success early in the library preparation workflow. Another quality control measure for a sample,

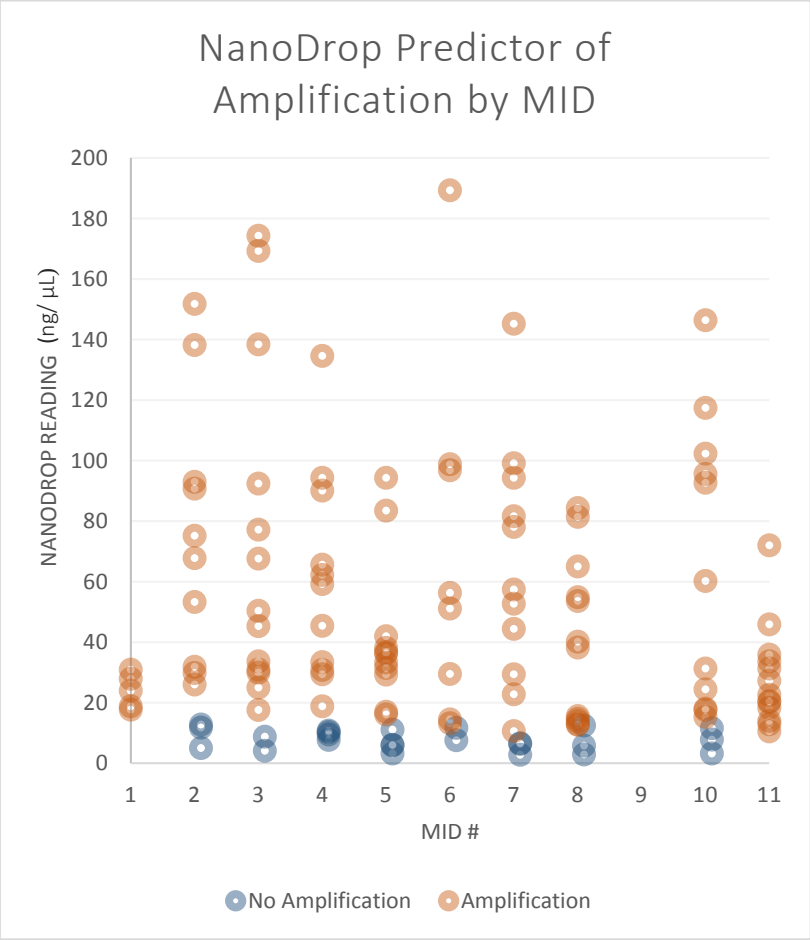


Figure 3.3: Scatterplot of NanoDrop reading as an indicator of potential amplification, separated by MID

NanoDrop measurement, comes after the gDNA extraction step and indicates potential for amplification. Low NanoDrop values (≤13 ng/μL) consistently represented samples that proved difficult for detection of quantifiable amplicons before pooling samples for sequencing. Low values for saliva samples could be caused by user errors during the extraction protocol (inadequate ethanol washing, failure to remove fully remove layers containing PCR inhibitors), or limited cell count in the sample source. Figure 3.3 shows the distribution of samples with a concentration of less than 200 ng/μL gDNA for all patient samples (further characterized in chapters 5 and 6). The average NanoDrop reading for samples that exhibited clear amplification of ITS1 fragments was 55.1 ng/μL (range 10.5 to 189.3, n=106). The average NanoDrop reading for samples that exhibited no signs of amplification was 7.75 ng/μL (range 2.8 to 12.8, n= 24). To eliminate MID sequence as a variable affecting amplification success, NanoDrop values were also plotted by MID number. Only MIDs 1 and 11 showed a consistent predisposition toward successful amplification, while all other MIDs evenly shared successful and unsuccessful amplification of ITS1.

A second assessment point is the relative amount of amplification product. In cases where PCR amplification is low, samples should be handled with extra care during artifact removal steps. For instance, saliva volumes at < 0.75 mL did not produce amplicon signals in gel electrophoresis, so quantities under this amount should be considered carefully before subjecting to library preparation steps that further reduce amplicon amount. We continued to notice that preferential amplification of dimers occurred when amplification of ITS1 products was low and hypothesize that if fungal gDNA content is low, increased care is needed to remove the high amounts of dimers that accompany limited ITS1 amplicons. Skipping the agarose gel and Pippin Prep™ steps increases the likelihood that more product, however little, may be retained. Proceeding directly to the Bioanalyzer can give an indication of whether a sample is expected to fail and is therefore not as good an investment of time and reagents as other samples might be.

For example, Bioanalyzer readings that continued to show predominant primer dimer peaks, after up to as many as five AMPure® XP purifications, consistently resulted in ITS1 fragment removal below the limit of detection in the Bioanalyzer, while small dimer peaks were still present. These samples were thus unable to be sequenced due to loss of quantitative data that allows for equimolar pooling of multiplexed samples. If dimers must be removed, an AMPure® XP purification could be added as long as the volume of elution material is decreased to help concentrate the amplicon in the eluate.

3.6 Database selection

One of the most important factors in presenting reliable interpretation of sequencing results is choosing a database that includes the best representation of relevant taxa. Ghannoum and colleagues used the Assembling Fungal Tree of Life (AFTOL) and NCBI databases for analysis. In order to assess other recommended databases[40] for reliability and ease of use, four databases were queried with a *C. albicans* ITS1 sequence (completed in February 2012): 1) BoldSystems ITS, 2) mycologylab.org, 3) UNITE, and 4) the Fungal Metagenomics Project (FMP). As a top hit, BoldSystem ITS returned *Mystrium oberthueri*, an ant, with an e-value of $6x10^{-23}$, while UNITE returned *Amanita virosa,* a poisonous basidiomycete fungus at an e-value of $3x10^{-30}$. The optional addition of INSD (including GenBank, DDJB, and EMBL) in UNITE returned uncultured Ascomycota at $1x10^{-138}$, which is the correct phylum for *C. albicans*, but is not specific enough as a top hit considering its biomedical relevance to this study. The FMP and mycologylab.org both returned *C. albicans* as a top hit with respective e-values of $1x10^{-130}$ and $2x10^{-88}$. Such differences in accuracies of these two databases compared to UNITE and BoldSystems ITS is likely due to the composition and focus of the databases. BoldSystems ITS offers a disclaimer: “There are very few ITS records on BOLD so most queries will likely not return a successful match”, while UNITE comprises plant pathogens and symbionts. Mycologylab.org comprises only curated human/animal pathogenic fungal species, while FMP

offers a curated search of Genbank, which eliminates sequences with designations to "uncultured" and "environmental" descriptors. The removal of primer regions from the ITS1 sequence returned the same results for both mycologylab.org and the FMP. For BoldSystems ITS, search results were improved with primer removal, so that at least the correct identification to Ascomycota was made, although at the very weak e-value of $8x10^{-8}$. Primer removal in UNITE resulted in a top hit to *Thelephora alnii* at e-value = $2x10^{-21}$ and with addition of INSD, *C. albicans* was returned at $1x10^{-117}$. These data made it clear that FMP would provide a consistent database that contained relevant results with strong e-values with and without primer removal.

3.7 Removing non-informative sequences by length

To be considerate of other FMP users and respect the Project's computational resources, we wanted to eliminate non-informative sequences as much as possible before submitting sequences to be classified. All sequences from the 6 healthy samples that were deep-sequenced for core mycobiome analysis, were partitioned into length classes (post-adaptor, primer and MID removal) to determine the relevance of each category by its assigned FMP e-values (Table 3.2).

For the smallest size class (0-99), there were no returns with e-values better than $1x10^{-15}$, a poor e-value threshold that represents alignment of reference sequences to short flanking regions. We determined that sequences less than 100 bp should be removed before submitting to FMP because filtering by length occurred *before* removal of the reverse primer in QIIME. A bioinformatic step to remove sequences in the 0-99 bp category was implemented using Galaxy to increase efficiency in assigning taxa in FMP. The sequences in the 200-299 bp and 300-399 bp categories were most fruitful for providing informative sequences. Sequences with lengths outside of these two categories are either generally a result of non-specific amplification or may represent new, medically relevant fungi. Distinguishing between the two hypotheses can be

revisited in the future and will be made easier as more reference fungal sequences are deposited into databases.

Range	Number	%of total	Evals <1E-15	%Evals <1E-15
0-99	7005	1.253281	0	0
100-199	5127	0.917283	2479	48.35186
200-299	382161	68.37331	365925	95.75153
300-399	163767	29.29993	162735	99.36984
400-499	787	0.140804	106	13.46887
500+	86	0.015386	3	3.488372
total	558933	100	531248	95.04681

Table 3.2: Distribution of sequences lengths obtained from 6 deeply sequenced healthy mycobiomes and their proportion of poorly aligning sequences.

3.8 Assessing reproducibility and stochastic effects

An additional concern in microbial sequencing is the potential for stochastic effects. While PCR is the main focus of chance events and bias in sequencing experiments[41], other variables still play a role in the randomness of results. To minimize the concern of stochastic effects due to PCR, reactions were performed in triplicate and a low annealing temperature was used to accommodate for potential mutations in primer binding sites. Four separate iterations of library preparation were performed on extracted fungal DNA from 6 different healthy individuals to assess variables introduced by differences in template amount, and individual scientist in a separate location for PCR preparation: 1) 125 ng gDNA template, preparer A, 2) 250 ng gDNA template, preparer A, 3) 125 ng gDNA template, preparer B, 4) 250 ng gDNA template preparer B. Between 52 and 76% of the taxa in the four iterations for a single sample (post-filtering through dimers script, DeconSeq, and QIIME) were not reproducible for more than one iteration (data not shown). At first glance this suggested that variables may indeed have an effect on the representation of taxa. In the cases of two of our samples, only 2% and 1% of taxa were present

in all iterations, respectively. However, when considering the number of sequences representing the taxa present in all iterations of a single sample, they constituted the majority of the data (Figure 3.4). As an example, the remaining sequences that appeared in at least 3 out of 4 iterations for sample 50 belonged to 120 taxa, but any of these individual taxa were at most represented by a total of 380 sequences (0.01% of total sequences collected for combined iterations of sample 50).

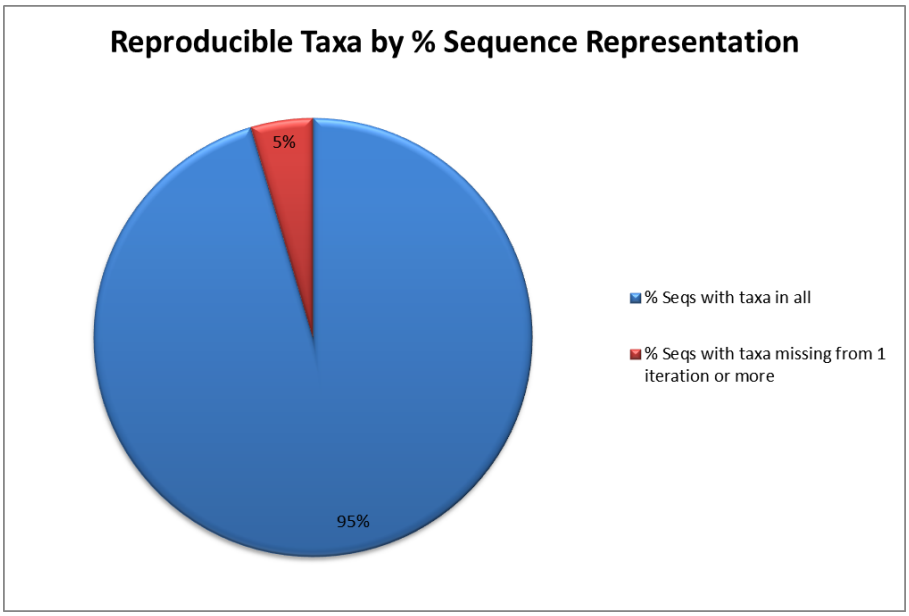


Figure 3.4: Pie chart of percentage of sequences representing taxa shared across all four iterations for 6 healthy subjects

To assess biases due to template amount vs. individual scientist and location of PCR preparation, iterations sharing a common parameter were compared, (for example, both 250 ng gDNA template amounts vs both 125 ng gDNA template amounts). The hypothesis was that if biases were created by introducing a new parameter, then there would be more shared taxa between two common parameters than between two random parameters. In every instance, the average shared number of taxa between two common parameters never exceeded the average number of shared taxa between random parameters (Figure 3.5). This meant that using a smaller starting amount of gDNA template would not significantly affect the outcome of the results, potentially allowing for more experiments to be performed with leftover gDNA, and

potentially beneficial if complications were experienced at any point in the pipeline. These data also provide evidence that sample handling in separate locations and by separate scientists will not affect the outcome of experimental results any more than random effects due to the nature of such experiments. These data, combined with the fact that these shared taxa represent less than 5% of all sequences obtained, provides support for removal of minor components from mycoprofiles to get an accurate representation of reproducible and legitimate community membership. In hindsight, these numbers were also influenced by poor taxonomic assignments to reference sequences and redundant genera (discussed in 3.9 and 3.10, respectively), which reduce their proportion to less than 5% of all data obtained for these samples and indicate that biases for our processing and handling were situationally minimal.

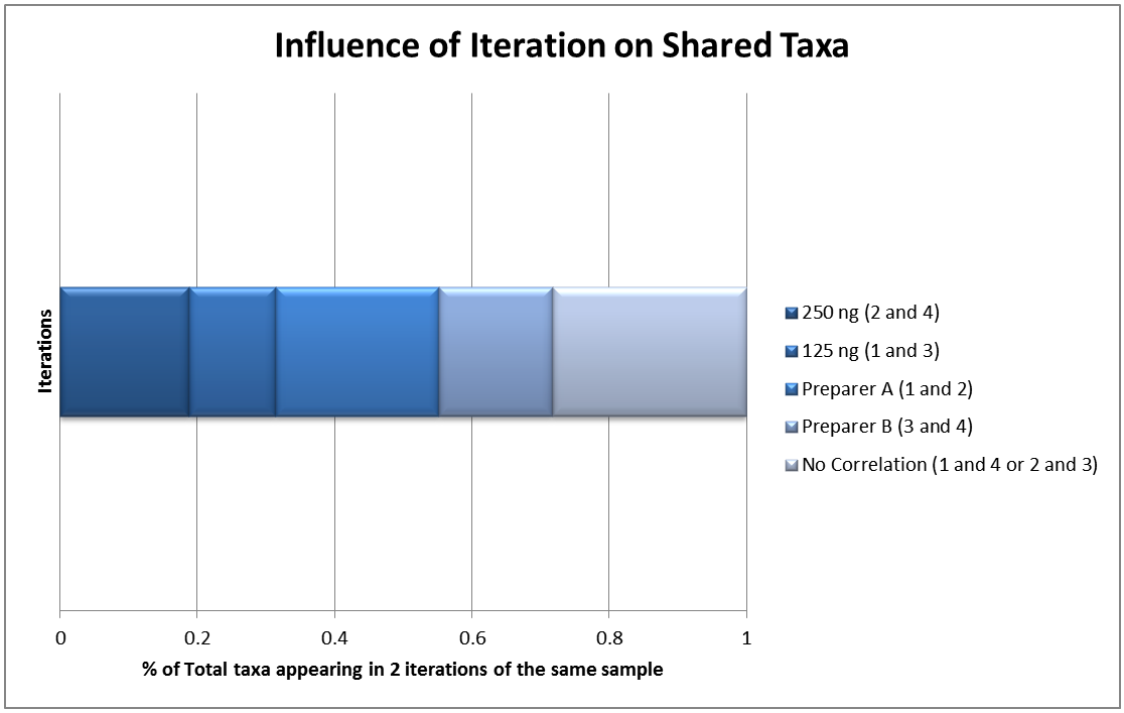


Figure 3.5: Bar graph of average representation of taxa shared between 2 iterations of a sample across common PCR preparation parameters tested

3.9 Identifying legitimate taxa

3.9.1 Using DeconSeq to remove contamination

As metagenomic amplicon surveys attempt to universally amplify from multiple cell sources, non-specific amplification often occurs. Degenerate primers and lower annealing temperatures are measures taken to broadly target all DNA sources of interest, but also promote irrelevant amplification of other sources that are commonly found in the niches with cells of interest. A tool to remove sequences matching well characterized sources of bacterial and human DNA is needed to remove such sequences as to simplify downstream analysis. DeconSeq was therefore implemented for removal of sequences matching custom alignment and query coverage thresholds. A plot of matches to DeconSeq databases showed contamination at >80% alignment with query coverage >5%. Two settings were chosen to see how they would affect the taxonomic representation of sequences 1) a “loose” setting at ≥80% alignment with ≥10% query coverage and 2) a “strict” setting at the default setting of ≥94% alignment identity and ≥90% query coverage. Loose settings removed some sequences matching to relevant oral fungi such as *Malassezia*, *Candida,* and *Alternaria* with e-values to these genera of as low as $1x10^{-168}$. Strict DeconSeq settings eliminated only sequences that were illegitimately classified (e-values for fungal matches as great as $1x10^{-4}$). This setting allowed for fine tuning of our dataset, while conservatively retaining relevant fungal taxa, and was applied to all downstream sample analyses.

3.9.2 Using a Blast statistic to develop a taxonomy based ID screen

While recent oral mycobiome studies have used alignment identity thresholds (generally 97–98%) to assign species identifications to ITS1 sequences, the suitability of this practice has been questioned[42]. We found that this standard resulted in reductions in representation of taxa that were abundant, frequent and potentially biologically meaningful. As an example, of the 18,914 sequences assigned to *Emericella nidulans*, an alignment threshold of 97% eliminated

13,601 (72%), leading to an underrepresentation of a known opportunistic pathogen. Thresholds of 90% are sometimes used for genus-level identifications, but these also have potential problems. We investigated E-value thresholds as a supplemental identification metric by mapping increasingly stringent thresholds (representing an arbitrary doubling, tripling and quadrupling of the exponent) onto a subset of sequence assignments for a single subject (#50, the most diverse individual sampled in our study, Figure 3.6).

Figure 3.6 summarizes the effects of increasing E-value thresholds, where Expect-value number denotation represents the exponent ($1x10^{-21}$ = E-21= 1E-21). The least restrictive E-21 removed 134 genera assignments, representing sequences that were at counts less than 4 and plant derived sequences. Other assignments in this interval were to the genera *Neopaxillus*, *Mortierella*, and *Ramicandelaber* and were characterized by poor E-values (>−21) driven by 18S and/or 5.8S alignments; these three taxa appeared in only marginally better intervals at E-24. Only 4% of the sequences removed by this threshold represented genus level assignments also identified by much stronger matches in the same individual (e.g. *Saccharomyces, Pichia, Cordyceps, Cortinarius*). We concluded that there was no loss of fungal genera by imposing the E-21 threshold in curation of the taxon assignment dataset. At the most restrictive interval of E-64 to E-84, neither sequences below the abundance threshold (4 counts), nor sequences derived from plants were present. Assignments that did not reach genus level resolution were present, but minimal (4%). The vast majority of genera assignments (96%) were also included in even stronger E-values, lending support to the conclusion that taxonomic assignments in this interval represented authentic fungal components. As evidence of assurance in taxon identifications, 99% of the sequence assignments in this entire dataset (sample #50) were stronger than the E-63 restrictive threshold, and 97% met an E-95 threshold suggested previously as a basis for confidence in genera assignments[43].

Our goal was to choose a threshold from intermediate intervals that would achieve both confident identifications and “conservative flexibility” for natural variants. In the E-22 to E-42 interval, sequences below the count thresholds and ones based on plant identities remained; almost half (46%) of the sequence assignments were to three genera with no E-value stronger than E-42 in the entire sequence set. Included in this group were the plants *Osmorhiza* (NCBI BLAST match to carrot,) and *Tilia* (NCBI BLAST match to tomato). The genus *Coniosporium* (a rock-inhabiting fungus) was assigned to sequences in this interval based on a poor alignment;

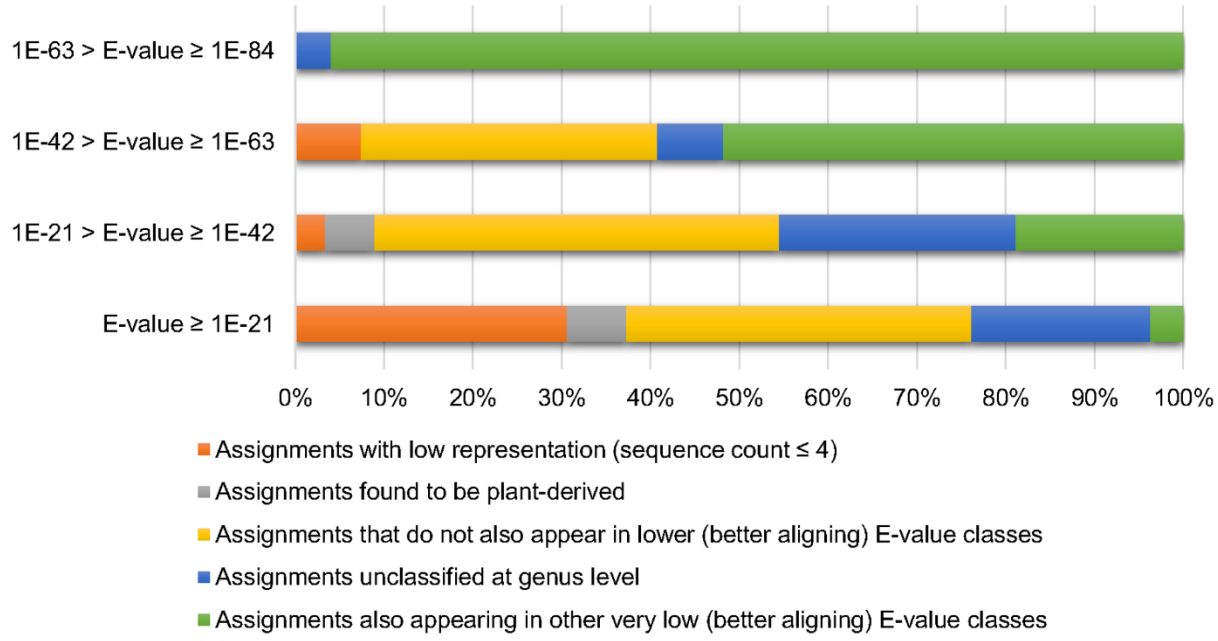


Figure 3.6: Bar graph of distributions of indicators for relevant and irrelevant taxonomic assignments at multiple e-value classes

however, based on strong alignment and E-value to an uncultured fungus, this most likely represented a legitimate fungal sequence that cannot be assigned with confidence to any genus based on current reference databases. About a quarter of the sequence assignments (27%) were to the less precise phylum level (Basidiomycetes, endophytic). The percentage of sequence assignments to genera also included in much stronger E-values (−52 through −135) increased to 19%, distributed over five genera (*Candida*, *Cortinarius*, *Cryptococcus*,

Mycosphaerella, and *Pichia*). In the next most stringent interval of E-43 to E-63, plant-derived assignments disappeared and phylum/class assignments were reduced, all to the Dothiodiomycetes. More than half (52%) of the sequence assignments were to five genera that were also included in much stronger E-value groups, *Candida, Malassezia, Mrakia, Mycosphaerella*, and *Pichia*. Based on these findings, we adopted an E-value threshold of ≤−42 for inclusion in the curated assignment list. We also note that the alignments that met this threshold had bit scores that were ≥ the 200 bit score filter adopted in a study for ITS2 amplicons[19]. We confirmed the validity of our E-value threshold by evaluating its performance against sequence sets for all five remaining subjects. Whether the subject represented individual variation similarly to the highly diverse subject 50 or a less diverse community such as subject 51, the results were still fully consistent with the details provided in this section with respect to the kinds of inappropriate taxa that were eliminated. The finding that E-values significantly lower than those routinely deemed as acceptable could still represent spurious assignments to fungal genera is an important one that can result in misleading interpretations about fungal community members. An intriguing aspect of the E-value threshold is that this single filter effectively removes low abundance representation, plant-derived amplicons, unclassifiable sequences, and those identifications based on short conserved sequences or otherwise poor alignments.

3.10 Creation of a roadmap for improved nomenclature results

Genera identified in the top 20 rankings also provided an opportunity to consider the challenges that nomenclature posed to the curation of taxon assignment datasets, subsumed under the “1N = 1F” (One Name One Fungus) initiative. We focused on genus level assignments because they represented very strong probabilities of non-random matches, and most of the taxonomic assignments in this level were derived from multiple reference sequences, often including type species. We collapsed genus assignments by considering alternate names, common knowledge

of the teleomorph (sexual form) and anamorph (asexual form) pairs, previously published recommendations, and the more specific taxonomy assignment in our dataset. We also created our own biblioinformatic examination of "common usage" (Table 3.3) as suggested by Hawksworth[30]. Our usage table was based on assignments in our own dataset, and is by no means comprehensive. In the nomenclature deconvolution process, we referred by necessity to species names as well as genera in considering sexual/asexual pairs. Moreover, given the human and biomedical orientation of this project, we added NIH NCBI publications to our biblioinformatic metrics, as well as consideration of those genera known to be common oral inhabitants, in deciding which genus to list as the "priority" one in conjoined groupings. In order to avoid the loss of information inherent in dual nomenclature[31], we listed major constituents of the conjoined genera.

