2013

Dynamics of Microeukaryotes and Archaea in the Mammalian Gut Microbiome

Serena Nicole Dollive
University of Pennsylvania, sdollive@mail.med.upenn.edu

Dynamics of Microeukaryotes and Archaea in the Mammalian Gut Microbiome

Abstract

The vertebrate microbiome consists of the bacteria, fungi, archaea, protozoans, and viruses that inhabit the body at diverse locations including the skin, mouth, upper airways, urogenital tract, and digestive tract. These microorganisms are known to synthesize vitamins, interact with and tone the immune system, and dramatically affect human health. A long list of diseases has been associated with imbalances in commensal microbiome communities. The work presented in this dissertation aims to characterize the microeukaryotic and archaeal components of the gut microbiome through development of wet lab techniques and in silico methods, and apply them to the study of response to antibiotics. These methods provided a picture of the healthy fungal and archaeal communities in the gut, with high prevalence of the yeast Saccharomyces and the archaeon Methanobrevibactor, along with several other species. These new tools were then used to investigate the longitudinal changes that the microbiome undergoes when treated with heavy antibiotics. Using an antibiotic cocktail containing ampicillin, neomycin, vancomycin, and metronidazole in a mouse model, we found that bacterial communities were effectively suppressed and fungi grew out by one to two orders of magnitude. After we discontinued antibiotics, bacterial and fungal cell counts returned to baseline levels within one week, but community composition was still significantly altered. Eight weeks after cessation of antibiotics, fungal community composition was not significantly different from non-treated controls, but several mice continued to have elevated levels of yeasts that had grown out during antibiotic treatment. The bacterial community composition was still significantly different from non-treated controls. Ultimately, this work demonstrated potentially deleterious long term effects of antibiotic use, and emphasizes how strong cage effects can be in mouse studies. The research performed in this dissertation will aid researchers looking to study all three domains of life and take into account the effects of commonly used antibiotics in future microbiome studies.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Frederic D. Bushman

Keywords

Archaea, Fungi, Gut, ITS, Microbiome, Microeukaryotes

Subject Categories

Bioinformatics | Microbiology

DYNAMICS OF MICROEUKARYOTES AND ARCHAEA IN THE MAMMALIAN GUT MICROBIOME

Serena Dollive

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

_____

Frederic D. Bushman, Ph.D., Professor, Microbiology

Graduate Group Chairperson

_____

Maja Bucan, Ph.D., Professor, Genetics

Dissertation Committee

Christian J. Stoekert, Ph.D., Research Professor of Genetics

Ronald G. Collman, M.D., Professor of Medicine

Hongzhe Li, Ph.D., Professor of Biostatistics

Paul D. Sniegowski, Ph.D., Assistant Professor of Biology

DYNAMICS OF MICROEUKARYOTES AND ARCHAEA IN THE MAMMALIAN GUT MICROBIOME

Dedication

To my mother, who has always been there for me.

ACKNOWLEDGMENTS

I would first like to thank my advisor, Rick Bushman for his scientific tutelage, support, and persistent championing for forward motion. Under Rick's mentorship, I have not just developed as a scientist but gained a wide set of life skills that will be invaluable going forward. I would also like to thank our collaborators in the Human Microbiome Project at Penn. I am especially obliged to Gary Wu and Jim Lewis for their help, support, advice, and insights. I am also grateful to my thesis committee members: Chris Stoeckert, Ron Collman, Hongzhe Li, and Paul D. Sniegowski, for their time, knowledge, recommendations, and investment in my graduate career.

I am thankful for the lessons and skills I have learned from fellow lab members. Chris Hoffmann taught me nearly all of the wet lab techniques I know today and had an immense amount of patience with me as a very green rotation student. Through my years in the Bushman Lab, Chris has been unbelievably helpful and a great friend. I am indebted to Kyle Bittinger for his advice, cheerleading, and persistent willingness to put aside his work and teach me. I am thankful to my former baymate Shoshanna Roth who took the time to help me navigate through the academic world as a beginning graduate student. I am also grateful to Ying-Yu Chen and Stephanie Grunberg for amount of work and effort they put into the collaborative projects in this dissertation. I am obliged to the microbiology graduate students in the Bushman lab whose graduate careers coincided with mine: Emily Charlson, Greg Peterfreund, and Sam Minot. Having companions with whom I could commiserate and celebrate from classes to dissertation was incredibly heartening. I am glad for the support and friendship from past and present lab members: Rohini Sinha, Nirav Malani, Aubrey Bailey, Young Hwang, Jimmy Hu, Sana Attar, Brendan Kelly, Frances Male, Troy Brady, Alexandra Bryson, Jacque Young, Troy Brady,

Rebecca Custers-Allen, Jennifer Hwang, Christel Chehoud, Erik Clarke, Karen Ocwieja, Mali Skotheim, Caitlin Greig, and Laurie Zimmerman.