Table 3.3: Common usage survey on pairs of competing genera names[1]

Name[2]	Google	Google Scholar	BSM	PubMed	Name[2]	Google	Google Scholar	BSM	PubMed
Trichosporon pullulans	7,300	977		32	Guehomyces pullulans	7,940	113		7
Trichosporon	223,000	18,300	356	1,473	Guehomyces	11,400	130	5	8
Lewia infectoria* (T)	13,200	125		1	**Alternaria infectoria (A)**	6,140	471		35
Chalastospora gossypii	691	7		1	**Alternaria malorum**	3,020	27		3
Chalastospora	3,660	21	5	3					
Lewia*	13,400	299	44	7	**Alternaria**	1,390,000	135,000	788	3,098
Pichia jadinii (T)	8,100	450		16	Cyberlindnera jadinii	84,400	11		0
Pichia	963,000	85,000	503		Cyberlindnera	106,000	30	1	
					Candida utilis(A)	120,000	14,500		813
Debaryomyces hansenii (T)	154,000	7,200		440	Candida famata (A)	34,500	2,730		130
Aspergillus nidulans (A)	501,000	45,200		4,079	Emericella nidulans (T)	90,400	3,160		282
Aspergillus	5,280,000	629,000	1,047	34,549	Emericella	144,000	5,330	149	371

[1] Searches conducted on 4/3 and 4/4 2013; Bold names are ones selected for this study. When taxa could be assigned to multiple genera, the default selection was concordance at lower taxonomic levels.
[2] (A) indicates anamorph; (T) indicates teleomorph

Cytospora chrysosperma	5,580	795		3	**Valsa sordida**	5,360	422		5
Cytospora translucens	839	14		0	**Valsa translucens**	541	11		0
Cytospora* (A)	105,000	3,140	95	33	**Valsa* (T)**	424,000	2,230	131	61
Filobasidium floriforme (T)	3,310	187		2	**Cryptocccus albidus (A)**	24,600	3,710		177
Filobasidium (T)	8,150	1,120	93	31	**Cryptococcus (A)**	1,220,000	92,400	868	9,192
Cystofilobasidium macerans (T)	1,060	17		1	**Cryptococcus macerans (A)**	5,270	168		3
Cystofilobasidium (T)	6,830	571	75	45					
Hypocrea koningii	11,600	86		0	**Trichoderma koningii**	37,100	5,590		98
Hypocrea (T)	149,000	5,550	273	290	**Trichoderma (A)**	952,000	160,000	515	4,083
Polyporus mikawai	1,170	13		0	Neofavolus mikawai	28	1		0
Polyporus	489,000	20,300	502	288	Neofavolus	389	1	0	0
Peyronellaea glomerata (A)	2,230	64		1	**Phoma glomerata(A)**	8,370	810		18
Peyronellaea	5,300	318	7	9	**Phoma**	792,000	84,500	665	425
Pyrenochaetopsis pratorum	30	1		0	**Phoma pratorum (A)**	645	3		0
Pyrenochaetopsis	1,980	9		1					
Talaromyces radicus (T)	71	1		0	**Penicillium radicum (A)**	5,290	251		0
Talaromyces	96,500	7,770	183	199	**Penicillium**	1,400,000	235,000	989	10,053
Uwebraunia commune	180	5		1	**Mycosphaerella communis**	1,300	9		1
Uwebraunia dekkeri	87	3		1	**Mycosphaerella lateralis (T)**	1,960	56		0
Uwebraunia	16,200	56	16	1	**Mycosphaerella**	395,000	24,300	700	418
Villosiclava virens	10,700	25		2	**Ustilaginoidea virens**	380,000	1,120		18
Villosiclava (T)	13,400	29	3	2	**Ustilaginoidea (A)**	389,000	1,310	28	19
Cochliobolus lunatus (T)	36,700	1,140		54	**Curvularia lunata (A)**	56,600	8,210		219
Cochliobolus verruculosus (T)	2,680	23		0	**Curvularia verruculosa (A)**	1,990	197		4
Cochliobolus (T)	194,000	17,300	194	358	**Curvularia (A)**	185,000	20,100	265	549
Coprinellus radians	2,910	71		5	**Coprinus radians**	5,370	193		2
Coprinellus flocculosus	9,850	18		0	**Coprinus flocculosus**	4,580	27		
Coprinellus micaceus	17,500	108		0	**Coprinus micaceus**	51,200	831		4
Coprinellus	59,700	522	36	24	**Coprinus**	642,000	20,600	766	642
Coprinopsis radiata	6,250	28		0	**Coprinus radiatus**	7,830	323		18
Coprinopsis	150,000	1,820	74		**Obsolete?**	Synonym			
Engyodontium album (A)	6,320	278		7	**Tritirachium album (A)**	18,800	1,910		36

Engyodontium	8,450	416	36	14	**Tritirachium**	30,000	2,720	21	49
Erythrobasidium hasegawianum	3,700	97		8	Rhodotorula hasegawae(A)	1,940	62		3
Erythrobasidium	10,500	229	55	19	Rhodotorula	304,000	27,200	369	1683
Funneliformis caledonium	6,020	6		0	**Glomus caledonium**	19,400	1,040		23
Funneliformis	16,600	126	1	7	**Glomus***	230,000	26,300	583	994
Gliomastix murorum (A)	11,300	419		3	Acremonium murorum (A)	6,740	194		1
Gliomastix	9,610	1,130	29	15	Acremonium	296,000	26,000	408	1,455
					Synonyms				
Lenzites betulinus(a)	20,701	1,111		15	Trametes betulina	1,640	26		0
Lenzites	64,000	4,050	147	27	Trametes	459,000	21,500	348	902
Ramularia grevillana (A)	1,200	17		0	**Mycosphaerella fragariae (T)**	11,400	587		3
Ramularia (A)	78,400	4,480	254	22	**Synonyms**				
Ramularia eucalypti	271	5		0					
Discostroma fuscellum (T)	476	13		0	**Seimatosporium lichenicola**	3,700	69		0
Discostroma (T)	9,130	145	37	1	**Seimatosporium (A)**	7,310	437	53	12
Sporidiobolus pararoseus (T)	3,310	478		11	Sporobolomyces shibatanus (A)	1,600	437		2
Sporidiobolus	17,000	478	116	58	Sporobolomyces	79,200	6,650	222	209
Dioszegia hungarica	3370	88		4	**Cryptococcus hungaricus**	3010	97		5
Dioszegia	6160	256	36	18					
Gibellulopsis nigrescens (A)	9410	36		0	**Verticillium nigrescens (T)**	6280	297		3
Gibellulopsis	11,700	59	2	0	**Verticillium**	712,000	63,400	496	787

3.11 Concluding remarks

Herein we have proposed a roadmap for analysis of ITS1 amplicons and have provided empirical support for recommendations made. Testing our methods through technical iterations showed that there was no effective bias in genera between PCR preparer and amount of template used. We have challenged the current standard for fungal mycobiome studies by implementing an extraction method that is efficient for obtaining DNA from spores, identifying an appropriate database for obtaining reliable fungal assignments, employing empirically determined parameters to remove and refine the genera represented in our dataset, and collapsing fungal genera to ensure genera were not inflated by synonyms names.

Chapter 4. Characterization of the healthy oral fungal microbiome

4.1 Introduction

At the time this study began, the healthy oral fungal microbiome (mycobiome) had been described by a single study by Ghannoum et al 2010, in which 13 components were presented as common occupants of the oral cavity. Following the methods in chapter 2, our results were concordant for the majority of genera, but also revealed several novel genera. This chapter summarizes our findings in the healthy mycobiomes of six individuals using a deep sequencing approach. The similarity of our findings to those in the study by Ghannoum and collaborators, although using a very different protocol, offers support for the analyses we employed.

4.2 Application of curation rules to healthy subjects

Raw reads totaling 853,653 provided 473,493 sequences after applying our bioinformatics pipeline; approximately 55.5% of original sequences remained. A first glance at the retainment percentage of Ghannoum and collaborators' that was reported (87% after applying their pipeline) suggested that our pipeline was very strict in comparison. However, after closer inspection, the percentages of unclassified fungi (sequences resulting in automated taxonomy that were only as specific as the family level) were drastically different between studies. Removal of unclassified sequences in our study resulted in a minor shift in the percentage of sequences retained (55.5% to 54.6%). Removal of unclassified sequences from Ghannoum and collaborators' analysis vastly impacted the percentage of sequences retained, reducing them from 86.8% to 48.5%. This suggests that our rate of correctly identifying fungal mycobiome components is on par with (or slightly better than) other similar studies. The reason for such a drop in reliable sequences in Ghannoum et al.'s study was likely due to a number of reasons: 1) Using a top-hit BLAST approach in reporting taxa caused matches to "unclassified and "uncultured" Genbank sequences often deposited from metagenomic studies, which would otherwise be classifiable 2) Using databases non-specific to fungi caused identifications to

Pongo and Campylobacter, misrepresenting the proportions of fungi in the oral cavity with non-fungal classifications. While our initial experiments were abundant in concatenated primer sequences (~95% of all sequences), implementing the pre-sequencing AMPure® XP and Pippin Prep™ steps greatly reduced the proportion of these artifacts in the dataset, leaving only 4.7% of sequences to be removed by our post-sequencing perl script. A closer look at the sequences filtered in this step shows a reduction in unclassified “Fungi” of ~80%. These modifications have vastly improved the productivity of sequencing space and reduced effort toward analyzing uninformative artifacts.

The DeconSeq step was used with a strict cutoff of 90% query coverage and 94% alignment identity so that only those sequences that matched well to the contamination databases would be eliminated. As a result, none of the sequences filtered by this method reached the e-value threshold of 1E-42, and are clearly nonspecific products of universal amplification from a saliva sample with different kinds of gDNA. Only 1,352 sequences were filtered with this method. Perhaps using a lower stringency for contamination would eliminate even more sequences with poor e-values before submitting to the fungal metagenomics project in the final stages of analysis. The top 20 taxa were not affected by this step.

The QIIME split_libraries step had the greatest effect on sequence numbers filtered (87% of sequences lost at this step). The majority of sequences removed in this step were due to fragment size <100 bp. The median e-value increased to $1x10^{-21}$ from $1x10^{-138}$ at the DeconSeq step, which was unexpected because filtering short sequences should improve the median match statistic and the median e-value for sequences filtered was $2x10^{-4}$ (appendix table 1). However, we found the e-value to be affected by the removal of trimming primer sequences, which acted as anchor sites to reference sequences, thereby weakening the match statistic to all sequences during the QIIME step. In the length filtering step, the pipeline restored the

median e-value to a more acceptable value ($1x10^{-123}$), suggesting that sequences <100 bp that still remained after QIIME were the greatest affecters of e-value post-QIIME.

Of all steps in the pipeline, the e-value filtering step had the greatest effect on the number of genera represented across samples. Only 32% of raw sequence genera were retained at this step, likely due to elimination of sequences that were illegitimately forced to match reference sequences to which they did not strongly align. This step was critical for removal of non-fungal sequences that contained 18S regions (plant), as evidenced by the elimination of *Lysurus,* which was the third most represented genus in the top 20.

Total genera were reduced from a maximum of 732 down to 144, eliminating the majority of singleton genera and reducing the representation of spurious assignments. Ten of the top 20 genera were eliminated after curation rules and 12 were affected by nomenclature deconvolution (Figure 4.1). Of the 17 genera listed in the fully curated top 20 (Figure 4.1), 12 were affected by nomenclature deconvolution. The genus *Cyberlindnera* was exclusively represented by its synonym, *Pichia jadinii*, so the former sequence counts were attributed to the genus *Pichia*. In turn, the genus *Pichia* was represented by three species: *jadinii, kudriavzevii*, and *membranifaciens*, all of which have the other names of *Candida utilis, Candida krusei*, and *Candida valida*, the respective anamorph forms. The *Pichia* sequence assignments were collapsed into *Candida*; the pair accounted for 0.2%–36% of sequences in individual subjects (Figure 4.3), a range in good agreement with the previously published study of Ghannoum and collaborators. While not every described species in the genus *Pichia* has a *Candida* counterpart, all *Pichia* identified in our sequence study did and were therefore appropriately combined. Across the six subjects in our study, sequences assigned to *Pichia* represented 99%, 43%, 81%, 7.7%, 6.7% and 0% of the combined *Candida* plus *Pichia* sequences (Figure 4.3). The teleomorphic genus *Gibberella* was often accompanied by its anamorphic genus *Fusarium* at identical E-values in the top 4–5 NCBI BLAST hits. In the vast majority of these cases, there

were no species assigned to *Gibberella*, but assignments to *Fusarium culmorum* were common. In other *Gibberella* assignments, the species have *Fusarium* anamorph pairs. *Gibberella* sequence assignments were collapsed into the genus *Fusarium*. The genus *Emericella* was exclusively represented by the species *nidulans*, a synonym of *Aspergillus nidulans*, so the former sequence counts were attributed to *Aspergillus*. Assignments to genus *Eurotium* were also reassigned to *Aspergillus*, its priority genus[44]. Likewise, since the genus *Lewia* was exclusively represented by the species *infectoria*, the teleomorph form of *Alternaria infectoria*, we collapsed these sequences into *Alternaria*. The synonymous teleomorph genus *Davidiella* was collapsed into its anamorph genus *Cladosporium*[30].

Nomenclature considerations for less abundant taxa also affected the top categories (Figure 4.1, panel B). The teleomorph genera *Filobasidium (F. floriforme)* and *Cystofilobasidium (C. macerans)* were collapsed into the more commonly used nomenclature of its paired anamorph genus *Cryptococcus*[30]. *Trichosporon* was represented by the species *pullulans*, another name for *Guehomyces pullulans*; both sequence assignments were included under the genus *Trichosporon* (common usage). The anamorph species *Cytospora chrysosperma* (also called *Valsa sordida*) and *Cytospora translucens* were combined into the teleomorph genus *Valsa*. The genera *Lenzites, Penicillium*, and *Phoma* also rise in the listings by cumulative abundances. Many other assignments that were also affected by nomenclature deconvolution, but not in the top 20, are included in Table 3.3.

A. Sequence Dataset	Raw data set	Following primer artifact removal and DeconSeq	Following QIIME	Following Length Filter	Following E-value Filter	Following Nomenclature Deconvolution
Total Sequence Counts	853653	812255	565350	557443	473493	N/A
% of Total Count Removed	N/A	4.90%	28.90%	0.90%	9.90%	N/A
# of Sequences Classified by FMP	853178	811796	546563	539727	473493	N/A
Average Length	240.4	248.5	257.3	259.6	257.2	N/A
Median E-value	1E-131	1E-138	1E-21	1E-123	1E-132	N/A
# of Genera (uncollapsed)	607	599	732	698	194	144*
B. Top 20 genera by sequence abundance following each curation step	*Malassezia*	*Malassezia*	*Malassezia*	*Malassezia*	*Malassezia*	*Malassezia*
	Epicoccum	*Epicoccum*	*Epicoccum*	*Epicoccum*	*Epicoccum*	*Epicoccum*
	Mortierella	*Mortierella*	*Lysurus*	*Lysurus*	*Cyberlindnera*	***Candida*** */Pichia*
	Ascomycota	*Ascomycota*	*Cyberlindnera*	*Cyberlindnera*	*Gibberella*	***Fusarium*** */Gibberella*
	Cyberlindnera	*Cyberlindnera*	*Gibberella*	*Gibberella*	*Emericella*	***Aspergillus*** */Emericella*
	Cortinarius	*Cortinarius*	*Emericella*	*Emericella*	*Alternaria*	***Alternaria*** */Lewia*
	Fungi	*Fusarium*	*Alternaria*	*Alternaria*	Fungi	Fungi
	Fusarium	*Calostoma*	*Cetrelia*	Fungi	*Lewia*	***Cladosporium*** */Davidiella*
	Calostoma	*Sydowia*	Fungi	*Lewia*	*Cladosporium*	*Ganoderma*
	Sydowia	*Emericella*	*Lewia*	*Cladosporium*	*Davidiella*	*Mrakia*
	Emericella	*Alternaria*	*Cladosporium*	*Davidiella*	*Ganoderma*	***Cryptococcus*** */Filo-, Cysto-filobasidium*
	Alternaria	*Orpinomyces*	*Davidiella*	*Ascomycota*	*Mrakia*	*Sporobolomyces/Sporidiobolus*
	Orpinomyces	*Cladosporium*	*Ascomycota*	*Ganoderma*	*Ascomycota*	*Irpex*
	Cladosporium	*Serpula*	*Ganoderma*	*Mrakia*	*Candida*	***Trichosporon*** */Guehomyces*
	Serpula	*Fungi*	*Mrakia*	*Candida*	*Fusarium*	*Phenophora*
	Ganoderma	*Ganoderma*	*Candida*	*Fusarium*	*Cryptococcus*	***Cytospora*** */Valsa*
	Leptosphaeria	*Funneliformis*	*Fusarium*	*Saccharomyces*	*Diversisporales*	***Lenzites*** */Trametes*
	Physciella	*Physciella*	*Saccharomyces*	*Periconia*	*Fusarium*	***Penicillium/*** *Talaromyces*
	Funneliformis	*Saccharomyces*	*Periconia*	*Cryptococcus*	*Sporobolomyces*	*Udeniomyces*
	Glomus	*Glomus*	*Cryptococcus*	*Diversisporales*	*Irpex*	***Phoma*** */Peyronellaea/Pyrenochaetopsis*

Figure 4.1: Stepwise effects of bioinformatics pipeline and nomenclature deconvolution on A) basic sequence statistics and B) on the top 20 represented genera of the six combined healthy individuals. Yellow highlighting represents genera removed that were found to have poor alignment. Gray represents plant-derived sequences incorrectly assigned fungal identity. Blue represents fungal assignments above the genus level.

4.3 Comparison of core mycobiomes

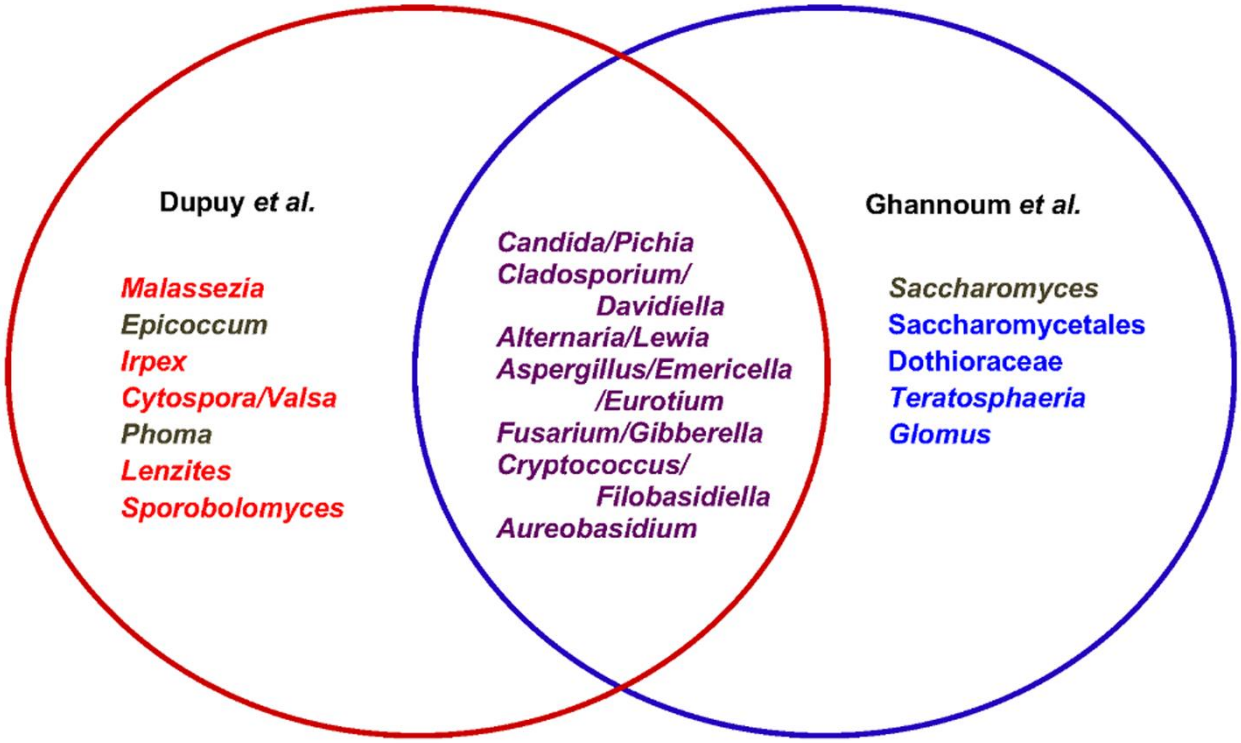
Ghannoum and collaborators[23] identified components of the healthy oral mycobiome using a 1% abundance cutoff with an average of 1,700 reads per sample, likely missing community members in low abundance. A deep sequencing approach must be used to identify core members observed frequently at both low and high abundances to assess their role in health. Of the 34,049 sequences, 36.1% used in Ghannoum's study were assigned to an ambiguous group of "unculturable" fungi. One possibility is that these assignments were made to other environmental survey-based submissions in the NCBI database that omit fungal taxonomy. We hypothesized that application of our extraction method and curation steps would discover novel

oral commensals, reduce the proportion of “unculturable” assignments, and refine the core oral mycobiome through more comprehensive profiles.

Fourteen genera met qualifications as core community members in health in our study (Figure 4.3). Included in these were all but three core oral genera present in Ghannoum’s study. The shared genera included *Candida*, *Cladosporium*, *Aureobasidium*, *Alternaria*, *Aspergillus/Eurotium*, *Fusarium/Gibberella*, and *Cryptococcus*. The three genera absent from our representation of the core mycobiome are *Teratosphaeria*, *Saccharomyces*, and *Glomus*. In this study, *Teratosphaeria* may have been subsumed in a category such as “unclassified Capnodiales” or the genus *Mycosphaerella*, which is polyphyletic and includes some species of *Teratosphaeria*. Even still, only 56 counts of *Mycosphaerella* and 7 counts of “unclassified Capnodiales” were observed. *Saccharomyces* was represented in 3 subjects, but accounted for only 0.1-0.5% of sequences in each sample. Twenty-four counts of *Glomus* were obtained from two samples. Relatively high counts of *Saccharomyces* and *Glomus* assignments were observed in our raw dataset, but the application of the sequence curation pipeline greatly reduced these numbers. Our findings suggest that *Glomus* may not be a legitimate core member as originally proposed by Ghannoum and colleagues, and that species of *Saccharomyces* are markedly underrepresented compared to other core genera.

The comparison between the results of Ghannoum and colleagues and our study is summarized in Figure 4.2. Ghannoum and colleagues reported thirteen components in the basal mycobiome: *Alternaria, Aspergilllus, Aureobasidium, Candida, Cladosporium, Cryptococcus*, Dothioraceae, *Eurotium, Fusarium, Glomus, Saccharomyces*, Saccharomycetales, and *Teratosphaeria*. Of the eleven that were identified at the genus level, our study also found eight of these in more than half of the subjects (genus followed by frequency and range): *Alternaria/Lewia* (100%, 0.01–

Figure 4.2: Venn diagram of the relationships between results from the two studies of the human oral mycobiome. Shared genera are indicated in the overlap (purple font) between the current study (Dupuy *et al*.[27], red font) and the previously published study (Ghannoum *et al*.[23], blue font). Genera in brown are shared between the two studies but failed to meet thresholds in one or the other.



7.07%), *Aspergilllus/Emericella/Eurotium* (100%, 0.001–10.27%), *Candida/Pichia* (100%, 0.12–35.86%), *Cladosporium/Davidiella* (100%, 0.06–8.26%), *Cryptococcus/Filobasidiella* (100%, 0.05–0.81%), *Fusarium/Gibberella* (83%, 0.01–18.35%), and *Aureobasidium* (67%, 0.004–0.08%). The genera *Saccharomyces* (50%), *Epicoccum* and *Phoma* were also shared, but were below thresholds in one study or the other. *Epicoccum* is found in indoor house dust samples[45], has been identified in air samples in buildings, including in the Northeastern U.S. where all of our subjects lived[46], and is a well-known air allergen. *Phoma* and *Epicoccum* were also identified as components of indoor fungal composition in temperate zones[47], and may represent environmental acquisitions specific to geography. While *Epicoccum* has not been associated with human infections, it has been identified as a source of allergens, and some species

possess antifungal activity against pathogenic plant fungi. *Phoma* species were found to be causative of infection in a transplant recipient[48].

GENUS	Freq.	Normalized %	Abundance	50 185228	51 128378	52 33969	54 3489	56 67452	57 54780
Malassezia	6	37.763 %	221719						
Epicoccum	6	33.176 %	114855						
Candida/Pichia	6	9.566 %	39747						
Cladosporium/Davidiella	6	3.012 %	4852						
Alternaria/Lewia	6	1.858 %	13551						
Unclassified	6	1.793 %	7524						
Aspergillus/Emericella/Eurotium	6	1.769 %	19196						
Irpex	6	0.879 %	709						
Cytospora/Valsa	6	0.541 %	521						
Cryptococcus/Filo-, Cystofilobasidium	6	0.220 %	1775						
Phoma/Peyronellaea, Pyrenochaetopsis	6	0.059 %	284						
Fusarium/Gibberella	5	3.726 %	35175						
Sporidiobolus/Sporobolomyces	5	2.262 %	1573						
Lenzites/Trametes	5	0.252 %	438						
Aureobasidium	4	0.029 %	159						

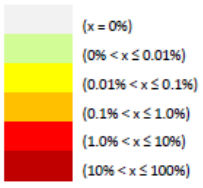


Figure 4.3: Frequency, abundance, and distribution of genera occurring in >50% of the six subjects. Genera ordered by frequency of occurrence, with normalized representation and sequence counts (columns 2, 3, 4). Heat map depiction (columns 5–10) summarizes qualitative and quantitative distribution of genera in six individuals (50, 51, 52, 54, 56, and 57) and depth of sequencing for each subject (row 2). Values within individual heat map cells are the percentage representation within that subject.

Four components of the core mycobiome proposed by Ghannoum and colleagues (*Glomus, Teratosphaeria*, Saccharomycetales and Dothioraceae) were absent from our high frequency listing. Several explanations could account for such discrepancies. First, the fungi were simply not present in the subjects sampled. Second, the identifications were spurious ones. Third, the taxonomic assignments were made to different levels in the two studies. In our case, sequences initially assigned to the genus *Glomus* were found in high abundance, but were eliminated following primer artifact and DeconSeq filters (Figure 4.1). Although *Teratosphaeria* was not detected in any of our subjects, we did have unclassified sequences in the order to which it belongs, Pleosporales (Figure 4.4). Two of the taxonomic assignments in the study by Ghannoum *et al.* were at the higher taxonomic ranks of order (Saccharomycetales) and family

(Dothioraceae). Saccharomycetales is a large monophyletic order with about 1,000 known species across many genera [49], several of which were identified in our study. We note that one member genus in the family Dothioraceae, *Sydowia*, was a prominent taxon assignment in our study before being eliminated by early step sequence curation filters (Figure 4.1).

Five genera (frequencies and ranges) were identified in high frequency in our study (Figure 4.3) but were not part of the basal oral mycobiome proposed by Ghannoum *et al*.: *Malassezia* (100%, 12.98–96.01%), *Irpex* (100%, 0.02–4.07%), *Cytospora/Valsa* (100%, 0.005–2.92%), *Lenzites/Trametes* (100%, 0.02–1.18%), and *Sporobolomyces/Sporidiobolus* (100%, 0.01–12.87%). *Malassezia* is discussed in the following section; all of the other four genera are common soil and/or plant pathogens that are widespread in common environmental sources in temperate zones. Members of three of these genera, *Irpex* [50], *Cytospora/Valsa* [51]and *Sporobolomyces/Sporidiobolus* [48] were previously identified as causative agents in infections in immune compromised persons. In this context, it seems prudent to consider these taxa worthy of attention in future studies.

Species of *Irpex* and *Cytospora*/*Valsa* were all identified at less than 1% of sequences and were found in all subjects. Whether introduced to the oral mycobiome as transients through ingestion or inhalation, the sensitivity of detection for these genera is particularly important because of their previous implications as opportunistic human pathogens. *Irpex lacteus* was identified in the case of a feverish 9-year-old girl being treated for acute lymphoblastic leukemia[50]. *Valsa sordida* was isolated from a 55 year old woman suffering sinusitis due to complications with treatment of acute myeloid leukemia[51]. While both were administered amphotericin B, an antifungal, only the former was successfully treated. The latter died two weeks after the antifungal was first given. In both cases, environmental strains of these fungi do not proliferate at temperatures above 30 °C[50,51]. The adaptability of *Irpex* and *Cytospora/Valsa* to increase maximum growth

temperatures to 37 °C illustrates their future potential in opportunistic mycoses, especially in instances such as these, where host immunity is compromised by cancer therapies.

Sporobolomyces/*Sporidiobolus* was represented in four of the six individuals, with 13% of sequences from one subject classified to these genera. Although considered a rare invasive pathogen, *Sporobolomyces* spp. have been identified in several cases. This genus was detected in a bone marrow biopsy from an AIDS patient and in other cases of cutaneous and nasal infections[52]. Its presence at high abundance in one of our subjects suggests that potential proliferation of *Sporobolomyces* would not be inhibited in the oral niche, nor is it inherently dangerous in an immunocompetent individual.

4.4 Malassezia

The most unexpected finding from our study was the presence in all six subjects, at high abundances from 13% to 96%, of the genus *Malassezia* (Figure 4.3), one that was not identified by Ghannoum *et al*.[23]. On the experimental side, the results from our negative controls strongly support the conclusion that *Malassezia* sequences were not introduced during the processing of samples. There is additional support from the literature for the argument to include *Malassezia*, a recognized commensal and pathogen in humans and other mammals[53], as a member of the basal mycobiome. Well known to cause a variety of skin disorders[54], *Malassezia* was recently identified by metagenomic sequencing as associated with scalp disorders such as dandruff[55]. More directly relevant to the oral cavity, one of the main entryways for microbes into the airways, *Malassezia* was also discovered by metagenomic sequencing in the sputum of cystic fibrosis patients[19]. The mouth is the point of entry into the gastrointestinal tract, and *Malassezia* was identified by culture-independent, Sanger sequencing of cloned 18S amplicons from human stool[24]. Directly relevant to the mammalian oral cavity, *Malassezia* species were shown to occupy the mouth of dogs[56] and underwent zoonotic transfer by health care professionals from their dogs to neonates where they were responsible for serious infections[57]. Since the four more

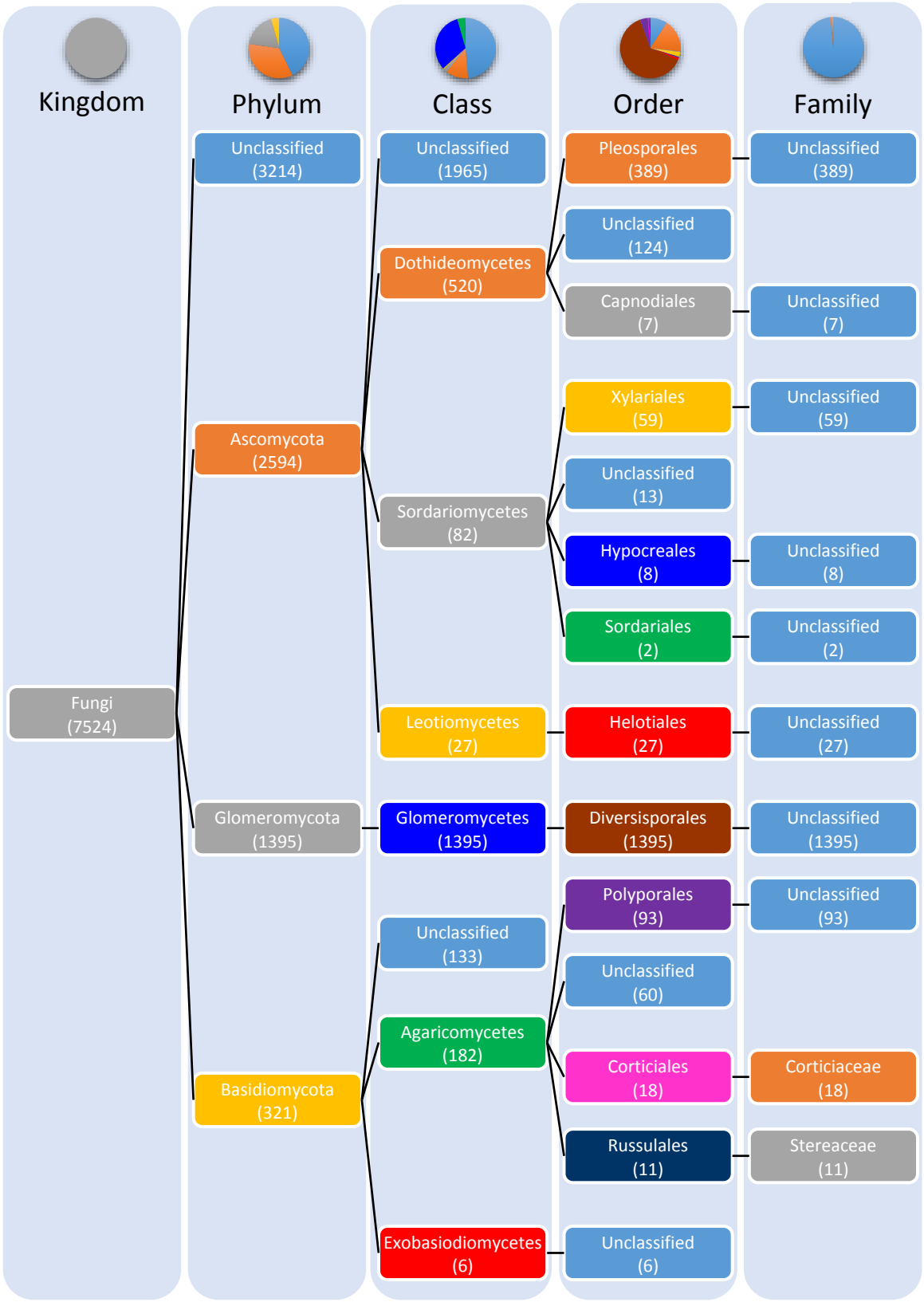
recent and culture-independent metagenomic studies that identified *Malassezia* in human biocompartments used subjects from worldwide geographies, different protocols for molecular biology, and different rules for sequence and taxon curation, the consensus on this genus was compelling evidence for its inclusion as a legitimate member of the basal oral mycobiome. It is noteworthy that each of the metagenomic studies reporting *Malassezia*, including ours, employed relatively harsh extraction protocols that were more likely to recover *Malassezia* DNA since species in this genus are known to have especially thick cell walls. The relatively harsh bead breaking step we employed did not appear to unfavorably impact recovery of other salivary genera given the agreement between our study and the prior report on the salivary mycobiome[23].

Additional characteristics of *Malassezia* species[53], have probably contributed to the previous failures to recognize *Malassezia* as a prominent oral commensal. First, culture-based methods may not have captured *Malassezia* species since most have growth requirements for lipids and require specialized culture media [58]. Second, the taxonomy and nomenclature issues also complicated studies of *Malassezia*, which are dimorphic fungi (yeast and mycelial phases) that have been placed in multiple genera. Although much of the taxonomy within the genus has been sorted out, studies undertaken before the mid-1990s and those without knowledge of the recent resolutions of nomenclature may have missed this genus.

4.5 Unclassified sequences

Only 1.6% of total sequences were deemed “unclassified fungi”, of which 57% were classified to family-level or above. Additional steps were taken to improve the representation of unclassified fungi by identifying the reference sequences to which they were matched and submitting those reference sequences (typically longer sequences containing 18S, ITS1, 5.8S, ITS2, and 28S)

Figure 4.4: Distribution of unclassified sequences at all taxonomic levels above genus for six healthy subjects



back to Genbank to see if more specific taxonomies would emerge. Resolved taxonomies were accepted if sequence alignments matched a sequence with ≥90 sequence coverage and ≥99% alignment identity, following strict recommendations[59]. Of all unresolved sequences, references matched to a more specific taxonomy than their original designations for ~60% of unclassified accession numbers identified. Measures were taken to redefine relative abundance data for the sequences affected. More often than not, we found that the resolved taxonomies were already present in the dataset, and we were able to remove unclassified designations, while boosting the numbers of a genus that was already identified. *Cladosporium* was one of the most strongly affected genera, whereby resolved counts could often exceed the number that were reported by our standard pipeline.

4.6 Concluding remarks

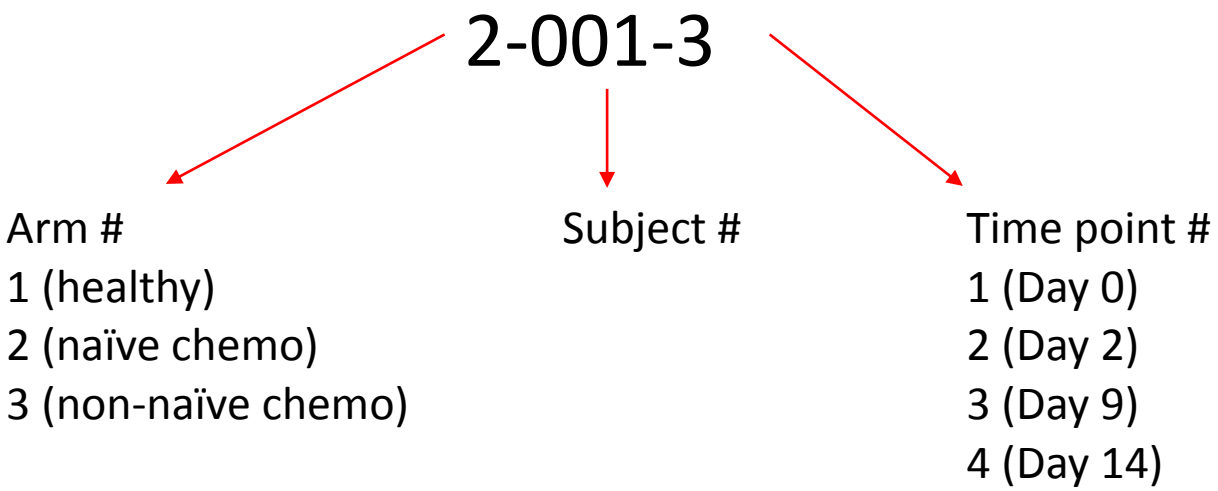
Herein we have assessed our pipeline using a deep sequenced dataset of ITS1 sequences from the saliva of 6 healthy subjects. We showed that each step creates an improved curated dataset with greater legitimate representation of fungi at the genus level. Contributions toward examining sequences classified more broadly than the genus level also improved the representation of legitimate sequence matches, decreasing the percentage of unclassified sequences compared to the study by Ghannoum and colleagues. We provided supporting evidence for new members of the oral community and confirmed the presence of those already known.

Chapter 5. Variation in healthy subjects

5.1 Introduction

The goal of this chapter is to demonstrate how healthy individuals vary in their oral mycobial profile between subjects and over the course of two weeks within a subject. This information will help us account for natural variation when considering results of a time course study for chemotherapy patients. First, we used healthy subjects to determine minimum sequencing effort required for capturing the majority of diversity in a sample, so that minor, but reproducible components are not overlooked. This was accomplished by generating theoretical rarefaction curves and conducting an empirical evaluation of the effect of sequencing depth on genus recovery. Second, we assessed for subject mycoprofiles that could be partitioned by mycobial content and prevalence. Grouping subjects is especially important for comparing trends in patients with similar profiles that are affected by disease states. Third, we examined the stability of mycoprofiles over a two time point (day 1, day 14) longitudinal study. Our hypothesis was that, in the absence of changes in subject lifestyle activities, the mycoprofile will stay essentially the same over 14 days. Saliva samples were given designations to denote the cohort to which they belong, the subject, and the time point (Figure 5.1).

Figure 5.1: Breakdown of sample notation for oral mucositis study.



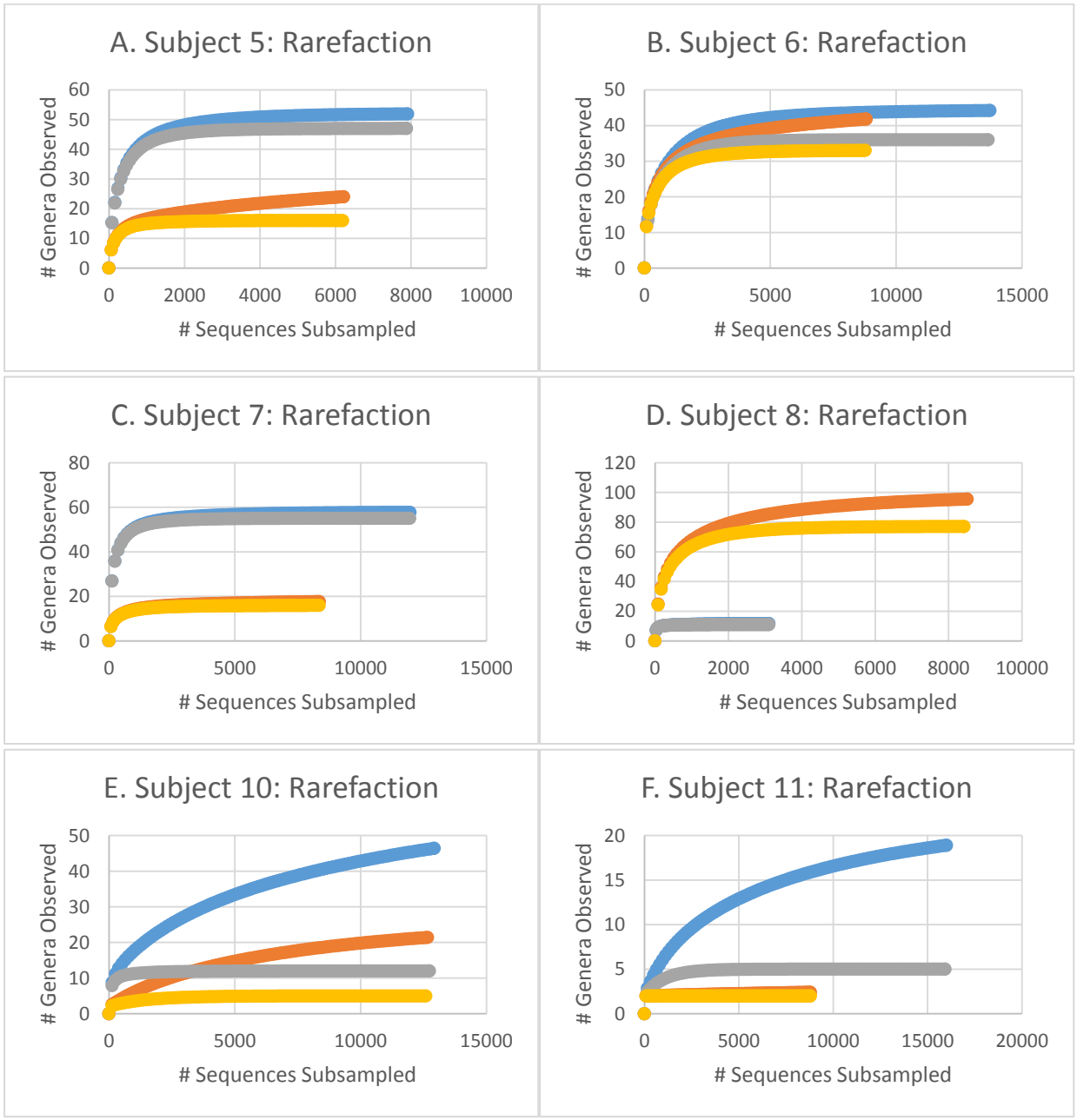
5.2 Minimum sequencing effort for saliva samples

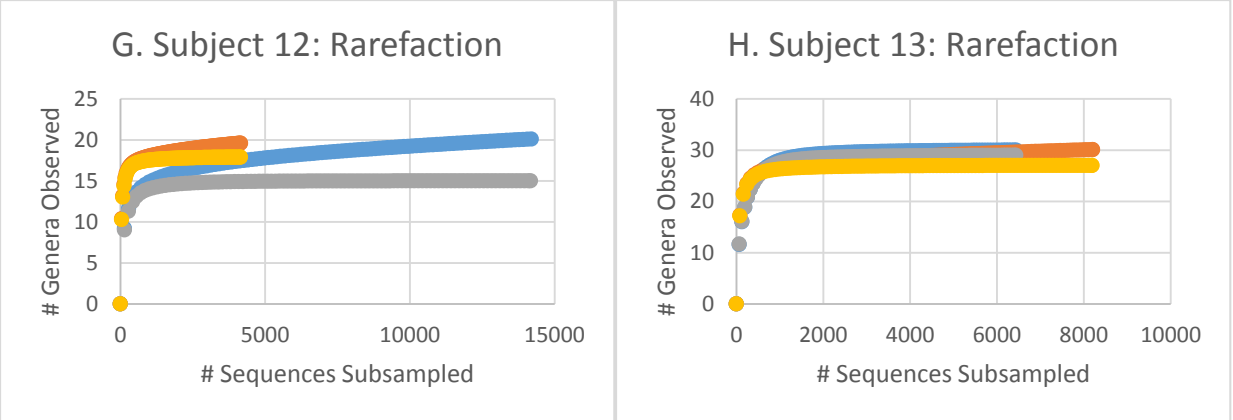
Since samples vary in sequencing efficacy, it is useful to determine the minimum sequencing effort required to capture inherent fungal diversity. Of the 30 arm 1-healthy subjects, 24 were adequate for DNA sequencing and analysis. Four subjects not suitable for sequencing were as follows. Subjects 1-4 and 15 yielded NanoDrop readings that were similar to values obtained for reagent blanks (Section 3.5) and were unable to be amplified. Subject 9 was not processed because no saliva remained after bacterial sequencing. Subject 22 was only successful in amplification and sequencing of the first time point. This left 47 time points to represent the remaining 24 subjects. After applying our pipeline, 480,121 sequences were retained with an average of 10,215 sequences per sample. However, some samples were represented by only a few hundred sequences. While some studies suggest as few as 40 sequences are needed to categorize a sample[60], our negative controls revealed non-legitimate, but high quality sequences that pass through the pipeline in low amounts (Chapter 3). Such spurious sequences likely misconstrue the diversity of a sample, especially when sequencing effort is low.

5.2.1 Rarefaction curves

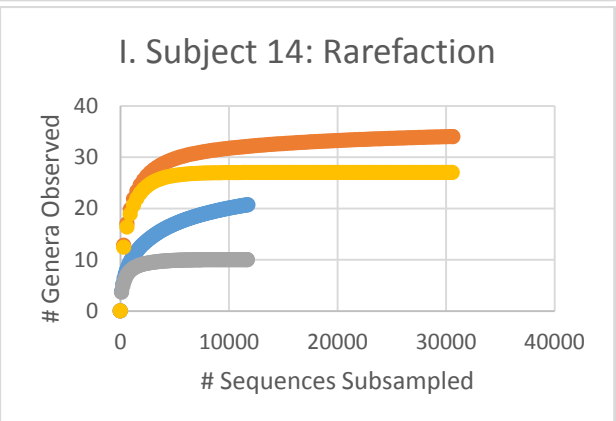
In order to determine the minimum amount of sequences required to capture the majority of the diversity in an oral sample, rarefaction curves were generated using an implementation of R in VAMPS. Ten iterations were performed for each of the 47 samples and average values were plotted to visualize a shared minimum sequence value where the curve reaches an asymptote (Figure 5.1 panels A-X: blue and orange). The point at which the curves commonly plateau is the required sequencing effort for oral samples to be confidently analyzed here and in future studies. Many of the curves never reached asymptote, including samples with some of the greatest sequencing efforts (time point 4 for subject 18 and 14 with 39,201 and 30,637 sequences respectively). This was likely due to occurrences of taxa in lower abundances that inflated the slope of the curve.

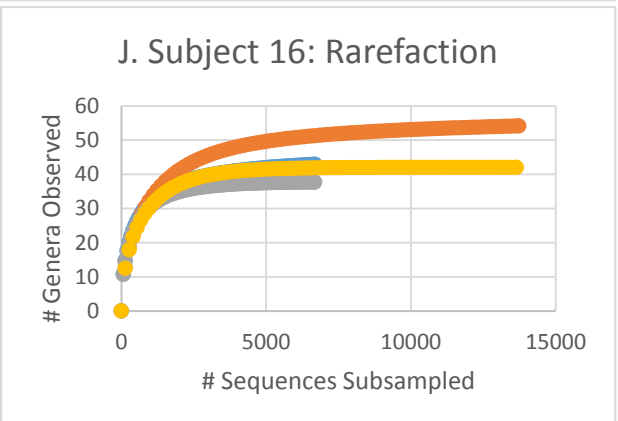
Figure 5.2: Panels A-X: Average of 10 rarefaction curves generated for arm 1-healthy subjects time points 1 (blue) and 4 (orange) with all genera and for time points 1 (gray) and 4 (yellow) after removal of genera with relative abundance < 0.1%.