I am also extremely grateful for the support of the GCB faculty and students. Their enthusiasm and eagerness to help made my graduate student carrier so much more enlightening and exciting. I am particularly indebted to Maja Bucan, who encouraged and supported me. I am very appreciative of my fellow GCB graduate students whose friendship and solidarity bettered my graduate school experience. I am thankful to have had collaborative and congenial cohort-mates in Ellen Tsai, Aleah Caulin, Nick Stong, and Sira Sriswasdi, whose comradery made prelims bearable. I am glad to have been able to get to know and spend time with many of my GCB classmates, including Scott Sherrill-Mix, Rithun Mukherjee, Hannah Dueck, Yih-Chii Hwang, Sarah Middleton, Varun Aggarwala, Fan Li, Kathleen Sprouffske, Rumen Kostadinov, Paul Ryvkin, Jun Chen, Anthony Olarerin, Miler Lee, Erik Clarke, Hannah Hutton, Ryan Coleman, Brett Hannigan, Najaf Shah, and others. I am especially grateful to Hannah Chervitz and Tiffany Barlow for their continual support and help over the past five years.

I would like to acknowledge the friends I have made at Penn who have made Philadelephia my home. Friends like Lauren Friedman, Kristy Skowronski, Shanshan Ding, Yiwen Song, Grace Chao, Katherine O'Brien, Siqing He, Viktoriya Syrovatkina, Judite Costa, and Anton Kutovoy made the years I have spent at Penn wonderful. I would like to thank my family: my mother, Janice, my siblings, Blake and Monica, my grandmother, Virginia, my Aunt Kathy, Uncle John, Uncle Cliff, cousins Brian and Tim, and partner and best friend Tom who have been my biggest fans and supporters.

ABSTRACT

DYNAMICS OF MICROEUKARYOTES AND ARCHAEA IN THE MAMMALIAN GUT MICROBIOME

Serena Dollive

Fredrick D. Bushman

The vertebrate microbiome consists of the bacteria, fungi, archaea, protozoans, and viruses that inhabit the body at diverse locations including the skin, mouth, upper airways, urogenital tract, and digestive tract. These microorganisms are known to synthesize vitamins, interact with and tone the immune system, and dramatically affect human health. A long list of diseases has been associated with imbalances in commensal microbiome communities. The work presented in this dissertation aims to characterize the microeukaryotic and archaeal components of the gut microbiome through development of wet lab techniques and *in silico* methods, and apply them to the study of response to antibiotics. These methods provided a picture of the healthy fungal and archaeal communities in the gut, with high prevalence of the yeast *Saccharomyces* and the archaeon *Methanobrevibactor*, along with several other species. These new tools were then used to investigate the longitudinal changes that the microbiome undergoes when treated with heavy antibiotics. Using an antibiotic cocktail containing ampicillin, neomycin, vancomycin, and metronidazole in a mouse model, we found that bacterial communities were effectively suppressed and fungi grew out by one to two orders of magnitude. After we discontinued antibiotics, bacterial and fungal cell counts returned to baseline levels within one week, but community composition was still significantly altered. Eight weeks after cessation of antibiotics, fungal community composition was not significantly different from non-treated controls, but several mice continued to have elevated levels of yeasts that had grown out during

antibiotic treatment. The bacterial community composition was still significantly different from non-treated controls. Ultimately, this work demonstrated potentially deleterious long term effects of antibiotic use, and emphasizes how strong cage effects can be in mouse studies. The research performed in this dissertation will aid researchers looking to study all three domains of life and take into account the effects of commonly used antibiotics in future microbiome studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

History of Microbiome Studies

For over a century, microbial life has been known to live in the healthy human gut [1]. Early methods to study the microbiome were restricted to culture based methods [2] and light microscopy [1]. These studies were extremely limited, because only a small fraction of bacteria are culturable [3]. Standard culturing methods may not detect species that are restrictive anaerobes or auxotrophic. Also, culture based studies are highly biased towards species with the ability to grow quickly and thrive in the selection media used in a specific study [4]. Basic microscopic techniques often have difficulty differentiating between species with similar size and morphology. Advanced phylogenetic stains using Fluorescent In Situ Hybridization (FISH) are available, but only provide visual information about a microbial community [5].