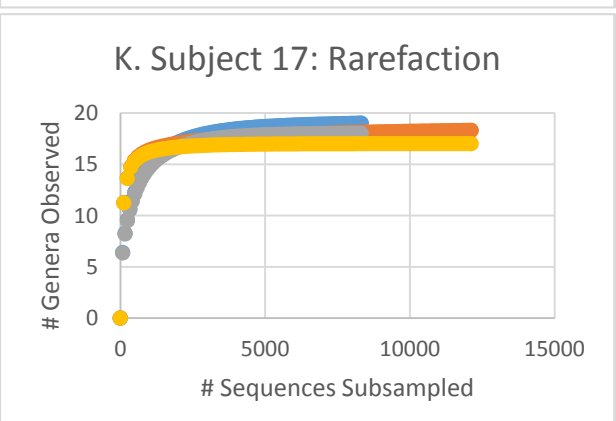


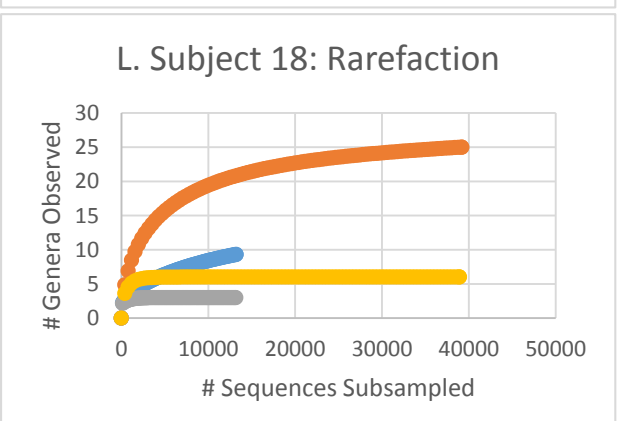

G. Subject 12: Rarefaction
Genera Observed
0
5
10
15
20
25
0
5000
10000
15000
Sequences Subsampled

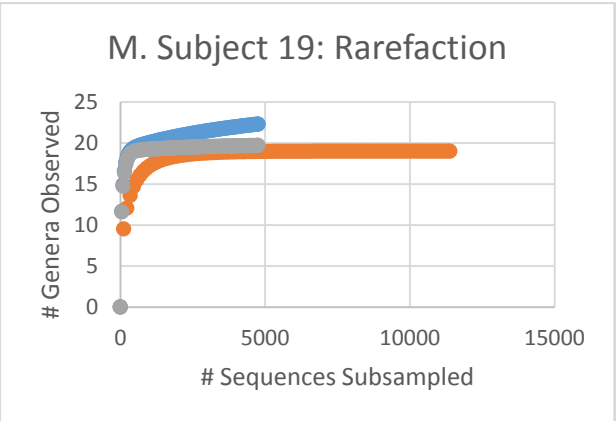
H. Subject 13: Rarefaction
Genera Observed
0
10
20
30
40
0
2000
4000
6000
8000
10000
Sequences Subsampled

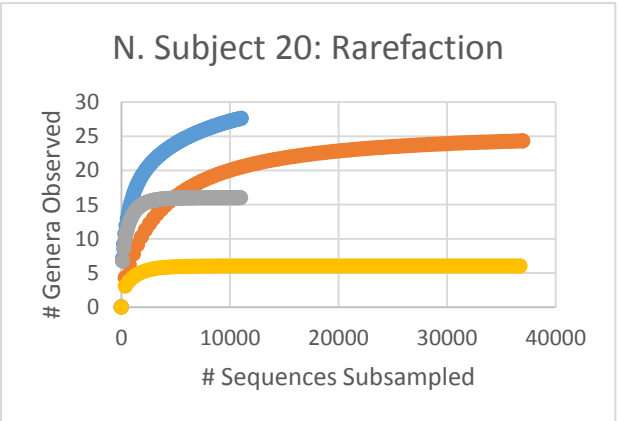

I. Subject 14: Rarefaction
Genera Observed
0
10
20
30
40
0
10000
20000
30000
40000
Sequences Subsampled

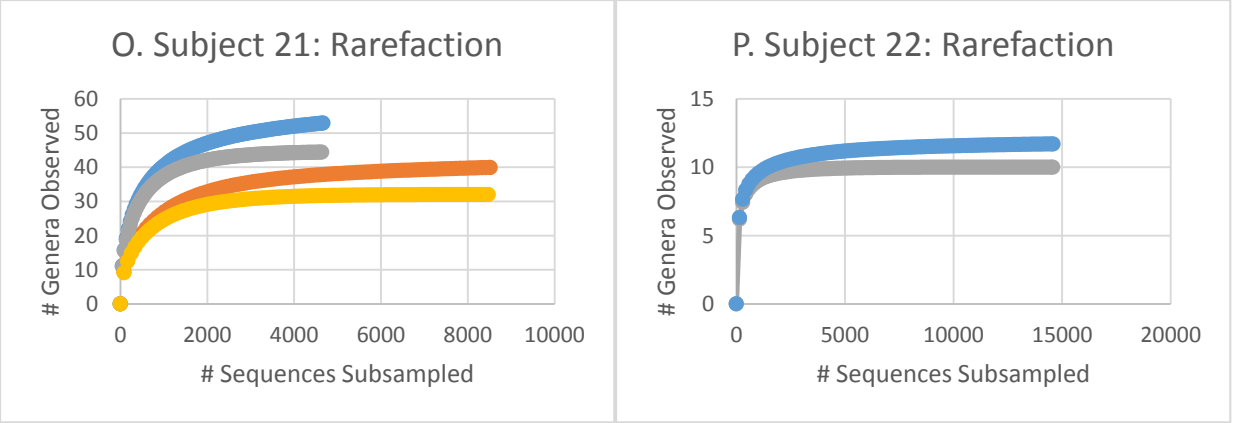

J. Subject 16: Rarefaction
Genera Observed
0
10
20
30
40
50
60
0
5000
10000
15000
Sequences Subsampled


K. Subject 17: Rarefaction
Genera Observed
0
5
10
15
20
0
5000
10000
15000
Sequences Subsampled

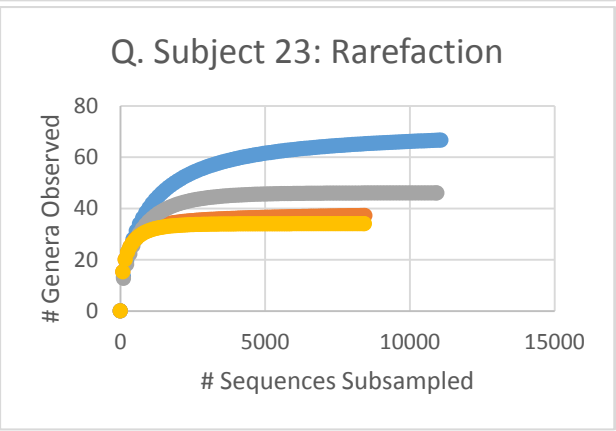

L. Subject 18: Rarefaction
Genera Observed
0
5
10
15
20
25
30
0
10000
20000
30000
40000
50000
Sequences Subsampled

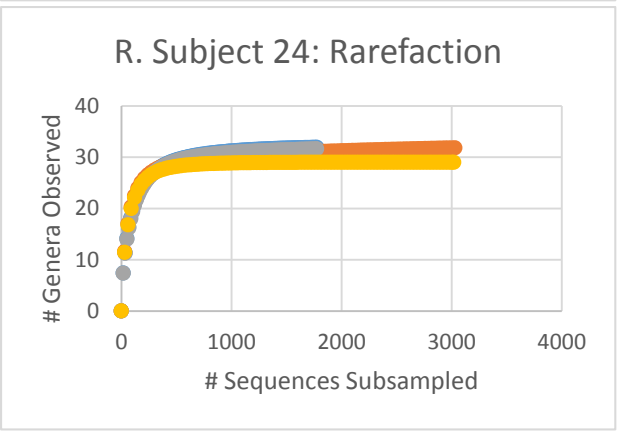

M. Subject 19: Rarefaction
Genera Observed
0
5
10
15
20
25
0
5000
10000
15000
Sequences Subsampled

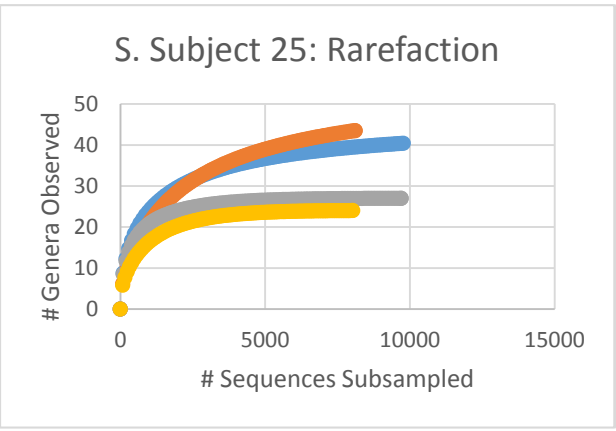

N. Subject 20: Rarefaction
Genera Observed
0
5
10
15
20
25
30
0
10000
20000
30000
40000
Sequences Subsampled

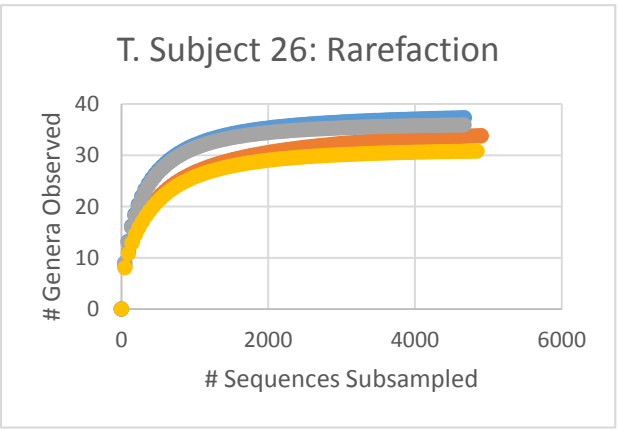

O. Subject 21: Rarefaction
Genera Observed
60
50
40
30
20
10
0
0
2000
4000
6000
8000
10000
Sequences Subsampled

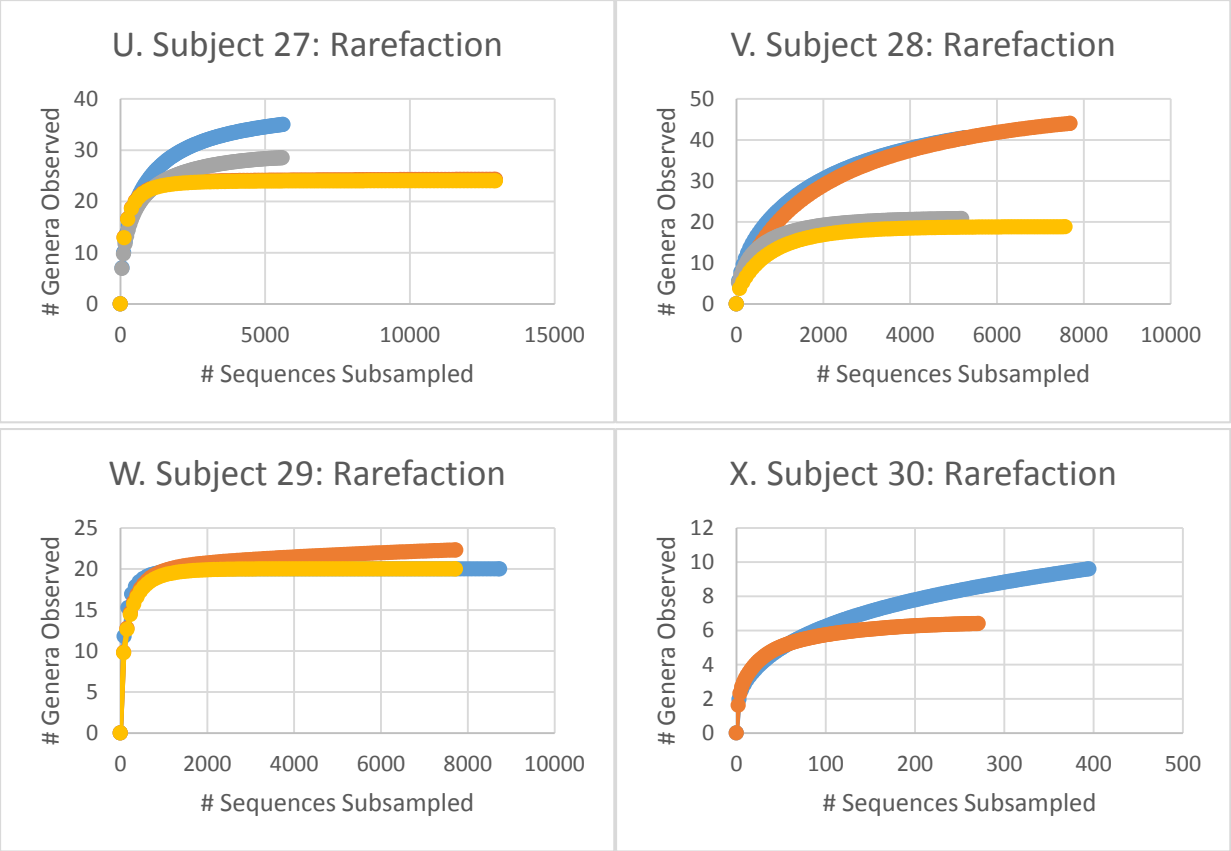
P. Subject 22: Rarefaction
Genera Observed
15
10
5
0
0
5000
10000
15000
20000
Sequences Subsampled


Q. Subject 23: Rarefaction
Genera Observed
80
60
40
20
0
0
5000
10000
15000
Sequences Subsampled


R. Subject 24: Rarefaction
Genera Observed
40
30
20
10
0
0
1000
2000
3000
4000
Sequences Subsampled


S. Subject 25: Rarefaction
Genera Observed
50
40
30
20
10
0
0
5000
10000
15000
Sequences Subsampled


T. Subject 26: Rarefaction
Genera Observed
40
30
20
10
0
0
2000
4000
6000
Sequences Subsampled



Where Ghannoum et al. eliminated taxa appearing at less than 1% of sample abundance, we predicted that by removing even fewer taxa (< 0.1% relative abundance), we could reach asymptote for our samples. At lower sequencing efforts, as in Ghannoum and colleagues' study (average 1,702), 1% relative abundance is only ~17 sequences, where eliminating taxa at 1% in our study meant removing a mycobial component represented by as many as 300+ sequences. Arguably, discounting such taxa from larger sequencing efforts ignores legitimate contributions to the oral mycobiome. We hypothesized that excluding less taxa would still eliminate artifactual singletons and other insignificant taxa, retain rare, but legitimate taxa, but also level off rarefaction curves, indicating the majority of the diversity had been captured. Singletons and other low abundance genera have the biggest impact on reaching the asymptote, and these would be removed with a 0.1% relative abundance cutoff. To test our hypothesis, an additional ten iterations were performed and averaged for each sample after the removal of taxa at < 0.1%

relative abundance, which we used as the criterion to define “abundant genera” (Figure 5.1 panels A-X: gray and yellow). For four of the samples there was no change in number of taxa represented (indicated by asterisks in Table 5.1). Two of these four (Figure 5.2 panel M orange, panel W blue) reached asymptote despite this, and so did not require such modifications. The other two came from subject 30, both of which yielded low sequence counts of 395 and 271 and did not reach asymptote, casting doubt on this low sequencing effort. Another suspicion about trusting sequence counts as low as these was raised during preliminary analysis; these two time points had the greatest beta diversity between them of any given subject. For sequencing efforts as low as this, even a singleton would be greater than 0.1% relative abundance and could also possibly be a legitimate taxon. It is clear by the rarefaction curve that the majority of the diversity for subject 30 time point 1 was not captured (Figure 5.2 panel X blue). Both time points for this subject would be greatly affected by additional sequencing effort; only 132 sequences are needed for one additional taxon time point 1 and 903 sequences for time point 4. Further evidence that sequencing effort in the hundreds is inadequate for analysis is seen in Figure 5.3, where sampling efforts were pooled after inadequate sampling effort.

Nearly all of the remaining 43 samples reached a clear asymptote when taxa < 0.1% were removed, and certainly the majority of diversity had been captured (Table 5.1). Most notable was how drastically rarefaction curves were impacted after this modification, with an average of 69.2% of taxa retained, while sequence counts remained relatively unchanged at an average of 99.6% retained. Therefore we addressed our hypothesis and showed that by removing fewer taxa (< 0.1% relative abundance) than comparable studies[23], we reached asymptote, eliminated insignificant taxa, and retained rare, but legitimate taxa, while capturing the majority of diversity in a sample.

The slight modification of eliminating taxa below 0.1% relative abundance allowed us to determine a minimum recommended sequence threshold. Visual inspection of the curves led us

Sample Time point 1	Effort for 90.0% taxa	Effort for 95.0% taxa	% Taxa at 3k Seqs	Effort for 1 Taxon Missing	# Seqs	# Taxa	Effort needed for ≥1 Taxon	Sample Time point 4	Effort for 90.0% taxa	Effort for 95.0% taxa	% Taxa at 3k Seqs	Effort for 1 Taxon Missing	# Seqs	# Taxa	Effort needed for ≥1 Taxon
5	1092	1638	98.8	3120	7915	46	78690.2	5	793	1281	99.2	1647	6217	15	1.2E+07
6	2040	2856	95.8	4352	13734	35	7.1E+09	6	1740	2610	96.3	4350	8814	32	28842.7
7	952	1547	98.6	3451	11967	54	3.3E+09	7	1411	2241	97	3320	8360	15	41372.9
8	186	372	100	403	3108	10	31050.0	8	1596	2436	96.8	5124	8510	76	16782.5
10	635	1270	98.6	1397	12928	11	1.3E+10	10	2570	3875	91.5	2625	12662	4	2.3E+09
11	1908	2703	96.1	1749	16005	4	undefined	11	1	1	100	1	8789	1	undefined
12	705	1269	98.8	1551	14200	14	1.2E+12	12	246	492	99.7	943	4152	17	41360.3
13	704	1024	99.2	1856	6443	28	31917.2	13	324	648	99.5	1296	8204	26	2.5E+10
14	1989	3159	94.8	3042	11762	9	undefined	14	2440	3660	92.9	5185	30637	26	undefined
16	1716	2508	96.7	5016	6695	37	16619.5	16	2312	3128	94.4	5032	13733	41	3.9E+07
17	1660	2573	96.5	3237	8327	17	82554.4	17	605	968	99.1	1452	12126	16	7.4E+10
18	1048	1572	99	524	13216	2	undefined	18	1556	2334	97.8	1556	39201	5	undefined
19	235	376	99.3	2538	4763	19	15835.5	19*	1017	1582	99	2034	11385	18	2.2E+15
20	1526	2180	97.6	2616	11051	15	8.1E+08	20	2569	3670	93.1	2569	36987	5	undefined
21	1426	2070	98.2	2852	4667	44	5059.5	21	1932	2772	96.1	4032	8519	31	1.7E+06
22	1015	1885	97.7	1740	14592	9	1.1E+12	22	N/A	N/A	N/A	N/A	N/A	N/A	N/A
23	1962	2834	95.9	4687	11073	45	1.1E+07	23	924	1344	99.4	2184	8447	33	2.2E+07
24	408	646	N/A	1615	1776	31	1610.7	24	270	390	100	600	3030	28	8.0E+08
25	2231	3298	94.2	4850	9776	26	48467.5	25	2640	3600	92.4	4720	8115	23	39277.2
26	1196	1840	98.4	3312	4677	35	5154.5	26	1488	2160	97.8	3744	4906	30	5359.3
27	2200	3245	94.2	3630	5617	28	3675.8	27	903	1290	99.1	1935	12959	23	8.0E+11
28	1734	2499	96.8	3774	5266	20	8501.6	28	2100	2850	95.7	4050	7686	18	73599.5
29*	435	609	100	783	8734	19	undefined	29	616	924	99.9	1155	7724	19	undefined
30*	282	336	N/A	381	395	12	131.6	30*	104	152	N/A	76	271	7	903.3

Table 5.1: Dataset statistics of arm 1-healthy samples post removal of < 0.1% taxa, Asterisks indicate no taxa were below 0.1% relative abundance, orange indicates reasonable additional sequencing effort is needed for ≥ 1 taxon (< 6,000 sequences), green represents < 95% taxa were subsampled

to propose a conservative minimum of 3,000 sequences to achieve a rarefaction plateau for all samples. Using sample iteration averages, a lower limit of 3,000 sequences yielded an average of 97.3% taxa subsampled, with only 8 time points subsampling at < 95% of taxa (Table 5.1 green). Analysis of the 8 poorly performing samples follows. Three of these, samples 10-4, 14-1, and 20-4, all had taxonomic counts between 4 and 9, thus their subsampling percentages above 90% are essentially 100%. The other 5 reached 95% of subsampled taxa by 3,660 counts, indicating that a sequence count of 3,700 sequences would be adequate to meet at least this percentage of total taxa.

On its own, sequencing effort at 95% sub-sampling is a misleading metric for some samples because it only considers taxa that were obtained during experimental sequencing. For example, based on results from subject 30, already shown to be unreliable due to its inability to reach asymptote (Figure 5.2, Panel X), only 336 sequences are required to reach 95% of taxa. However, confidence cannot be had in this value alone, especially when the rarefaction curve has not plateaued and the effort needed to acquire an additional legitimate taxon is not unreasonable (>1,000 sequences in the case of sample 30).

In addition to visual inspection, delta values (Δ sequencing effort/Δ sub-sampled taxa) were calculated between the last two sequencing efforts represented in the rarefaction curves. While an exact number cannot be inferred, this metric represents an estimate of the additional number of sequences required to add one taxon to the sample. This value is a minimum estimate, as it is expected that the slope of the line will continue to approach zero and eventually reach asymptote past the point of our maximum sequencing effort. The median delta value obtained for all time points was 1.2E+07 sequences, which is an unreasonable amount of experimental sequencing required for a negligible gain of only one taxon for the time and labor required. Nine of the samples levelled off completely (undefined, or infinite, value) and seven of the samples showed a reasonable effort of sequencing could be made for additional taxa (< 6,000

sequences). Aside from the two time points from sample 30, time point 24-1 indicated a need of only 1,611 sequences to acquire an additional taxon, suggesting that a sequencing effort as low as 1,776 is not adequate to represent the majority of the diversity in this oral sample. These data suggest that sequencing effort should be greater than 2,000 counts post-pipeline, a less conservative estimate than the 3,700 we found to necessary to capture the majority if diversity. While the remaining four samples only require 3,675- 5,359 sequences to gain an additional taxon, the percentage of subsampled taxa would only decrease to 92.6% at a minimum, with diminishing returns produced for sequencing efforts beyond this point. Application of a minimum sequencing effort of 2,000 affected three time points in our healthy dataset. Therefore, time points 24-1, 30-1, and 30-4 were excluded from further analysis, leaving 44 time points representing 23 subjects. Two of the arm 1-healthy time points fell between 2,000 and 3,700 sequences, but analysis by their delta and 95% sub-sampling values show that the majority of diversity had been captured, so they were retained in analysis.

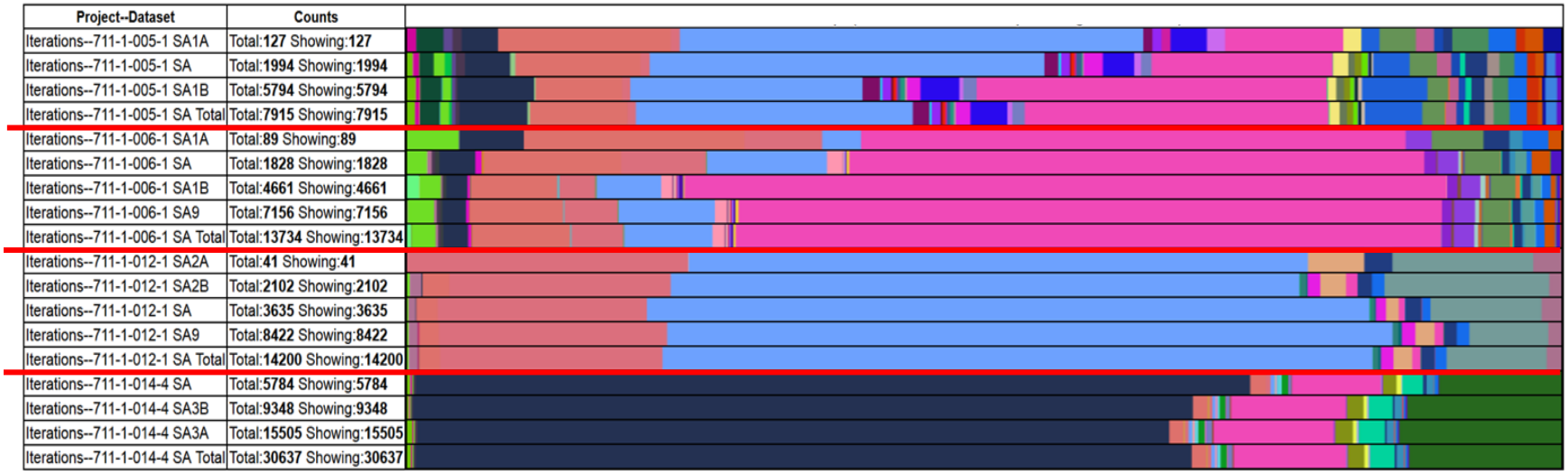
Based on these results, we recommend that a minimum of 3,700 sequences are collected for all saliva samples to obtain at least 95% of taxa. In future studies, samples with counts below 2,000 should be excluded and for those falling within 2,000-3,700, caution should be taken by creating supporting material in the form of rarefaction curves, delta values, and subsampling requirements to reach the majority of taxa at ≥ 0.1% relative abundance.

5.2.2 Empirical measurement of richness as a function of sequencing depth

Due to the imperfect nature of experimental sequencing, our first attempts for particular samples yielded low sequence counts and required additional sequencing. Ultimately, all replicate sampling efforts were combined for analysis. Failure in obtaining a large number of sequences on the first try provided a unique opportunity for empirical evaluation of our required estimate of 3,700 sequences for capturing diversity. Figure 5.3 shows four samples chosen to compare

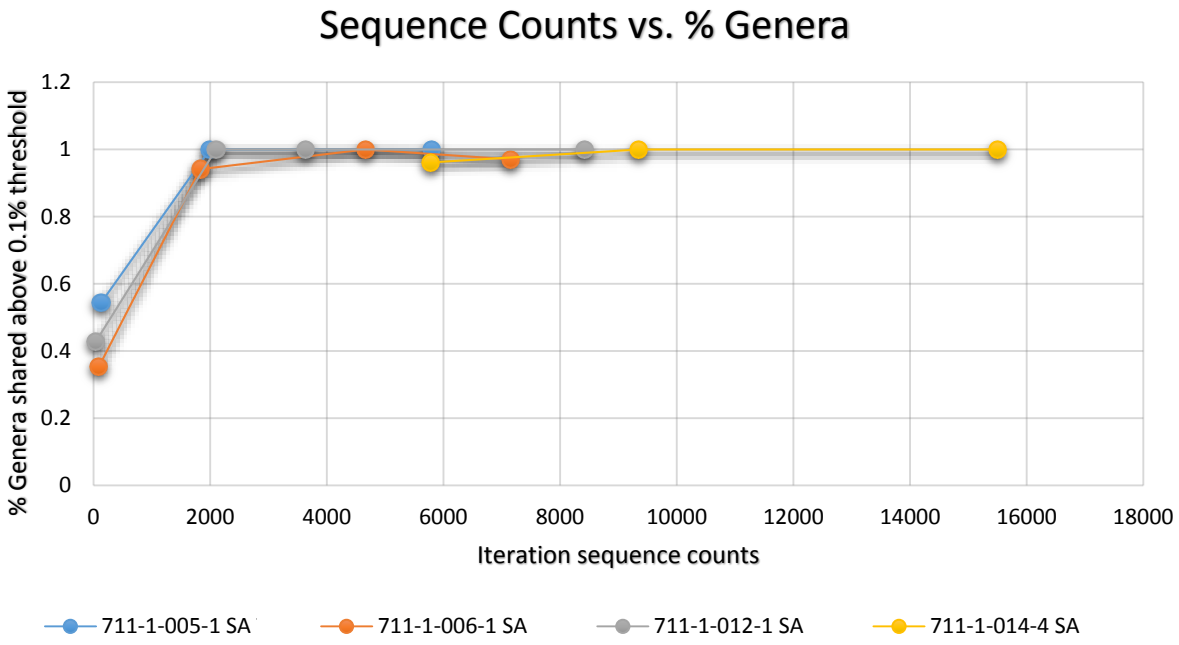
Figure 5.3: Experimental richness comparisons of sequencing iterations of four arm 1-healthy samples (as described in Figure 5.1) at multiple sequencing depths using abundant genera (genera ≥ 0.1% relative abundance). Panel A: Bar graph showing effects on genera representation as sequencing effort increases and on combining all sequences from multiple efforts (Iterations Totals, rows 4, 9, 14, 18). Red lines separate each sample. Highly represented genera are pink (*Malassezia),* blue (*Pichia*), orange (*Cladosporium*), salmon (*Cryptococcus*), and navy blue (*Candida*). Panel B: Scatterplot of empirical rarefaction curve showing percentage of total abundant genera from combined efforts (y-axis) found at each sequencing depth (x-axis) for the four samples.

A



Project--Dataset	Counts
Iterations--711-1-005-1 SA1A	Total:**127** Showing:**127**
Iterations--711-1-005-1 SA	Total:**1994** Showing:**1994**
Iterations--711-1-005-1 SA1B	Total:**5794** Showing:**5794**
Iterations--711-1-005-1 SA Total	Total:**7915** Showing:**7915**
Iterations--711-1-006-1 SA1A	Total:**89** Showing:**89**
Iterations--711-1-006-1 SA	Total:**1828** Showing:**1828**
Iterations--711-1-006-1 SA1B	Total:**4661** Showing:**4661**
Iterations--711-1-006-1 SA9	Total:**7156** Showing:**7156**
Iterations--711-1-006-1 SA Total	Total:**13734** Showing:**13734**
Iterations--711-1-012-1 SA2A	Total:**41** Showing:**41**
Iterations--711-1-012-1 SA2B	Total:**2102** Showing:**2102**
Iterations--711-1-012-1 SA	Total:**3635** Showing:**3635**
Iterations--711-1-012-1 SA9	Total:**8422** Showing:**8422**
Iterations--711-1-012-1 SA Total	Total:**14200** Showing:**14200**
Iterations--711-1-014-4 SA	Total:**5784** Showing:**5784**
Iterations--711-1-014-4 SA3B	Total:**9348** Showing:**9348**
Iterations--711-1-014-4 SA3A	Total:**15505** Showing:**15505**
Iterations--711-1-014-4 SA Total	Total:**30637** Showing:**30637**

B



experimental richness between sequencing iterations. We found that only 35-55% of abundant genera from pooled efforts were represented in sequencing attempts comprised of 41-127 sequences. In sequencing efforts this low, missing genera included core members such as *Aspergillus*, *Aureobasidium*, *Cladosporium*, *Epicoccum*, *Fusarium*, *Malassezia*, *Penicillium*, and *Saccharomyces*. In contrast, iterations with sequences near 2,000 or greater shared at least 94% with pooled abundant genera. In all four instances, increasing sequence counts above 2,000 did not significantly add abundant genera, and pooling unequal attempts had little to no impact on their overall quantitative representation. Therefore, 3,700 is a sound estimate for required sequencing effort, and as little as 2,000 sequences are needed for capturing the majority of diversity. These results also suggest that the mycoprofile of a sample can be determined with sequences counts in the tens to hundreds, but the ability to capture representation of opportunistic pathogens and even core members in lower abundance is lost.

5.3 Healthy subjects partition into distinct mycoprofiles

We asked the question of whether healthy subjects could be grouped by similarities in their mycobial content (mycoprofiles), determined by frequent prevalence of specific genera. The Morisita Horn metric was chosen for comparing arm 1-healthy samples' beta diversity (differences between samples), as it is not strongly affected by differences in sequencing effort across time points. Samples were clustered in a dendrogram using VAMPS's R Vegan tree and pie charts were placed to visually explain clusters by taxonomic representation (Figure 5.4). Designations to mycoprofiles were made based on relative abundances of the most prevalent genus; the requirements for placement into a mycoprofile groups were 1) prevalent genus had to be >40%, and 2) no other genus could be >25%. Two main groups were visualized in the dendrogram, with the largest of the groups predominated by Malassezia in amounts from 42-95%. The second group is predominated by Candida species (excluding Pichia) at 60-100%. There were no other genera that met qualifications to constitute a mycoprofile, so the rest of the samples were deemed the "Diverse mycoprofile". This third group often contains *Malassezia*, but little to no *Candida*.

Next, we asked whether these mycoprofiles were stable. Of the 21 subjects which had an additional time point, 13 of them were in the same group or "mycoprofile", after a period of 2 weeks. Only 3 of the 9 subjects with *Malassezia* mycoprofiles stayed in this type, with 6 meeting qualifications for the Diverse mycoprofile after 2 weeks and 5 of these retaining *Malassezia* as a prevalent component. In contrast, only 2 of the 8 patients with *Candida* mycoprofiles for at least one time point met qualifications for the diverse mycoprofile group within two weeks. This suggests that *Malassezia* mycoprofiles are more fluid in their diversity over time, while *Candida* mycoprofiles are more stable. Furthermore, a plot of beta diversities in a heat map show two blue groups representing the *Malassezia* and *Candida* mycoprofiles, as well as a red group of sequences that do not resemble each other or any other profiles, representing the Diverse

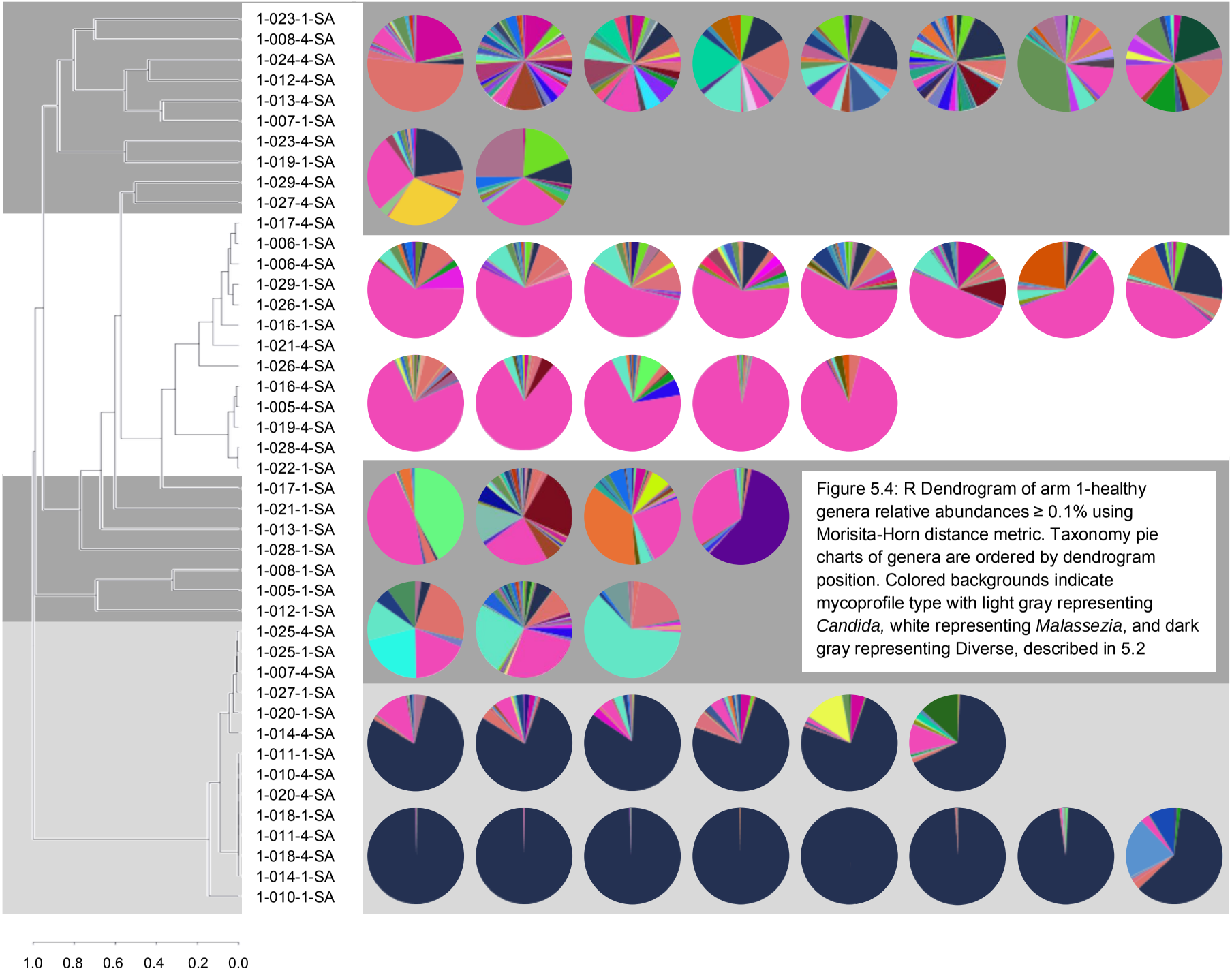
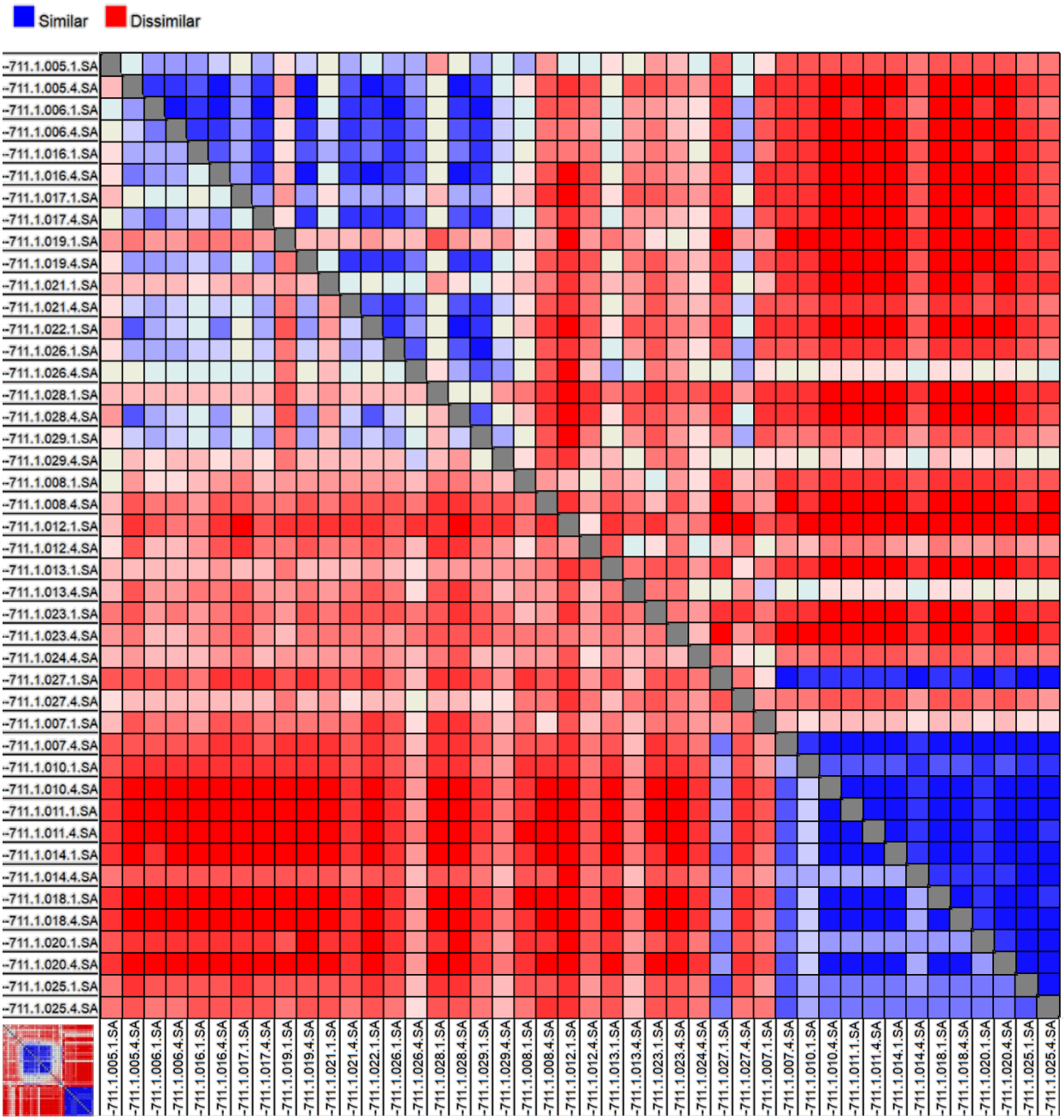


Figure 5.4: R Dendrogram of arm 1-healthy genera relative abundances ≥ 0.1% using Morisita-Horn distance metric. Taxonomy pie charts of genera are ordered by dendrogram position. Colored backgrounds indicate mycoprofile type with light gray representing *Candida,* white representing *Malassezia*, and dark gray representing Diverse, described in 5.2

Figure 5.5: Heat map of arm 1-healthy time points showing mycoprofile group remains relatively consistent after a two week period using Morisita-Horn distances (upper right, above gray boxes) and Bray-Curtis distances (bottom left triangle, below gray boxes). Samples were ordered by subject and mycoprofile, where mycoprofile for a subject was determined by at least one time point belonging to a *Malassezia* or *Candida* mycoprofile, described in 5.2. Subjects were grouped: *Malassezia*- 5, 6, 16, 17, 19, 21, 22, 26, 28, and 29; Diverse- 8, 12, 13, 23, and 24; and *Candida*- 27, 7, 10, 11, 14, 18, 20, and 25. Bottom left inset shows Morisita-Horn distances for samples ordered according to figure 5.4 dendrogram.



mycoprofile (Figure 5.5 inset). This confirms the groupings suggested by the dendrogram are driven primarily by *Malassezia* and *Candida* content. Ordering the heat map by subject number (including time points 1 and 4) shows similar groupings to the inset and suggests that mycoprofiles typically constitute the same genera over time.

5.4 Conclusions

In this chapter we explored the healthy fungal mycobiome of 23 patients in order to determine how healthy individuals vary in their oral mycobial profile inter- and intra-subject. We supported our hypothesis that removing taxa at less than 0.1% sequence abundance would eliminate artifactual taxa and retain rare but legitimate taxa. Using the rarefaction curves and empirical evaluation of iterations we determined the minimum sequencing depth needed for capturing 95% of genera richness. We found that approximately 2,000 would suffice for most samples, but 3,700 sequences were required to capture richness for all samples. We recommend collecting 3,700 ITS1 sequences for oral mycobial samples for future studies. To our knowledge, this is the first study to determine such an estimate for metagenomic analysis of fungi. We also discovered that there are typically 3 mycoprofiles that represent healthy subjects and ~60% of individuals stay within their original mycoprofile after a period of two weeks. Of the *Candida* and *Malassezia* mycoprofiles, we found the *Malassezia* mycoprofile to be the least stable. This implies that *Malassezia* types may be more susceptible to change due to transient microorganisms. It also implies that individuals with *Candida* mycoprofiles will remain relatively stable. With *Candida* sp. known to be opportunistic pathogens, this suggests that those harboring *Candida* in a healthy state may be more vulnerable during states of decreased immunity than individuals with Diverse or *Malassezia* mycoprofiles.

Chapter 6. Fungi in oral mucositis

6.1 Introduction

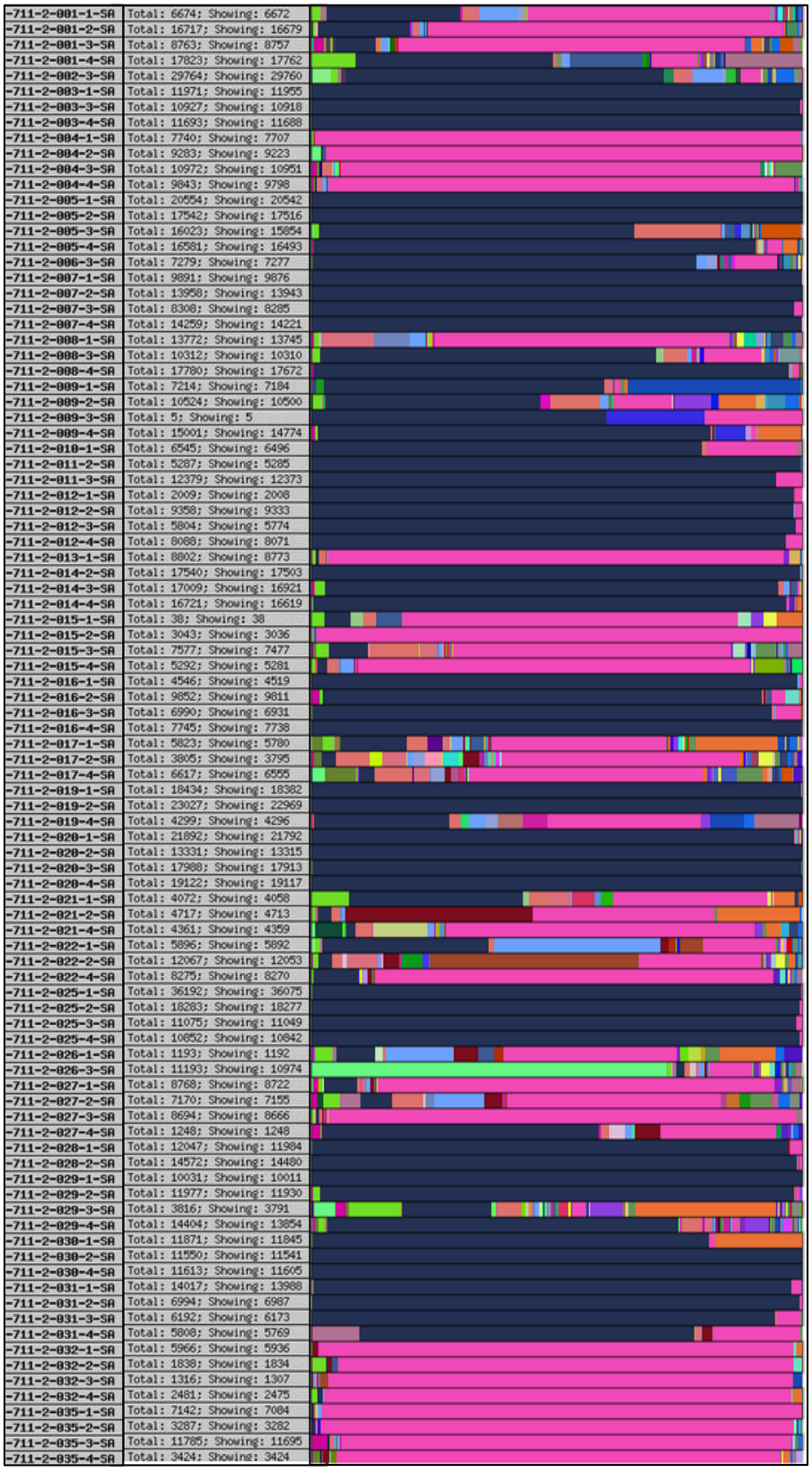
After our development of a robust protocol and new knowledge that healthy subject mycoprofiles remain relatively constant over time, we were able to explore fungi as they relate to disease. For this study we focused on oral mucositis, a side effect of chemotherapy in which fungi have previously been implicated. We predicted that mycoprofiles appearing mostly as *Candida* at day 0 would be considered the least healthy and would be those most susceptible to oral mucositis during chemotherapy, as *Candida* spp. are the most common cause of infections in cancer patients[61]. While the *Candida* mycoprofile has only been identified in approximately one-third of all healthy subjects studied in our lab for this project and others, a community predominated by *C. albicans* species from the start, has natural potential for opportunistic infections. We predicted that those with the majority of taxa represented by *Malassezia* and other core members would be the least likely to develop oral mucositis and candidiasis, as *Malassezia* (and some other genera) potentially compete for space and resources with *Candida* to maintain a healthier equilibrium. The goals of this chapter are twofold: First, to evaluate trends in community membership as oral mucositis develops and progresses for each subject using standard WHO (World Health Organization) scores. Second, to explore general changes within individuals between communities at day 0, day 2, day 9, and day 14 time points.

6.2 Mycoprofiles of cancer patients

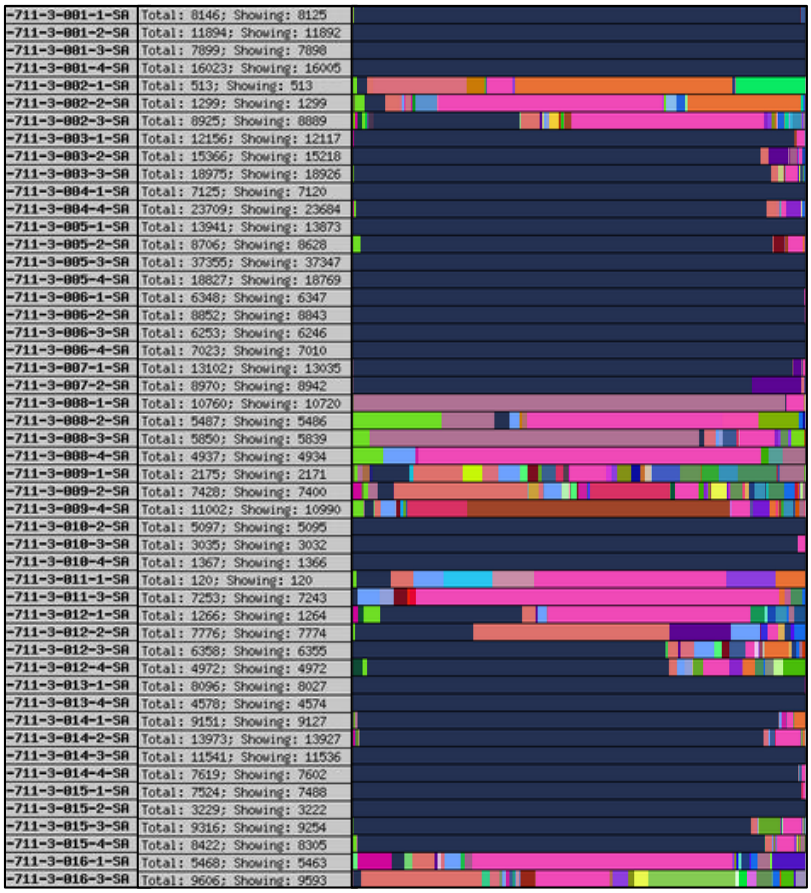
Forty-nine cancer patients participated and 40 completed the study with all four time points and enough remaining saliva for fungal ITS1 sequencing. Figure 6.1 shows genera distributions for all sequenced time points of 46 patients, representing 1,411,136 sequences. All three of the mycoprofiles previously recognized in healthy subjects, *Candida, Malassezia,* and Diverse, are also in evidence in the cancer patients. Only 21 of these subjects produced sequences at all four time points due to the unpredictable nature of clinical samples; it is possible that many

Figure 6.1: Bar graphs of all sequences obtained for cancer chemotherapy patients arms 2-naïve chemo (panel A) and 3-non naïve chemo (panel B). Mycoprofiles are visualized by distributions of colored bars. *Malassezia* mycoprofiles are predominantly pink and *Candida* mycoprofiles are predominantly navy blue.

A



B



samples did not contain enough fungi to provide adequate template for PCR and sequencing. Moreover, our methods likely exacerbated this by additional loss of PCR product when removing primer dimer artifacts. Seventeen of these 21 patients met the minimum number of required sequences (explained in section 4.2) for all time points.

Of these 17 subjects, 11 came from the arm 2-naïve chemo cohort and 6 came from the arm 3-non-naïve chemo cohort (recurring cancer patients). The average sequencing effort per time point for these 17 subjects was 11,118 (range 2,009-37,355), coverages that were more than adequate, as discussed in the preceding chapter. Eleven additional patients outside of the complete 21 were successfully sequenced with adequate depth for at least two time points and

could provide useful information for future analyses about specific time points. Over 75% of subjects stayed within their original genus mycoprofile during the entire course of treatment, even more than the 60% for healthy subjects (Chapter 5).