Microbial community characterization was greatly advanced by the invention of Sanger sequencing technology in 1977 [6], which did not have the extreme biases seen in culture based methods and enabled researchers to study a much larger proportion of a microbial population. Early studies used clone based Sanger sequencing, which provided a larger view of microbiome communities but was still hampered by low throughput limitations [7]. The advent of high throughput DNA sequencing in the mid 2000s with early next generation sequencing technologies such as 454 Pyrosequencing[8], Solexa sequencing [9], and SOLiD sequencing [10], enabled researchers to produce a large number of reads at a much lower cost than the older Sanger method and generated read coverages large enough to reflect the true microbial populations in diverse microbial communities.

Characterization of Microbiome Communities

There are several strategies used to characterize complex microbial communities. The most popular method for simple analysis is amplicon based sequencing. Amplicon based sequencing consists of targeted amplification of a conserved, yet polymorphic region that appears in all members of the target clade. Ideally, the primer landing sites should be sufficiently conserved to match the primers and amplify without bias for different species, and the genomic DNA that lies between the primers must be sufficiently variable between related species to allow differentiation [11]. Many different genes are used for amplicon based sequencing in different clades. The 16S ribosomal RNA gene is used almost exclusively in bacteria [12], although other amplicons have been proposed [13]. The 16S rRNA and rpoB genes [14] are used in archaeal sequencing. The 18S rRNA gene, 28S rRNA gene, and Internal Transcribed Spacers (ITS 1 and 2) are often used to characterize fungi [15] and protozoa [16,17,18]. Amplicon based sequencing allows the investigator to identify community members rigorously, but does not provide any details about community gene content more broadly.

Shotgun metagenomic sequencing has become increasingly popular for characterizing microbiome communities. Metagenomic sequencing of communities is performed similarly to genome sequencing: DNA is extracted, shredded, and sequenced [19]. Once sequencing is complete, data is assembled into contigs through traditional genome assembly methods [20] or specialized methods, such as de Bruijn graph assembly [21] or binning with self organizing maps [22]. Data can be analyzed to determine which genes and metabolic capabilities are present in a community. Nevertheless, taxonomic attribution is difficult with shotgun sequencing, because many of the genes found may not be present in available databases. Some researchers use 16S genes mined from sequencing data to estimate the relative abundance of species [19]. Others have proposed using many phylogenetic markers simultaneously to calculate the relative abundance of

species found [23]. Even with improved classification methods, a bias against the detection of rare community members still exists. In recent years, shotgun sequencing has become increasingly popular with increases in read length and throughput in Solexa sequencing [19], which enables better sample coverage and more thorough community characterization.

Recently sequencing of bulk RNA (RNA-Seq) from microbiome samples has become popular [24]. Such studies allow researchers to determine which genes are being actively transcribed within a microbial community, which is not possible with DNA sequencing. These studies can elucidate community dynamics and help determine which community members are responsible for metabolic activities [25]. Despite these useful insights, RNA-Seq does offer new bioinformatic challenges. RNAs cannot be assembled into large contigs like genomic DNA: reads are often mapped back to reference genomes, and analysis quality is more dependent on the number and quality of reference genomes [26].

The Gut Microbiome

The microbiome consists of the bacteria, archaea, fungi, protozoa, and viruses that live on and in the human body. It is estimated that microbial cells outnumber human host cells by 10 to 1 [27], and the cumulative gene content is approximately 3.3 million protein coding genes [19]. Commonly studied body sites include the mouth [28], skin [29], upper airways [30], urogenital tract [31], and digestive tract [32]. Different body sites have been documented to contain unique communities [33], mostly attributable to differences between sites such as pH, oxygen content, salinity, temperature, nutrients available at the site, and other factors [33,34,35,36].

The digestive tract is of particular interest, because it contains the majority of commensal microbiome cells [19]. A majority of the host immune system cells line the digestive tract and interact with the microbiome [37]. Furthermore many digestive diseases such as Crohn's disease

[38], ulcerative colitis [39], inflammatory bowel disease [19,40], colorectal cancer [41], and *Clostridium difficile* infection [42] have been associated with disruptions in the gut microbiome. Interestingly, a wide spectrum of non-digestive diseases including diabetes [43], obesity [32], liver diseases [41], rheumatoid arthritis [44], and autism [45] have also been associated with changes in the microbiome as well.