Because we had determined that mycoprofiles were relatively stable in healthy subjects, we understood the importance of determining the mycoprofile composition at the day 0 time point. We asked the question of whether mycoprofile type and frequency in the cancer cohort would differ from the three mycoprofiles identified in healthy subjects. Both the *Malassezia* mycoprofile and *Candida* mycoprofile were present, but at altered frequencies, and a new profile emerged as well, containing mostly *Aureobasidium* (Figure 6.2). Not a single one of the patients started with the Diverse mycoprofile, though one (subject 2-001), had a more diverse *Malassezia* mycoprofile than the other subjects with the *Malassezia* type. The absence of a Diverse mycoprofile in the cancer patients is interesting because 57% of the healthy cohort showed Diverse mycoprofiles for at least one of the two time points. Greater than 75% of the cancer subjects had *Candida* mycoprofiles at Day 0, while only 38% of healthy subjects showed a *Candida* mycoprofile. A different study on 33 additional healthy subjects performed in our lab also places the number of *Candida* mycoprofiles at a yet lower frequency (24% with > 50% *Candida albicans*, and 3% with greater than 75%, Linda Strausbaugh, personal communication) than observed for our chemotherapy cohort. One explanation we propose is that differing frequencies could be due to the difference in average in age between the two groups (48.1 years for the healthy subjects sequenced vs 60 years for the 17 cancer subjects in question). This supports published evidence that oral carriage of *Candida* spp. is more apparent in individuals who are in advanced age groups (74.2% for ages 71-92, and 35.0% for ages 56-70)[62]. However, there was only a slight difference in the average age for chemotherapy subjects with *Malassezia* mycoprofiles at day 0 (58 years) and the average for *Candida* mycoprofile

subjects (59.6 years). We interpret this to mean that there is no age-related influence on the ability of an individual to carry oral *Malassezia.*

One contributing factor to the worldwide rise in groups of individuals with lowered immune function is the increase in life expectancy for the human population. Urine proteomes revealed 19 proteins that are differentially expressed in younger age groups (19-26 and 45-54 years) compared to older ages (72-90 years), many of which were involved in tissue remodeling and increased immune dysregulation[63]. As the most common oral pathogens are *Candida* spp., and fungal infections predominate in individuals with lowered immune function, increased frequency

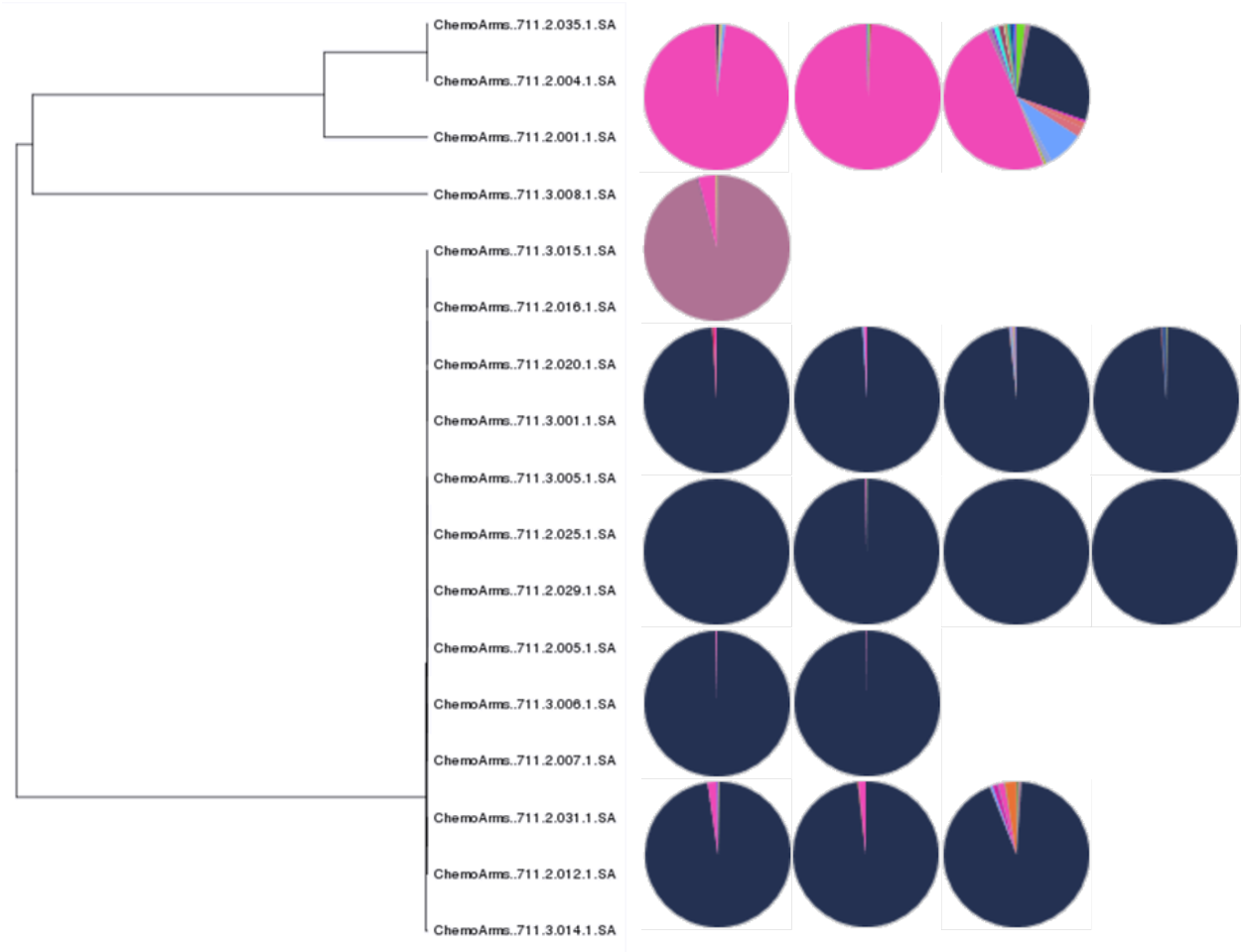


Figure 6.2: Dendrogram and corresponding pie charts for abundant genera (≥ 0.1% relative abundance) observed at day 0 time points for arms 3- naïve chemo and arm 3-non-naïve chemo patients for which the complete longitudinal dataset was available. Mycoprofiles are visualized by *Malassezia* (pink), *Aureobasidium* (mauve), and *Candida* (navy blue).

of *Candida* mycoprofiles in older groups in our study may also be due to an age–related decline in immunity.

Among this cohort, age is not the only explanation for decline immunity. We propose that a differential representation of *Candida* mycoprofiles for chemo arms may also be due to immune-compromised immune systems caused by the cancer itself. Another study reports that in mice, during early stages of tumor growth, a rise in H_2O_2 and TNF-α levels were detected, but after progression of the tumor, the effectiveness of macrophage activity decreased against *C. albicans*[64]. Between arm 2-naïve chemo and 3-non-naïve chemo cohorts, none of the arm 3 participants displayed *Malassezia* mycoprofiles. It is possible that merely the state of having cancer, or having prior chemotherapy, contributes to a perpetual state of susceptibility to opportunistic infections. The sole *Aureobasidium* mycoprofile came from an arm 3 subject, and was the first we encountered of its kind.

6.3 *Candida* mycoprofile

To make associations between changes in community membership during disease states, it is often beneficial to group like types of patients; in small studies such as this one, this strategy might better reveal co-occurrence of mycobiome features and produce similar disease phenotypes. Thus, we grouped the 13 subjects according to mycoprofile. The thirteen subjects that had *Candida* mycoprofiles at day 0 were compared to the rest of the time points within that subject, and also across subjects using the Morisita-Horn distance metric as a beta-diversity estimator (Figure 6.3). In general, there were no differences at the genus level either within a *Candida* mycoprofile subject over the course of the four time points or between cancer patients at the majority of time points. Three time points from the patient cohort had mycoprofiles that were atypical in that they displayed high contributions (>5%) of genera other than *Candida* (2-029-3, 2-005-3, and 2-031-4, with the first being the most different of any time point). Time point 2-029-3 was characterized by the largest beta diversity distance from any other sample with

high levels of *Saccharomyces* (28.5%), *Aspergillus* (10.4%), and *Penicillium* (6.7%). Time point 2-005-3 showed the next greatest beta diversity distances and was characterized by *Cladosporium* (17.4%) and *Trichosporon* (7.9%).Time point 2-031-4 exhibited high levels of *Malassezia* (17.5%) and *Aureobasidium* (9.3%). These non-*Candida* genera are common mycobial components found in healthy subjects[27]. All other 49 time points had no genera other than *Candida* with a contribution of more than 5% abundance. This suggests that genera components in cancer patients with *Candida* mycoprofiles can be expected to remain relatively unchanged from their baseline community membership.

WHO scores were used to measure the severity of oral mucositis, which range from 0-4, where 0 is no change in oral health, 1 is soreness or erythema (reddening) of mucosa, 2 is erythema and ulcers with the ability to eat solid foods, 3 is ulcers and the necessity of a liquid diet, and 4 is necessity of a feeding tube as alimentation is not possible[65]. Nine of the thirteen patients experienced OM at a WHO score of 2 at any given time point (2-005, 2-007, 2-020, 2-025, 2-029, 2-031, 3-006, 3-014, 3-015). Two subjects acquired WHO grade 1 OM (3-001 and 3-005), and two subjects were unaffected by oral mucositis (2-012 and 2-016). At the genus level there were no obvious mycoprofile differences separating these subjects.

While the *Candida* genus level mycoprofiles did not demonstrate informative changes during chemotherapy, they did display an important day 0 association to candidiasis. Of the 7 patients who acquired candidiasis, all of them had original day 0 communities comprised mostly of *Candida*. Due to non-broad spectrum effects of antifungals, species level analysis is particularly important for the biomedical community and for evaluating trends in subject populations. Day 0 relative abundances at the species level, is primarily comprised of *albicans* (medium blue 14.3-99.9%) and *dubliniensis* (gray 0- 79.6%). Four groups emerged out of these 7 subjects (Figure 6.4). The first group (subjects 2-005 and 2-029) started out almost entirely as *C. albicans*, which became the source of infection at time point 2 in both subjects. By time point 3, the infection had

cleared and *albicans* was replaced with many other non-*Candida* and non-*Malassezia* species. At time point 4, *C. albicans* began to increase again, but there was no recurrence of infection. The second candidiasis group that emerged (subject 2-020 and 3-001) showed a species majority of *C. dublienensis* (79.6- 57.9%) with the rest occupied by *C. albicans,* at time point 0. However, by time point 2, *C. albicans* counts increased by an average of 10%, and an additional average of 40% by time point 3. For both patients in this group, infection persisted from time point 2 through time point 4, and *C. albicans* remained above Day 0 abundance thresholds. The third candidiasis type was made up by a single subject (2-007) and similar to the second group, sequence counts primarily consisted of *C. dubliniensis.* However, this group contained a small percentage (8.2%) of *C. glabrata*, which grew to 46.8% by time point 3, potentially being the cause of candidiasis for this subject rather than the typical *C. albicans. C. glabrata* has been shown to colonize approximately 5% of cancer patients in a review of three studies[15]. The fourth group (subjects 3-005 and 2-011) appeared to have strong *C. albicans* representation throughout all time points where data was obtained, but only exhibited candidiasis at time point 3. In all cases we note that the pathogen implicated in infection was present at day 0 and that it remained detectable after infection had resolved. As this is a small subset of individuals, a larger study is required to determine if these groupings represent all possible types.

Antifungal drugs were administered to patients in this study who acquired infections. Six of the seven subjects presented here were prescribed Nystatin, while only subject 2-007 was prescribed fluconazole. It is presumed that the physicians did not test for *Candida* species before prescribing these drugs because fluconazole has lower effectiveness against *C. glabrata*, which is the main source of ITS1 sequences for subject 2-007.

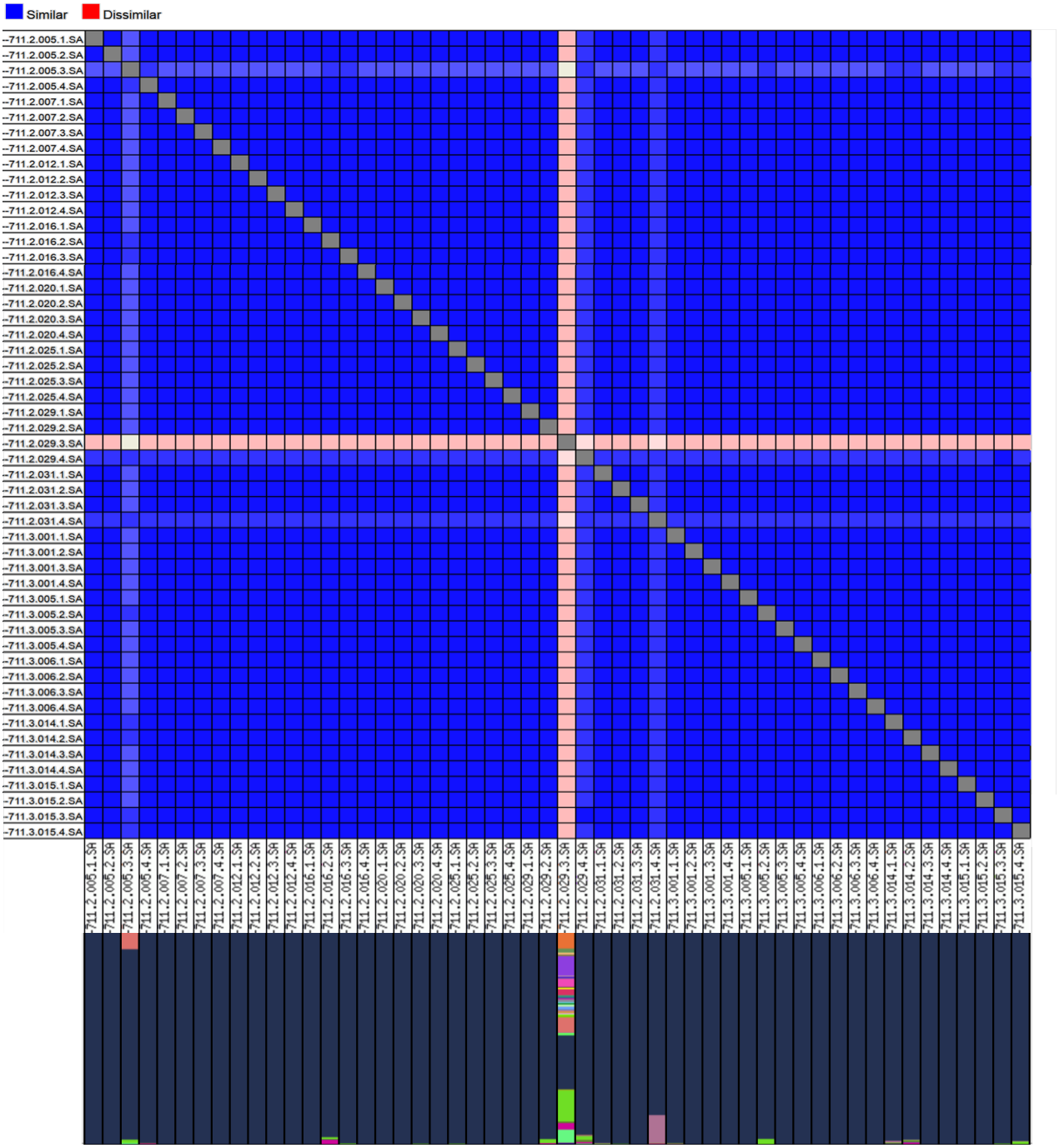


Figure 6.3: Heat map and genera distribution through all time points of patients with day 0 *Candida* mycoprofiles from arm 2-naïve chemo and 3-non-naïve chemo cohorts. Morisita-Horn metrics were used to generate the heat map. Abundant genera (≥ 0.1% relative abundance) were used to generate bar graphs, with navy blue representing *Candida*.

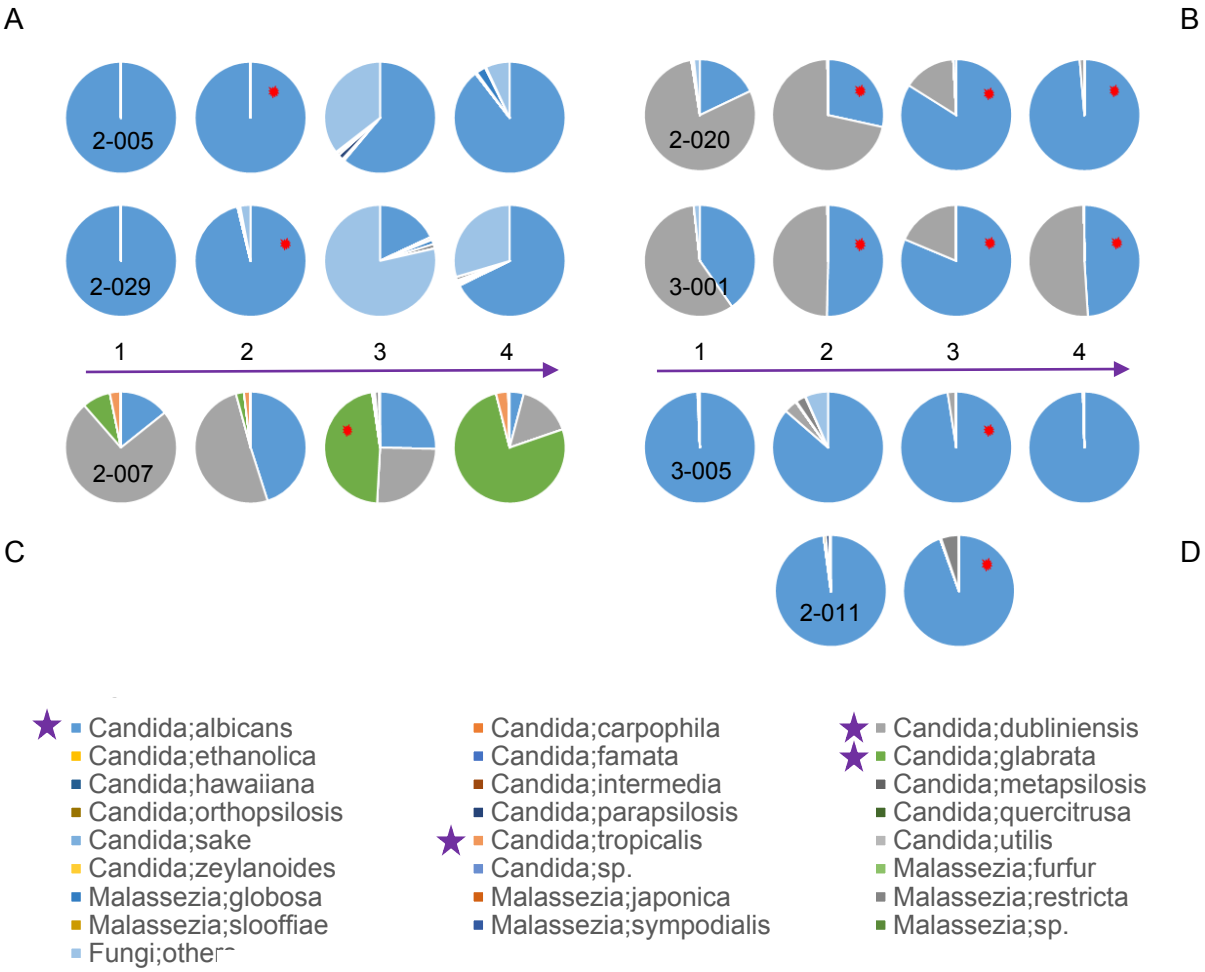


Figure 6.4: Species distributions of patients (time points 1-4) who develop candidiasis during the course of chemotherapy fall into four categories: A- *Candida albicans* predominates at baseline and during infection at time point 2, with reduction of *C.* albicans at time point 3 and an increase at time point 4. B- *Candida dubliniensis* predominates at baseline and then decreases as predominance of *C. albicans* increases with infection at time points 2-4. C- *Candida dubliniensis* predominates at baseline and *C. glabrata* prevalence dominates by time point 3 with infection, as well as into time point 4. D. *C. albicans* prevalence dominates throughout time points 1-4, with infection only occurring at time point 3. Red asterisks indicate time points where patients show candidiasis. Purple stars in key indicate predominant *Candida* species from pie charts.

6.4 *Malassezia* mycoprofile

Of the 3 subjects with day 0 *Malassezia* mycoprofiles, all stayed within their relative type over the course of chemotherapy, with the exception of subject 2-001, who originally showed a more

diverse *Malassezia* mycoprofile than the other two subjects, and included 27% *Candida* as well as 49% *Malassezia.* For the duration of the study, subject 2-001 showed *Candida* proportions ranging from 8-40% and *Malassezia* proportions from 9%-70% (Figure 6.5).

This subject also was affected by the worst grade of OM of the three samples with a WHO score of 2 at time point 3. Though N is limited for this dataset, subject 2-035 kept a consistent WHO score of 0, while subject 2-004 only developed OM at a WHO grade of 1 at time point 3. One reason might be that *Malassezia* can exert an ameliorating effect such that a *Malassezia* mycoprofile with little presence of *Candida* may be less affected by OM. Another reason that this subject may have been more strongly affected by OM at time point 3 was that his oral mycobial profile is most different from its baseline state at this time point. Alternatively, as an effect of developing OM, his mycobial profile is most altered at this point. While it appears that healthy mycobial states allow a wide range of *Candida* membership (up to 100%), the relative abundances of this genus in healthy subjects do not typically change within the period of two weeks. Major deviations from the baseline profile may be indicators that the subject is experiencing OM.

None of the *Malassezia* mycoprofile patients acquired infections, which is interesting because *Malassezia* has been shown to be an opportunistic infectious agent on skin of a wide variety of mammals. *Malassezia* is a genus that can be resolved at the species level with ITS1, and we found that the majority of *Malassezia* sequences were represented by the species *restricta* and *globosa,* with some profiles containing *sympodialis,* in all arms 1-3, whether healthy or chemo. Our results were concordant at the species level with other human samples where *Malassezia* has been found scalp, stool, skin, and sputum [55,24,66,19]. The skin and the mouth are very different biocompartments with radically different surface characteristics, so *Malassezia* may not have evolved to cause infections in the oral cavity, even though this fungi is well acclimated to

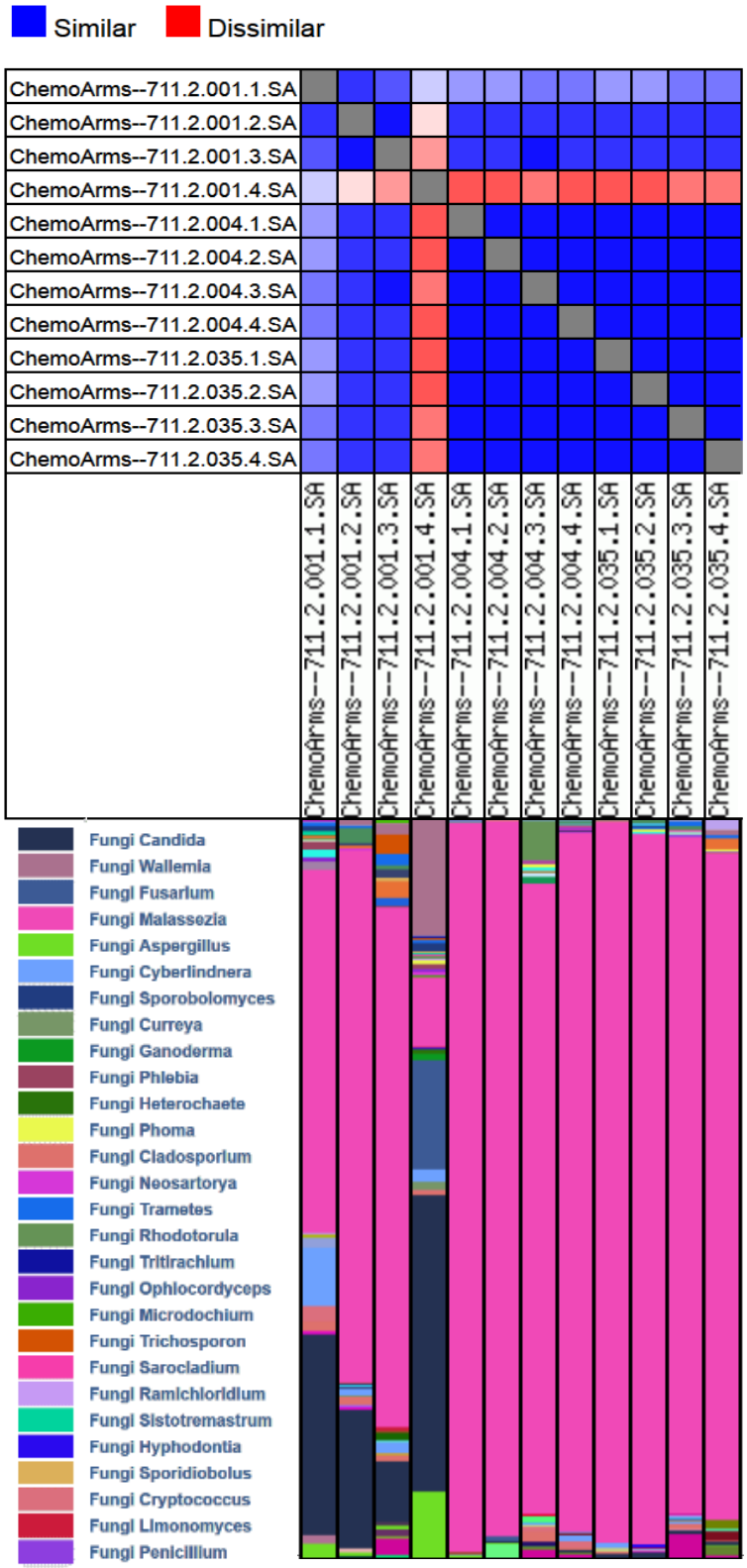


Figure 6.5: Heat map and genera distributions for arm 2-naïve chemo and 3-non-naïve chemo patients with *Malassezia* mycoprofiles

growth at human body temperatures. If this is true, *Malassezia* may act as a viable competitor for resources against species of *Candida* that do cause infections, providing a type of protection against *Candida*, and perhaps other pathogens.