Each unique section of the digestive tract has a distinct community [46,47]. pH [47], bile salt type [27], and oxygen content [48] all vary in different digestive organs, shaping the contents of the microbial communities. Commonly, stool samples are used to study the gut microbiome, because they are less invasive than collection methods using an endoscope or surgery, and have been documented as an adequate surrogate for the communities of the lower gastrointestinal tract [49].

Most gut microbiome studies have been restricted to analyzing the bacterial communities exclusively. This is due to the fact that the majority of microbiome cells are bacterial [50] and robust characterization methods for bacteria have been developed [49]. Several sets of primers targeting the bacterial 16S are widely used [51]. There is no single core bacterial community in the gut, but more recent analysis has revealed the presence of enterotypes: sets of specific bacteria that co-occur in individuals [52]. The initial enterotype study proposed three unique enterotypes [52], but others have postulated the existence of only two [53]. The two accepted enterotypes are dominated by *Prevotella* and *Bacteroides, respectively*. The more controversial third enterotype is postulated to be dominated by *Ruminococcus* [52]. All three enterotypes perform core microbiome functions including vitamin synthesis, energy regulation, and conditioning the host immune system [54]. Regardless of the enterotype, the gut bacterial community consists primarily of *Bacteroidetes* and *Firmicutes* [55]. *Proteobacteria* are common but a minority component

[33]. Several other phyla including *Actinobacteria* and *Fusobacteria* appear at very low levels as well [41].

Fungi

The fungi are a distinct group of micro- and macro-eukaryotes. Both fungi and animals are members of the opisthokonts, which are characterized by a single flagellum [56] and conserved insertions in elongation factor-1α and enolase, and supported by several single gene phylogenies [57]. Fungi have organelles and cell metabolism similar to those in animals, and the yeast *Saccharomyces cerevisiae* is often used as a model organism to study genetics and molecular biology for insights into human biology [57]. Nevertheless, fungi contain a rigid cell wall structure, which is not seen in animals [58]. Fungal cell walls consist of layers of mannan, β-glucan, and chitin in different configurations across fungal species and are difficult to lyse with traditional DNA extraction methods [59]. Many fungal species have variable ploidy. Many species have haploid, diploid, and polyploid states [60] and genders or mating types within the haploid state [61]. Also, some fungal species can be multinucleated [62,63]. Fungi have been observed in nearly all environments with which human come into contact. They can live in soil [64], water [65], and inside living organisms [66], and can become airborne [67].

Fungi and the Microbiome

Pathogenic fungi are a public health problem and have become increasingly so in the past several decades with increases in susceptible populations. Specifically, patients with diseases such as HIV/AIDS [68] and hematological cancers that weaken the immune system [69] are at elevated risk. Similarly, patients that have had solid organ [70,71] or bone marrow transplants [72], antibiotic exposure [73], antifungal exposure [73], received systemic corticosteroids [74], or used a catheter [73] are also at risk. Many fungi can cause invasive fungal infections such as *Candida*, *Aspergillus*, *Cryptococcus*, *Histoplasma*, *Mucor*, and *Coccidioides* spp., among others

[75,76,77]. Infections can be local and relatively minor, as seen in some cutaneous infections [78], but many are severe and widespread as in fungal pneumonia [79] or invasive systemic infection [80]. Fungal infections are particularly serious because they have a higher mortality rate, over a third of cases in some studies [81,82]. Recently, epidemiologists have seen the rise of strains of *Candida* [83], *Aspergillus* [84], and *Cryptococcus* spp. [85] that are resistant to commonly used azole drugs. These isolates are a growing health concern, especially because the populations of susceptible patients have been increasing [86].

Fungi also inhabit the human body as commensal microbes without causing disease. *Candida* spp. have been found in the gut of a majority of humans with culture based methods, and early sequencing methods indicated that more uncultured fungi are present [87]. The majority of fungal cells on the skin are *Malassezia* spp., but culture based evidence suggests that several other species inhabit the skin including *Debaryomyces* and *Cryptococcus* spp. [29]. Culture independent methods indicate that a large diversity of species occupy the oral cavity, including *Candida* and *Cladosporium* spp. [88]. Commensal fungi usually do not cause acute infection, but in some instances have been linked to disease severity. In patients with ulcerative colitis, the presence of *Candida* has been associated with increased disease severity in specific host genotypes [39]. Outgrowth and infection of commensal *Candida* commonly occurs when non-pathogenic yeast forms grow hyphae and invade local tissue. Differences between the host response to yeast and hyphal states are not well understood, but the host immune system recognizes the fungal cell wall as a pathogen-associated molecular pattern (PAMP) and has specific receptors, including Dectin-1 and Dectin-2, that detect and respond to fungi [89].