6.5 Diverse mycoprofile/Aureobasidium mycoprofile

In this group of subjects, only one mycoprofile belonged to a type outside of the *Candida* and *Malassezia* mycoprofiles. However, instead of being characterized by a diverse community of fungi, the majority of membership was composed of the genus *Aureobasidium* at 95% relative abundance*. Aureobasidium* is found ubiquitously in the environment, and is only considered to be pathogenic when isolated from immunocompromised patients[67]. However, ongoing exposure to this environmental fungus may influence mycoprofiles. Testing of air samples in the patient's home in addition to saliva from their cohabitors could provide support for *Aureobasidium* origin from environmental sources unrelated to cancer, such as household dust. Figure 6.6 shows a heat map of the four time points collected for the single patient with an *Aureobasidium* mycoprofile and the distribution of fungal species at each time point. This subject was affected by OM at a WHO scale of 1 at time point 2 and a WHO scale of 2 at time points 3 and 4. *Aureobasidium* levels decreased at time point 2, making way for a larger presence of *Malassezia* and the appearance of *Aspergillus,* the latter of which is a common cause of fungal infections*.*

6.6 Conclusions

Here we investigated the effects of cancer chemotherapy on the oral mycoprofiles of 17 subjects by assessing for changes in genera representation over time. We found that there is a large increase in patients with *Candida* mycoprofiles in cancer as opposed to in health and that only subjects with day 0 *Candida* mycoprofiles developed infections during chemotherapy, likely due to this genus's implications of becoming pathogenic in immune compromised individuals[68,69]. We discovered little to no change in mycoprofiles during the course of

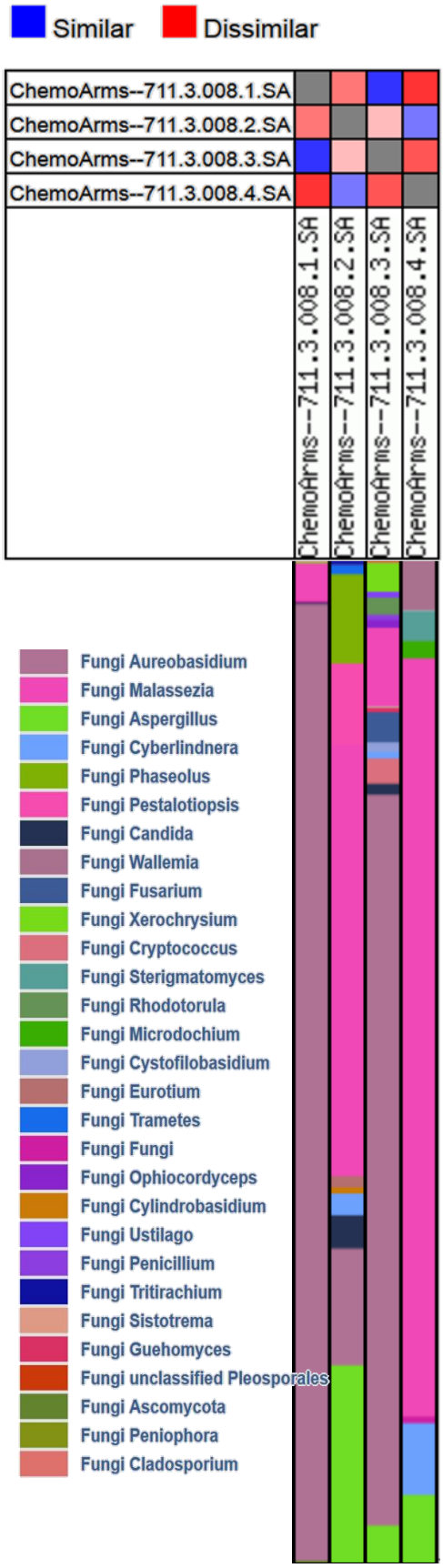


Figure 6.6: Heat map and distribution of fungal communities from an arm 3-non-naïve chemo subject with an *Aureobasidium* mycoprofile

chemotherapy at the genus level. Albeit using a small number of subjects, we revealed that patterns and changes begin to emerge at the species level in patients who develop candidiasis. Patients with *Malassezia* mycoprofiles at day 0 never developed infection, suggesting that starting with a *Malassezia* mycoprofile may lead to a healthier experience during chemotherapy. We report the first case of an *Aureobasidium* mycoprofile in a cancer patient in a survey study of metagenomic sequencing of ITS1. This mycoprofile presents a potential susceptibility for developing non-*Candida* infection[67].

Chapter 7. Discussion and Future Directions

7.1 Summary

While bacterial microbiome projects have been developed for high-throughput sequencing and analysis for the last decade, fungal metagenomic studies, specifically in human niches, have infrequently been attempted. A July 2015 Google Scholar search of “bacterial identification with 16S rRNA” yields 218,000 all-time hits and 11,500 hits for 2015. Whereas a search for “fungal identification with 18S rRNA” yields 31,800 all-time hits and 2,160 hits for 2015. Even with all of the caveats of literature searches, this difference is vast and likely reflects truths about the difficulties of working with fungi. If less than 20% of microbiome studies are investigating fungal communities, it is no wonder that reliable tools and databases for bacterial sequence assignments are much more advanced than for fungi. Assigning fungi to taxa is daunting due to names based on phenotype, redundancies in nomenclature, and lack of type specimens to validate user-deposited sequences in databases[59]. This study is the first that attempts to remedy such challenges with a biomedically relevant patient cohort, by using a curated database and manually adjusting the dataset by removing poorly aligning sequences, collapsing redundant genera, and rectifying unclassified sequences to the most reliable sequence match. The impact of our pipeline greatly simplifies the representation of taxa by eliminating members that were likely spurious and as a result, places greater emphasis on legitimate community members and potential pathogens.

7.2 Improvements to sequence generation and analysis

Because fungal analysis is still an evolving process, it is an ongoing necessity to re-evaluate the data that has been collected against updated databases. Upon bioinformatics reprocessing, at least some sequences with poor alignment to FMP references may match new fungal entries and can then be added back to the curated dataset with high confidence. In the future, at the same time that FMP is re-queried, it would be useful to compare results to other databases that

have become more fully developed since the start of this project. For example, the UNITE database has expanded and now redesignates “uncultured fungus” sequences, to more specific taxonomy where representative sequences provided this information[70]. The newly formed RefSeq database also would be useful for confirming accurate assignments against verified type specimens and for confirming accurate fungal nomenclature resolution[59].

The evolutionary age of fungi has allowed for vast differences in species and even strains within a genus. However, non-resolvable species are still apparent. Future studies should take this into account not only when assigning taxonomy to ITS1 fragments, but also in the experimental design phase. As it stands, ITS1 is shown to have a better overall rate of species identification than ITS2, but this is not the case for all genera. For example, 73% success identifications to fungal sequences were made using ITS1 vs 69% success with ITS2 across 5,407 species of Ascomycetes, but the genus *Alternaria,* one of the members we identified in the core mycobiome, includes species that are statistically identified with greater accuracy using ITS2[71]. Other regions for fungal identification have been used that may provide further resolution: large and small rRNA subunits (28S, 18S), translation elongation factor EF1-α, β-tubulin, actin, and RNA polymerase II subunits. Ribosomal genes are the most commonly used fungal identifiers, and so make up a majority of the sequences in fungal databases. The next most commonly used is EF1-α, though it still comprises a relatively small amount of fungal sequences in mycobial databases[72]. If different gene fragments are to be used in sequencing, the discontinuation and expense of support for 454 pyrosequencing must be considered. Sequencing costs associated with the Illumina platform have been reported as 75% lower than those for 454[73]. Now that we have provided a baseline framework for expectations in community membership with long reads in 454, comparisons using shorter fragments with Illumina can be made to validate its performance against our ITS1 genera distributions. Although large and small subunit genus level identification accuracies were only 60-75% for Illumina-sized

fragments (compared to ~80% for 454-sized fragments)[74], advances in small fragment mate pair sequencing on the Illumina platform could be adequate for such continuing studies.

While we propose many adjustments and modifications to protocols, there are still many limitations to be considered as part of fungal metagenomics studies. We can never be sure that we are capturing all legitimate sequences. Fungi at minimal levels may never amplify in PCR. In addition, this is still a semi-quantitative approach as the number of ITS1 copies per genome has not been taken into account. We are also limited by the databases available, and so unknown fungi cannot be identified until more medically relevant fungi have been cloned and sequenced. Resolution of nomenclature also continues to provide challenges. With database names based on their submitter preferences, it still requires hand curation to collapse redundant genera. Another issue is that there are still subjective steps in library preparation due to the unpredictable nature of clinical samples. It can be difficult to retain uniformity across samples to ensure that they are comparable. All in all fungal sequencing and identification is an evolving process.

7.3 *Candida* and *Malassezia* as oral community members

This is the first metagenomic attempt to deeply sequence the core mycobiome of saliva and offer a minimum sequencing effort to capture the majority of species richness in this niche. While our community membership was generally in concordance with other studies[23], mycobial datasets were carefully curated to give the most accurate representation of fungal profiles, given the current status of ITS1 databases. We have also illuminated a genus that was not previously known to be a commensal of the oral cavity: *Malassezia.* While *Candida* sp. have been extensively studied and are well-known as oral colonizers, little is known about *Malassezia* species due to their complicating factors in DNA isolation, special lipid requirements in culturing and growth rate of hyphae in vitro.

As the most abundant novel member of the oral community, *Malassezia* was represented by 47% of all sequences in our deep sequencing analysis of 6 healthy subjects. *Malassezia* spp. have been identified as human pathogens, causing pityriasis versicolor, seborrheic dermatitis, and dandruff[53]. *Malassezia* has also been identified in healthy human biocompartments in four other culture-independent molecular surveys: scalp[55], skin[66] , airways[19], and gastrointestinal tract[24]. *Malassezia* species are known to have especially thick cell walls (~0.12 μm)[53] that may well have contributed to Ghannoum's inability to recover *Malassezia* DNA. It is noteworthy that each of the aforementioned studies that detected *Malassezia* employed relatively harsh extraction protocols. With culturing methods, *Malassezia* has now been confirmed as a common isolate from healthy oral samples (Patricia Diaz, personal communication). Based on its predominance, *Malassezia* is likely an important oral commensal.

We found the representation of *Malassezia* and *Candida* to be inverse in nature, resembling an antagonistic relationship between the two genera. Potentially, both genera compete for the same carbon source, dextrose[75], while *Malassezia* requires additional lipids for growth. Both fungi have evolved to withstand human body temperatures while retaining the ability to switch between yeast and hyphal forms. This gives them the propensity to occupy various niches that allow one form or the other to propagate and to out-colonize other fungal types. *Malassezia globosa,* one of the two common *Malassezia* species we found in saliva, is normally found as nonpathogenic yeast on surface layers of skin, but has been shown to form hyphae that extend into deeper layers of skin to spread yeast into multiple areas[76]. Several questions remain to be answered on the importance of finding *Malassezia* in the oral cavity: Are *Malassezia* and *Candida* equilibrium maintained by host factors? Does host genotype predispose toward a *Malassezia* vs *Candida* mycoprofile? Do external factors, such as season, exposure to antifungals, and bacterial community composition have an effect on the balance of these genera?

Additionally, future studies should include quantitative efforts to assess changes in fungal load in interesting species such as *Candida albicans, Candida glabrata, Malassezia restricta, Malassezia globosa*, *Aureobasidium pullulans,* and *Cryptococcus neoformans* at all time points. Patients that developed candidiasis always started with *Candida* mycoprofiles, but from relative abundance data it is unclear whether in some subjects whether Candida load increased to contribute to infection, as the profile remains stable for the duration of the two weeks. Knowing whether or not there was an increase in fungal load of certain species could provide more information on whether species are transients or occupiers that are indeed contributing to oral mucositis and its complications. The data here is normalized to 100%, but basic gel results show that not all amplifications are created equal. Some samples, even within a subject, produce very faint ITS1 bands, while others produce very bright signals. Measuring fungal loads would be difficult because they require knowledge of ITS1 copy numbers as well as the ploidy of species that become pathogenic. Classical microscopic staining and identification could be implemented to assist in this approach for species of interest. To provide information beyond presence and absence of community members, sequencing beyond the rDNA region could provide further information about pathogenicity during the development of oral mucositis. Fungal transcripts could be sequenced to compare changes in gene expression during the disease state, since ITS1 does not necessarily provide information on if a commensal has become pathogenic. It would be interesting to see if virulence factors are more highly expressed during stages of oral mucositis even if the mycoprofile remains relatively stable. Combined with fungal load information, powerful evidence for contributors to disease would be certain to surface.

Future projects should consider more sophisticated analyses for trends in minor components. As it stands there is no clear indication of consistent community membership in the Diverse mycoprofile group and the beta diversity measurements between two time points from the same subject is very high. The non-predictive longitudinal nature of healthy subjects in this group

currently inhibits insight into changes that may occur due to disease state. Potentially the sequences gathered here could be used to combine taxa at the family, order, or class levels to see if this contributes any clarity to such inconsistencies between time points of a subject, further defining mycoprofiles that may be more predictable longitudinally.

7.4 *Candida* and the immune system

In this study we provided evidence that there is an increase in the frequency of *Candida* mycoprofiles among individuals with cancer (35% in healthy subject vs >75% in chemotherapy cohorts). Lowered immune functionality and increased frequency of *Candida* mycoprofiles for our chemotherapy cohort is likely due to factors associated with aging and with acquiring cancer[62,63]. Compounding the predispositions affiliated with cancer and age is the tendency for *Candida* sp., specifically *C. albicans,* to become pathogenic as a result of lowered immunity[68]. In mouse models, *C. albicans* colonization, combined with gastrointestinal mucosal damage and neutropenia resulted in dissemination of fungi and 100% mortality[69]. This is particularly noteworthy for chemotherapy patients who often undergo mucosal damage and explains the occurrence of *Candida* infections in individuals with *Candida* mycoprofiles in this study. *C. albicans* commensals are not harmful, but have evolved the ability to become pathogenic when appropriate conditions are met. In the GI tract for example, *C. albicans* colonize heterogeneously with a mixture of cells expressing high and low levels levels of transcription factor Efg1p, where absence of Efg1p causes hyper-colonization of the GI tract, but higher susceptibility to the host immune system[77]. When the immune system is healthy, the balance of Efg1p expressing cells tends to be toward higher expression of Efg1p, while in an immune compromised state, the selective pressure is placed on cells expressing low levels of Efgp1. With host immune defenses lowered, increased susceptibility of low-expressing Efg1p1 cells is not able to trigger a sufficient immune response, and virulent variants are able to quickly populate the niche[5].

Although it is debatable whether use of antifungals will alleviate symptoms of oral mucositis, they have been shown to reduce mucositis-induced lesion sizes in one-third of individuals displaying oral candidiasis[61]. It is especially important in a diagnostic setting to identify the species causing infection. For instance, a common treatment for *Candida* infections is the antifungal, Fluconazole. However, species of *C. glabrata* and *C. krusei* (*Pichia kudriavzevii),* two of the most prominent causes of non-*albicans* candidiasis, are resistant to this drug due to effective efflux pumps and an altered cytochrome P450, respectively[78,79]. Alternative treatments that are effective against these species must be administered.

Here we have predominantly characterized the oral mycobiome at the genus level, but it may not yet be sufficient for clinical use as anti-fungals are less broad and more specific to pathways in certain species. As part of the study (data not shown), we have used reference sequences from GenBank to assay the value of ITS1 sequences as a unique identifier for species within a given genus. For the two most prevalent community genera in this study, *Candida* and *Malassezia*, the ITS1 sequence resolves all of the species reported. In other genera, however, species cannot be resolved by ITS1. For example, three *Cladosporium* species that were assigned in our dataset were identical in the ITS1 region: *cladosporioides* (3 reference database sequences used), *tenuissimum* (3 references) and *cubutense* (1 reference). Although automated pipelines can, and often do report results at a lower taxonomic rank, it is important to realize that the phylogenetic power of resolution may well be at a higher taxonomic level, so that assumptions aren't made about the species in question.

7.5 Conclusions

This work bridges the gap in fungal metagenomic studies by providing a roadmap to handling the challenges of fungi from bench side to analysis and offers recommendations for unpredictable samples and minimum sequencing efforts, while implementing current guidelines for resolving genera and unclassified sequences[30,59].With an improved pipeline and careful

measures implemented to enhance the reliability of sequence classifications with empirically determined parameters, such as our custom e-value threshold, we have characterized the core oral mycobiome using results from deep sequencing of ITS1 amplicons and discovered a novel oral genus, *Malassezia.*

We are also the first to attempt a metagenomic study of chemotherapy patients, which allowed us to use our pipeline to characterize the frequency and presence of oral mycoprofiles. We discovered that there is a vast difference in instances of *Candida* mycoprofiles for cancer patients vs. healthy individuals and that patients with these profiles are highly susceptible to infections. With the implications of *Candida* as opportunistic pathogens[5], this is particularly important for implementing prophylactic measures that may not only keep *Candida* at bay, but also reduce the severity of oral mucositis[61].

Appendix 1. Statistics for ITS1 sequences from saliva removed during bioinformatics pipeline of 6 deeply sequenced healthy subjects

Experimental Samples			STEP 1- Dimer Removal		STEP 2- DeconSeq		STEP 3- QIIME split_libraries.py	
	Raw		Dimers		Contamination		QIIME Fail	
Total Sequences	853653		40046		1352		246905	
% of Sequences At Each Step	100.0%		4.7%		0.2%		30.4%	
% of Total Sequences Lost	N/A		4.7%		0.2%		28.9%	
# Sequences Classified by FMP	853178		40044		1338		246476	
Max Length	816		535		578		816	
Average Length	240.4		70.3		465.2		86.7	
Min Length	40		46		72		40	
Max E-value	9.8		0.076		9.1		9.8	
Median E-value	1.00E-131		1.00E-06		0.003		2.0E-04	
Min E-value	0		0		2.00E-33		0	
Number Genera Represented (uncollapsed)	607		126		103		514	
Top 10 taxa and Counts								
1	Malassezia	233862	Fungi	29493	Acaulospora	206	Mortierella	65817
2	Epicoccum	119656	Leptosphaeria	5338	Glomus	190	Cortinarius	38418
3	Mortierella	66009	Eurotium	1267	Sydowia	173	Calostoma	27738
4	Ascomycota	39878	Physciella	825	Orpinomyces	154	Sydowia	24872
5	Cyberlindnera	39574	Gongronella	339	Eurotium	57	Orpinomyces	13241
6	Cortinarius	38799	Aspergillus	271	Cladosporium	56	Malassezia	11812
7	Fungi	37092	Suillus	240	Glomeraceae	49	Cladosporium	10512
8	Fusarium	36510	Ascomycota	196	Tulasnellaceae	34	Alternaria	6449
9	Calostoma	27770	Athelia	193	Stenocarpella	29	Fungi	5511
10	Sydowia	25081	Menegazzia	135	Alternaria	24	Epicoccum	4550

STEP 4- Length Filter		STEP 5- E-value Filter			
Less than 100 bp		Unclassifiable		E-values > E-42	
7907		17716		66234	
1.4%		3.2%		11.9%	
0.9%		2.1%		7.8%	
6836		N/A		66234	
99		541		600	
92.1		223.3		286.2	
4		100		100	
8.1		N/A		10	
7.5		N/A		9.00E-21	
2.00E-13		N/A		2.00E-42	
109		N/A		596	
Cetrelia	6063	N/A		Lysurus	55362
Neopaxillus	220	N/A		Saccharomyces	1766
Saccharomyces	104	N/A		Periconia	1643
Lecanora	47	N/A		Fungi	348
Discoxylaria	39	N/A		Ascomycota	325
Volvariella	31	N/A		Microstroma	263
Phialophora	23	N/A		Cryptococcus	229
Tuber	23	N/A		Mutinus	188
Bulbothrix	15	N/A		Hygrophorus	164
Lactarius	14	N/A		Podosphaera	155

Appendix 2. Individual subject demographics for arms 1-3

Key
Collection Date- Date saliva was collected
Sample ID Entry-
Age- Age in years
Gen- Gender: M= male, F= female
Eth- Ethnicity: 0= Hispanic, 1= Non-Hispanic
Race: 1= American Indian, 2= Asian, 3= Black, 4= Other/ Pacific Islander, 5= White
OU= OMAS Ulcers
OE= OMAS Erythemas
OT= OU+OE
COC= Candidiasis
Who= WHO Scale

Highlighted samples met requirements for sequencing efforts at all four time points

Collection Date	Sample ID Entry	Age	Gen	Eth	Race	OU	OE	OT	COC	Who
1/20/2011	711-1-001-1-SA	52								
2/3/2011	711-1-001-4-SA									
1/26/2011	711-1-002-1-SA	56								
2/9/2011	711-1-002-4-SA									
2/15/2011	711-1-003-1-SA	44								
3/1/2011	711-1-003-4-SA									
2/16/2011	711-1-004-1-SA	25								
3/2/2011	711-1-004-4-SA									
2/17/2011	711-1-005-1-SA	61								
3/3/2011	711-1-005-4-SA									
2/23/2011	711-1-006-1-SA	34								
3/9/2011	711-1-006-4-SA									
4/12/2011	711-1-007-1-SA	46								
4/26/2011	711-1-007-4-SA									
7/13/2011	711-1-008-1-SA	34								
7/26/2011	711-1-008-4-SA									
7/14/2011	711-1-009-1-SA	31								
7/28/2011	711-1-009-4-SA									
9/21/2011	711-1-010-1-SA	72								
10/5/2011	711-1-010-4-SA									

2/22/2012	711-1-011-1-SA	77								
3/7/2012	711-1-011-4-SA									
2/23/2012	711-1-012-1-SA	61								
3/8/2012	711-1-012-4-SA									
2/23/2012	711-1-013-1-SA	49								
3/6/2012	711-1-013-4-SA									
5/8/2012	711-1-014-1-SA	51								
5/22/2012	711-1-014-4-SA									
6/19/2012	711-1-016-1-SA	62								
7/3/2012	711-1-016-4-SA									
7/11/2012	711-1-017-1-SA	26								
7/25/2012	711-1-017-4-SA									
8/30/2012	711-1-018-1-SA	49								
9/13/2012	711-1-018-4-SA									
7/18/2012	711-1-019-1-SA	40								
7/18/2012	711-1-019-4-SA									
7/19/2012	711-1-020-1-SA	37								
8/2/2012	711-1-020-4-SA									
9/13/2012	711-1-021-1-SA	53								
9/27/2012	711-1-021-4-SA									
9/13/2012	711-1-022-1-SA	58								
9/27/2012	711-1-022-4-SA									
10/2/2012	711-1-023-1-SA	52								
10/16/2012	711-1-023-4-SA									
10/4/2012	711-1-024-1-SA	51								
10/18/2012	711-1-024-4-SA									
10/3/2012	711-1-025-1-SA	23								
10/17/2012	711-1-025-4-SA									
10/3/2012	711-1-026-1-SA	45								
10/17/2012	711-1-026-4-SA									
10/9/2012	711-1-027-1-SA	26								
10/23/2012	711-1-027-4-SA									
10/10/2012	711-1-028-1-SA	47								
10/24/2012	711-1-028-4-SA									
10/11/2012	711-1-029-1-SA	65								
10/25/2012	711-1-029-4-SA									
10/10/2012	711-1-030-1-SA	38								
10/25/2012	711-1-030-4-SA									
1/25/2011	711-2-001-1-SA	45	1	0	5	0	0	0	0	0
1/27/2011	711-2-001-2-SA					0	0	0	0	0
2/1/2011	711-2-001-3-SA					2	1	3	0	2

2/10/2011	711-2-001-4-SA					0	0	0	0	0
2/7/2011	711-2-002-1-SA	49	1	0	5	0	0	0	0	0
2/11/2011	711-2-002-2-SA					3	8	11	0	2
2/17/2011	711-2-002-3-SA					7	6	13	0	2
2/21/2011	711-2-002-4-SA					3	6	9	0	2
4/1/2011	711-2-003-1-SA	68	1	0	5	0	0	0	0	0
4/4/2011	711-2-003-2-SA					0	1	1	0	1
4/12/2011	711-2-003-3-SA					5	5	10	0	2
4/15/2011	711-2-003-4-SA					1	1	2	0	2
4/1/2011	711-2-004-1-SA	77	1	0	5	0	0	0	0	0
4/4/2011	711-2-004-2-SA					0	0	0	0	0
4/11/2011	711-2-004-3-SA					0	2	2	0	1
4/15/2011	711-2-004-4-SA					0	0	0	0	0
5/6/2011	711-2-005-1-SA	51	1	0	5	0	0	0	0	0
5/9/2011	711-2-005-2-SA					0	0	0	1	0
5/16/2011	711-2-005-3-SA					7	7	14	0	2
5/19/2011	711-2-005-4-SA					3	4	7	0	2
8/18/2011	711-2-006-1-SA	54	1	0	5	0	0	0	0	0
8/22/2011	711-2-006-2-SA					0	0	0	0	0
8/29/2011	711-2-006-3-SA					2	6	8	0	2
9/1/2011	711-2-006-4-SA					4	3	7	0	2
10/3/2011	711-2-007-1-SA	56	1	0	5	0	0	0	0	0
10/5/2011	711-2-007-2-SA					0	0	0	0	0
10/14/2011	711-2-007-3-SA					6	5	11	1	2
10/17/2011	711-2-007-4-SA					5	5	10	0	2
10/11/2011	711-2-008-1-SA	32	2	0	5	0	0	0	0	0
10/12/2011	711-2-008-2-SA					0	0	0	0	0
10/20/2011	711-2-008-3-SA					0	0	0	0	0
10/24/2011	711-2-008-4-SA					0	0	0	0	0
10/11/2011	711-2-009-1-SA	48	2	0	5	0	0	0	0	0
10/13/2011	711-2-009-2-SA					0	0	0	0	0
10/20/2011	711-2-009-3-SA					1	3	4	0	2
10/24/2011	711-2-009-4-SA					0	2	2	0	0
10/12/2011	711-2-010-1-SA	50	1	0	5	0	0	0	0	0
10/14/2011	711-2-010-2-SA					0	0	0	0	0
1/17/2012	711-2-011-1-SA	80	1	0	5	0	0	0	0	0
1/20/2012	711-2-011-2-SA					0	0	0	0	0
1/26/2012	711-2-011-3-SA					8	7	15	1	2
2/2/2012	711-2-011-4-SA					0	0	0	0	0
1/19/2012	711-2-012-1-SA	53	2	0	5	0	0	0	0	0
1/23/2012	711-2-012-2-SA					0	1	1	0	0

1/30/2012	711-2-012-3-SA					0	0	0	0	0
2/2/2012	711-2-012-4-SA					0	0	0	0	0
2/6/2012	711-2-013-1-SA	58	2	1	5	0	0	0	0	0
2/8/2012	711-2-013-2-SA					0	0	0	0	0
2/16/2012	711-2-013-3-SA					0	0	0	0	0
2/21/2012	711-2-013-4-SA					0	0	0	0	0
2/27/2012	711-2-014-1-SA	32	2	0	5	0	0	0	0	0
2/28/2012	711-2-014-2-SA					0	0	0	0	0
3/8/2012	711-2-014-3-SA					0	0	0	0	0
3/13/2012	711-2-014-4-SA					0	0	0	0	0
3/2/2012	711-2-015-1-SA	55	1	0	5	0	0	0	0	0
3/5/2012	711-2-015-2-SA					0	0	0	0	0
3/9/2012	711-2-015-3-SA					2	9	11	0	2
3/15/2012	711-2-015-4-SA					3	5	8	0	2
3/16/2012	711-2-016-1-SA	91	2	0	5	0	0	0	0	0
3/19/2012	711-2-016-2-SA					0	0	0	0	0
3/27/2012	711-2-016-3-SA					0	0	0	0	0
3/30/2012	711-2-016-4-SA					0	0	0	0	0
3/28/2012	711-2-017-1-SA	54	1	0	5	0	0	0	0	0
4/2/2012	711-2-017-2-SA					0	2	2	0	1
4/5/2012	711-2-017-3-SA					3	2	5	0	2
4/12/2012	711-2-017-4-SA					0	1	1	0	0
3/30/2012	711-2-018-1-SA	63	1	0	5	0	0	0	0	0
4/2/2012	711-2-018-2-SA					1	6	7	0	2
4/9/2012	711-2-018-3-SA					3	4	7	0	2
4/12/2012	711-2-018-4-SA					1	1	2	0	2
5/16/2012	711-2-019-1-SA	66	2	0	3	0	0	0	0	0
5/17/2012	711-2-019-2-SA					0	0	0	0	0
5/23/2012	711-2-019-3-SA					0	2	2	0	1
6/1/2012	711-2-019-4-SA					0	0	0	0	0
7/3/2012	711-2-020-1-SA	50	1	0	5	0	0	0	0	0
7/5/2012	711-2-020-2-SA					0	0	0	1	0
7/12/2012	711-2-020-3-SA					1	2	3	1	2
7/18/2012	711-2-020-4-SA					1	9	10	1	2
7/9/2012	711-2-021-1-SA	60	2	0	5	0	0	0	0	0
7/13/2012	711-2-021-2-SA					2	2	4	0	2
7/18/2012	711-2-021-3-SA					5	6	11	0	2
7/23/2012	711-2-021-4-SA					0	6	6	0	1
9/26/2012	711-2-022-1-SA	67	1	0	3	0	0	0	0	0
10/2/2012	711-2-022-2-SA					0	0	0	0	0
10/4/2012	711-2-022-3-SA					0	2	2	0	1

10/10/2012	711-2-022-4-SA					1	2	3	0	0
11/7/2012	711-2-025-1-SA	80	1	0	3	0	0	0	0	0
11/12/2012	711-2-025-2-SA					0	0	0	0	0
11/14/2012	711-2-025-3-SA					1	2	3	0	2
11/21/2012	711-2-025-4-SA					0	0	0	0	0
11/14/2012	711-2-026-1-SA	35	2	0	5	0	0	0	0	0
11/16/2012	711-2-026-2-SA					0	0	0	0	0
11/26/2012	711-2-026-3-SA					2	3	5	0	0
11/29/2012	711-2-026-4-SA					0	0	0	0	0
12/12/2012	711-2-027-1-SA	63	1	0	5	0	0	0	0	0
12/13/2012	711-2-027-2-SA					0	0	0	0	0
12/20/2012	711-2-027-3-SA					6	6	12	0	2
12/27/2012	711-2-027-4-SA					1	1	2	0	2
2/25/2013	711-2-028-1-SA	65	2	0	5	0	0	0	0	0
2/28/2013	711-2-028-2-SA					0	2	2	0	0
3/7/2013	711-2-028-3-SA					0	0	0	0	0
3/13/2013	711-2-028-4-SA					0	0	0	0	0
3/13/2013	711-2-029-1-SA	63	1	0	5	0	0	0	0	0
3/15/2013	711-2-029-2-SA					1	2	3	1	2
3/21/2013	711-2-029-3-SA					0	0	0	0	0
3/27/2013	711-2-029-4-SA					0	0	0	0	0
4/19/2013	711-2-030-1-SA	53	1	0	5	0	0	0	0	0
4/23/2013	711-2-030-2-SA					0	0	0	0	0
5/3/2013	711-2-030-4-SA					0	0	0	0	0
5/13/2013	711-2-031-1-SA	62	1	0	5	0	0	0	0	0
5/16/2013	711-2-031-2-SA					0	1	1	0	0
5/21/2013	711-2-031-3-SA					3	7	10	0	2
5/24/2013	711-2-031-4-SA					2	2	4	0	2
7/16/2013	711-2-032-1-SA	82	1	0	5	0	0	0	0	0
7/18/2013	711-2-032-2-SA					0	1	1	0	0
7/23/2013	711-2-032-3-SA					1	1	2	0	2
7/29/2013	711-2-032-4-SA					2	3	5	0	2
8/5/2013	711-2-035-1-SA	52	1	0	5	0	0	0	0	0
8/7/2013	711-2-035-2-SA					0	1	1	0	0
8/12/2013	711-2-035-3-SA					0	0	0	0	0
8/19/2013	711-2-035-4-SA					0	0	0	0	0
7/24/2012	711-3-001-1-SA	50	1	0	5	0	0	0	0	0
8/8/2012	711-3-001-2-SA					0	0	0	1	0
8/16/2012	711-3-001-3-SA					0	5	5	1	1
8/22/2012	711-3-001-4-SA					0	0	0	1	0
8/6/2012	711-3-002-1-SA	60	2	0	5	0	0	0	0	0

8/8/2012	711-3-002-2-SA					0	0	0	0	0
8/15/2012	711-3-002-3-SA					6	4	10	0	3
10/12/2012	711-3-003-1-SA	42	2	0	5	0	0	0	0	0
10/15/2012	711-3-003-2-SA					0	0	0	0	0
10/19/2012	711-3-003-3-SA					0	0	0	0	0
10/22/2012	711-3-004-1-SA	64	2	0	5	0	0	0	0	0
10/24/2012	711-3-004-2-SA					0	0	0	0	0
11/5/2012	711-3-004-4-SA					7	5	12	0	2
10/24/2012	711-3-005-1-SA	45	2	0	5	0	0	0	0	0
10/25/2012	711-3-005-2-SA					0	0	0	0	0
10/31/2012	711-3-005-3-SA					0	4	4	1	0
11/7/2012	711-3-005-4-SA					0	2	2	0	1
11/2/2012	711-3-006-1-SA	46	2	0	5	0	0	0	0	0
11/5/2012	711-3-006-2-SA					0	0	0	0	0
11/9/2012	711-3-006-3-SA					1	4	5	0	2
11/15/2012	711-3-006-4-SA					0	0	0	0	0
11/29/2012	711-3-007-1-SA	70	2	0	5	0	0	0	0	0
11/30/2012	711-3-007-2-SA					0	0	0	0	0
11/30/2012	711-3-008-1-SA	71	1	0	3	0	0	0	0	0
12/5/2012	711-3-008-2-SA					0	1	1	0	1
12/7/2012	711-3-008-3-SA					1	3	4	0	2
12/13/2012	711-3-008-4-SA					1	2	3	0	2
12/13/2012	711-3-009-1-SA	36	2	0	5	0	0	0	0	0
12/18/2012	711-3-009-2-SA					0	0	0	0	0
12/20/2012	711-3-009-3-SA					0	0	0	0	0
12/27/2012	711-3-009-4-SA					1	1	2	0	1
1/16/2013	711-3-010-1-SA	59	2	0	5	0	0	0	0	0
1/18/2013	711-3-010-2-SA					0	0	0	0	0
1/23/2013	711-3-010-3-SA					0	0	0	0	0
1/30/2013	711-3-010-4-SA					0	0	0	0	0
1/16/2013	711-3-011-1-SA	63	2	0	5	0	0	0	0	0
1/18/2013	711-3-011-2-SA					0	0	0	0	0
1/24/2013	711-3-011-3-SA					2	2	4	0	2
1/30/2013	711-3-011-4-SA					1	3	4	0	2
3/14/2013	711-3-012-1-SA	64	2	0	3	0	0	0	0	0
3/15/2013	711-3-012-2-SA					0	0	0	0	0
3/21/2013	711-3-012-3-SA					0	0	0	0	0
3/28/2013	711-3-012-4-SA					0	0	0	0	0
4/10/2013	711-3-013-1-SA	63	1	0	5	0	0	0	0	0
4/26/2013	711-3-013-4-SA					0	0	0	0	0
4/10/2013	711-3-014-1-SA	66	2	0	5	0	0	0	0	0

4/11/2013	711-3-014-2-SA					0	0	0	0	0
4/18/2013	711-3-014-3-SA					1	1	2	0	2
4/24/2013	711-3-014-4-SA					0	0	0	0	0
6/3/2013	711-3-015-1-SA	62	1	0	5	0	0	0	0	0
6/6/2013	711-3-015-2-SA					0	0	0	0	0
6/13/2013	711-3-015-3-SA					14	10	24	0	2
6/19/2013	711-3-015-4-SA					4	4	8	0	2
6/26/2013	711-3-016-1-SA	49	2	1	5	0	0	0	0	0
6/28/2013	711-3-016-2-SA					0	0	0	0	0
7/8/2013	711-3-016-3-SA					0	0	0	0	0
7/9/2013	711-3-016-4-SA					0	0	0	0	0

References

1. Hawksworth, D. L. The magnitude of fungal diversity : the 1.5 million species estimate revisited *. *Mycol. Res.* **105,** 1422–1432 (2001).

2. Blackwell, M. The fungi: 1, 2, 3 ... 5.1 million species? *Am. J. Bot.* **98,** 426–38 (2011).

3. Fisher, M. C. *et al.* Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484,** 186–94 (2012).

4. Casadevall, A. Fungal virulence, vertebrate endothermy, and dinosaur extinction: is there a connection? *Fungal Genet. Biol.* **42,** 98–106 (2005).

5. Pierce, J. V. & Kumamoto, C. a. Variation in Candida albicans EFG1 expression enables host-dependent changes in colonizing fungal populations. *MBio* **3,** 1–8 (2012).

6. Litvintseva, A. P., Brandt, M. E., Mody, R. K. & Lockhart, S. R. Investigating fungal outbreaks in the 21st century. *PLoS Pathog.* **11,** e1004804 (2015).

7. Naggie, S. & Perfect, J. R. Molds: hyalohyphomycosis, phaeohyphomycosis, and zygomycosis. *Clin. Chest Med.* **30,** 337–53 (2009).

8. Steenbergen, J. N. & Casadevall, A. The origin and maintenance of virulence for the human pathogenic fungus Cryptococcus neoformans. *Microbes Infect.* **5,** 667–75 (2003).

9. Zaragoza, O. *et al.* Capsule enlargement in Cryptococcus neoformans confers resistance to oxidative stress suggesting a mechanism for intracellular survival. *Cell. Microbiol.* **10,** 2043–57 (2008).

10. Casadevall, A., Steenbergen, J. N. & Nosanchuk, J. D. 'Ready made' virulence and 'dual use' virulence factors in pathogenic environmental fungi - The Cryptococcus neoformans paradigm. *Curr. Opin. Microbiol.* **6,** 332–337 (2003).

11. Bicanic, T. & Harrison, T. S. Cryptococcal meningitis. *Br. Med. Bull.* **72,** 99–118 (2004).

12. Scully, C., Sonis, S. & Diz, P. D. Oral mucositis. *Oral Dis.* **12,** 229–41 (2006).

13. Ye, Y. *et al.* Oral bacterial community dynamics in paediatric patients with malignancies in relation to chemotherapy-related oral mucositis: a prospective study. *Clin. Microbiol. Infect.* 1–9 (2013). doi:10.1111/1469-0691.12287

14. Sonis, S. T. Mucositis: The impact, biology and therapeutic opportunities of oral mucositis. *Oral Oncol.* **45,** 1015–20 (2009).

15. Lalla, R. V *et al.* A systematic review of oral fungal infections in patients receiving cancer therapy. *Support Care Cancer* **18,** 985–92 (2010).

16. Allison, R. R. *et al.* Multi-institutional, randomized, double-blind, placebo-controlled trial to assess the efficacy of a mucoadhesive hydrogel (MuGard) in mitigating oral mucositis

symptoms in patients being treated with chemoradiation therapy for cancers of the head and neck. *Cancer* **120,** 1433–1440 (2014).

17. Tuite, N. & Lacey, K. Overview of Invasive Fungal Infections. *Methods Mol. Biol.* **968,** 1–23 (2013).

18. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 6241–6 (2012).

19. Delhaes, L. *et al.* The airway microbiota in cystic fibrosis: a complex fungal and bacterial community--implications for therapeutic management. *PLoS One* **7,** e36313 (2012).

20. Taylor, D. L. & Houston, S. A Bioinformatics Pipeline for Sequence-Based Analyses of Fungal Biodiversity. *Methods Mol. Biol.* **722,** 141–155 (2011).

21. Bellemain, E. *et al.* ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiol.* **10,** 1–9 (2010).

22. Bazzicalupo, A. L., Bálint, M. & Schmitt, I. Comparison of ITS1 and ITS2 rDNA in 454 sequencing of hyperdiverse fungal communities. *Fungal Ecol.* **6,** 102–109 (2013).

23. Ghannoum, M. a *et al.* Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog.* **6,** 1–8 (2010).

24. Hamad, I., Sokhna, C., Raoult, D. & Bittar, F. Molecular detection of eukaryotes in a single human stool sample from Senegal. *PLoS One* **7,** 1–8 (2012).

25. Hawksworth, D. L. *et al.* The amsterdam declaration on fungal nomenclature. *IMA Fungus* **2,** 105–12 (2011).

26. Hawksworth, D. L., McNeill, J., de Beer, Z. W. & Wingfield, M. J. Names of fungal species with the same epithet applied to different morphs: how to treat them. *IMA Fungus* **4,** 53–6 (2013).

27. Dupuy, A. K. *et al.* Redefining the human oral mycobiome with improved practices in amplicon-based taxonomy: discovery of Malassezia as a prominent commensal. *PLoS One* **9,** e90899 (2014).

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

29. White TJ, Bruns T, Lee S, W. J. in *PCR Protocols: A Guide to Methods and Applications* (ed. Innis MA, Gelfand DH, Sninksy JJ, W. T.) 315–322 (New York: Academic Press, 1990).

30. Hawksworth, D. L. Managing and coping with names of pleomorphic fungi in a period of transition. *IMA Fungus* **3,** 15–24 (2012).

31. Braun, U. The impacts of the discontinuation of dual nomenclature of pleomorphic fungi: the trivial facts, problems, and strategies. *IMA Fungus* **3,** 81–6 (2012).

32. Huse, S. *et al.* VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics* **15,** 41+ (2014).

33. Mseddi, F., Jarboui, M. A., Sellami, A., Sellami, H. & Ayadi, A. A rapid and easy method for the DNA extraction from Cryptococcus neoformans. *Biol. Proced. Online* **13,** 5 (2011).

34. Bolano, A. *et al.* Rapid methods to extract DNA and RNA from Cryptococcus neoformans. *FEMS Yeast Res.* **1,** 221–224 (2001).

35. Rogers S, B. A. in *Plant Molecular Biology Manual D1* 183–190 (Kluwer Academic Publishers, 1994). doi:10.4319/lo.2013.58.2.0489

36. Raeder U, B. P. Rapid preparation of DNA from filamentous fungi. *Lett. Appl. Microbiol.* **1,** 17–20 (2008).

37. Weiland, J. J. Rapid procedure for the extraction of DNA from fungal spores and mycelia. *Fungal Genet. Newsl.* **44,** 60–63 (1997).

38. Fredricks, D. N., Smith, C. & Meier, A. Comparison of Six DNA Extraction Methods for Recovery of Fungal DNA as Assessed by Quantitative PCR Comparison of Six DNA Extraction Methods for Recovery of Fungal DNA as Assessed by Quantitative PCR. *Society* **43,** (2005).

39. Chen, Y. C. *et al.* Polymorphic internal transcribed spacer region 1 DNA sequences identify medically important yeasts. *J. Clin. Microbiol.* **39,** 4042–4051 (2001).

40. Irinyi, L. & Meyer, W. DNA barcoding of human and animal pathogenic fungi: the ISHAM-ITS database. *Microbiol. Aust.* **36,** 44 (2015).

41. Gonzalez, J. M., Portillo, M. C., Belda-Ferre, P. & Mira, A. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS One* **7,** (2012).

42. Nilsson, R. H., Kristiansson, E., Ryberg, M. & Hallenberg, N. Intraspeci fi c ITS Variability in the Kingdom Fungi as Expressed in the International Sequence Databases and Its Implications for Molecular Species Identi fi cation. 193–201 (2008).

43. LaTuga, M. S. *et al.* Beyond bacteria: A study of the enteric microbial consortium in extremely low birth weight infants. *PLoS One* **6,** 1–10 (2011).

44. Hubka, V., Kolarík, M., Kubátová, A. & Peterson, S. W. Taxonomic revision of Eurotium and transfer of species to Aspergillus. *Mycologia* **105,** 912–37 (2013).

45. Nonnenmann, M. W. *et al.* Utilizing pyrosequencing and quantitative PCR to characterize fungal populations among house dust samples. *J. Environ. Monit.* **14,** 2038 (2012).

46. Shelton, B. G., Kirkland, K. H., Flanders, W. D. & Morris, G. K. Profiles of Airborne Fungi in Buildings and Outdoor Environments in the United States. *Appl. Environ. Microbiol.* **68,** 1743–1753 (2002).

47. Amend, A. S., Seifert, K. a, Samson, R. & Bruns, T. D. Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 13748–53 (2010).

48. Everett, J. E., Busick, N. P., Sielaff, T., Wahoff, D. C. & Dunn, D. L. A deeply invasive Phoma species infection in a renal transplant recipient. *Transplant. Proc.* **35,** 1387–1389 (2003).

49. Suh, S.-O., Blackwell, M., Kurtzman, C. P. & Lachance, M. -a. Phylogenetics of Saccharomycetales, the ascomycete yeasts. *Mycologia* **98,** 1006–1017 (2007).

50. Buzina, W. *et al.* The Polypore Mushroom Irpex lacteus, a New Causative Agent of Fungal Infections. *J. Clin. Microbiol.* **43,** 2009–2011 (2005).

51. Kalkanci, A. *et al.* Fulminating fungal sinusitis caused by Valsa sordida, a plant pathogen, in a patient immunocompromised by acute myeloid leukemia. *Med. Mycol.* **44,** 531–9 (2006).

52. Morris, J., Beckius, M. & McAllister, C. Sporobolomyces Infection in an AIDS Patient. *J. Infect. Dis.* **164,** 623–624 (1991).

53. Ashbee, H. R. & Evans, E. G. V. Immunology of Diseases Associated with Malassezia Species. *Clin. Microbiol. Rev.* **15,** 21–57 (2002).

54. Saunders, C. W., Scheynius, A. & Heitman, J. Malassezia fungi are specialized to live on skin and associated with dandruff, eczema, and other skin diseases. *PLoS Pathog.* **8,** e1002701 (2012).

55. Park, H. K. *et al.* Characterization of the fungal microbiota (mycobiome) in healthy and dandruff-afflicted human scalps. *PLoS One* **7,** 1–6 (2012).

56. Yoshikawa T. Characterization of Malassezia spp. in Oral Cavity of Dog. *Oral-Med Sci* **7,** 72–76 (2008).

57. Chang, H. *et al*. An epidemic of Malassezia pachydermatis in an intensive care nursery associated with colonization of health care workers' pet dogs. *New Englahd J. Med.* **338,** 706–711 (1998).

58. Nagano, Y. *et al.* Comparison of techniques to examine the diversity of fungi in adult patients with cystic fibrosis. *Med. Mycol.* **48,** 166–76.e1 (2010).

59. Schoch, C. L. *et al.* Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database (Oxford).* **2014,** 1–21 (2014).

60. Diaz, P. I. *et al.* Using high throughput sequencing to explore the biodiversity in oral bacterial communities. *Mol. Oral Microbiol.* **27,** 182–201 (2012).

61. Leung, W. K., Dassanayake, R. S. & Yau, J. Y. Y. Oral Colonization, Phenotypic , and Genotypic Profiles of Candida Species in Irradiated, Dentate, Xerostomic Nasopharyngeal Carcinoma Survivors. *J. Clin. Microbiol.* **38,** 2219–2226 (2000).

62. Zaremba, M. L. *et al.* Incidence rate of Candida species in the oral cavity of middle-aged and elderly subjects. *Adv. Med. Sci.* **51 Suppl 1,** 233–236 (2006).

63. Bakun, M. *et al.* Urine proteomes of healthy aging humans reveal extracellular matrix (ECM) alterations and immune system dysfunction. *Age (Omaha).* 1–13 (2013). doi:10.1007/s11357-013-9562-7

64. Venturini, J., De Camargo, M. R., Félix, M. C., Vilani-Moreno, F. R. & De Arruda, M. S. P. Influence of tumour condition on the macrophage activity in candida albicans infection. *Scand. J. Immunol.* **70,** 10–17 (2009).

65. Palmer, M. K. WHO Handbook for Reporting Results of Cancer Treatment. *British journal of cancer* **45,** 484–485 (1982).

66. Findley, K. *et al.* Topographic diversity of fungal and bacterial communities in human skin. *Nature* **498,** 367–70 (2013).

67. De Oliveira, L. R. *et al.* Aureobasidium pullulans infection in a patient with chronic lymphocytic leukemia. *Rev. Soc. Bras. Med. Trop.* **46,** 660–662 (2013).

68. Rüping, M. J. G. T., Vehreschild, J. J. & Cornely, O. a. Patients at high risk of invasive fungal infections: when and how to treat. *Drugs* **68,** 1941–1962 (2008).

69. Koh, A. Y., Köhler, J. R., Coggshall, K. T., Van Rooijen, N. & Pier, G. B. Mucosal damage and neutropenia are required for Candida albicans dissemination. *PLoS Pathog.* **4,** (2008).

70. Kõljalg, U. *et al.* Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* **22,** 5271–5277 (2013).

71. Wang X, Liu C, Huang L, Bengtsson-Palme J, Chen H, Zhang J, Cai D, L. J. ITS1: a DNA barcode better than ITS2 in eukaryotes? *Mol. Ecol. Resour.* **15,** 573–586 (2015).

72. Mahé, S. *et al.* PHYMYCO-DB: A Curated Database for Analyses of Fungal Diversity and Evolution. *PLoS One* **7,** (2012).

73. Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7,** (2012).

74. Liu, K. L., Porras-Alfaro, A., Kuske, C. R., Eichorst, S. a. & Xie, G. Accurate, rapid taxonomic classification of fungal large-subunit rRNA Genes. *Appl. Environ. Microbiol.* **78,** 1523–1533 (2012).

75. Saigal S, Bhargava A, Mehra SK, D. F. Identification of Candida albicans by using different culture medias and its association in potentially malignant and malignant lesions. *Contemp. Clin. Dent.* **2,** 188–193 (2011).

76. Brand, A. Hyphal growth in human fungal pathogens and its role in virulence. *Int. J. Microbiol.* **2012,** (2012).

77. Pierces, J. V., Dignard, D., Whiteway, M. & Kumamoto, C. a. Normal adaptation of Candida albicans to the murine gastrointestinal tract requires Efg1p-dependent regulation of metabolic and host defense genes. *Eukaryot. Cell* **12,** 37–49 (2013).

78. Tsai, H. F., Krol, A. a., Sarti, K. E. & Bennett, J. E. Candida glabrata PDR1, a transcriptional regulator of a pleiotropic drug resistance network, mediates azole resistance in clinical isolates and petite mutants. *Antimicrob. Agents Chemother.* **50,** 1384–1392 (2006).

79. Orozco, A. S. *et al.* Mechanism of fluconazole resistance in Candida krusei. *Antimicrob. Agents Chemother.* **42,** 2645–2649 (1998).