Protozoa

The Protozoa consist of a diverse paraphyletic group of single celled and multicellular organisms spanning the domain Eukarya. Protozoa contains many high level groups, including

rhizaria, amoebozoa, alveolata, stramenopiles, and excavates [57]. Protozoan organisms are highly diverse. While all protozoa contain the same basic eukaryotic cell structure, many species have evolved specialized organelles [90,91,92]. Determination of the protozoal and the eukaryotic phylogeny has proven difficult, and numerous arrangements of high taxonomic levels and rootings have been proposed in recent years [57,93,94].

Protozoa and the Microbiome

Many protozoan species are human pathogens and the cause of public health problems in developing and tropical nations [95]. Many Protozoan diseases such as leishmania [96], malaria [97], Chagas disease [98], and African sleeping sickness [99] are blood borne pathogens that are transferred through an insect vector and cause millions of deaths and hundreds of millions illnesses annually [100]. Other protozoan pathogens have an oral route of infection: *Giardia* [101], *Entamoeba* [102], *Blastocystis* [103], and *Balantidium* [104] and cause gastrointestinal disease.

Protozoans have also been documented to live in the gut microbiome without causing disease. Several species of amoeba [105] along with recognized pathogens like *Giardia* [106] and *Blastocystis* [107] have been found to live commensally in the gut. Nevertheless, it is generally believed that commensal protozoans are a minority community component. Commensal *Blastocystis* and fungi far outnumber other microeukaryotes [108]. Nevertheless, further studies are needed to understand the occurrence and role of non-fungal microeukaryotes in the gut microbiome in health and disease.

Archaea

The archaea are a unique domain of prokaryotic life more closely related to eukaryotes than to bacteria [109]. Even though members of the archaea were known and isolated in the

1930s [110], archaeal species were initially classified within bacteria [111] and were not recognized as belonging in a clade apart from bacteria until 1977, when Carl Woese proposed that archaea were a distinct lineage through phylogenetic analysis of 16S ribosomal (rDNA) sequences [112]. Currently, there are two main archaeal kingdoms consisting of the majority of known species: Euryarchaeota and Crenarchaeota [113]. Additional kingdoms have been proposed to accommodate fringe species such as Nanoarchaeota (consisting of the species *Nanoarchaeum equitans*) [114] and Korarchaeota [115,116] (consisting of *Korarchaeum cryptofilum*)[117]. Recent phylogenetic analysis of sequenced archaeal genomes has led to the proposal of a third main kingdom: Thaumarchaeota, consisting of mesophilic species formerly classified within Crenarchaeota [118]. More deep branching archaeal kingdoms have been proposed as recently as 2011 [119], and the taxonomic and phylogenetic structure of the archaea and its exact relation to the Eukarya are still debated [120].

Cellular analyses of archaea have revealed that the domain has a unique cell structure and distinct metabolic capabilities. For example, the cell wall architecture is diverse across archaeal clades but significantly different from the bacterial cell wall [121], often difficult to lyse [122], and believed to be widely resistant to the enzyme lysozyme [123]. Additionally, archaea have even evolved exclusive cell wall structures such as the cannulae and hami [121], and archaeal membranes consist of ether linked lipids instead of ester linked lipids [124]. Methanogenesis is exclusive to the archaea and requires a complicated cascade of enzymes [125]. Similarly, several archaeal species have been documented to use highly modified enzymatic pathways in central glycolysis reactions [126]. The archaea are truly a separate clade from bacteria, confirmed by structural and biochemical distinctiveness.

Archaeal Detection in Environmental Samples

Archaea have been found in a wide range of environments and commonly contain specific adaptations for their environment. Several species have been found in extremely hot and acidic environments [127,128]. These organisms have evolved heat and acid resistant enzymes [129], long, branched, and highly saturated membrane lipids [124], and proton pumps to keep the pH within the cell up to 5 orders of magnitude lower than the exterior environment [130]. Conversely, archaea have been found in psychrophilic environments below 0°C and are believed to be able to survive at lower temperatures [131]. These archaea adjust to cold temperatures by having a cell membrane consisting of less saturated lipids [124] and reducing the use of charged amino acids [132]. Archaea have also been found in hypersaline environments such as the Dead Sea [133] and the Great Salt Plains [134]; these archaeal species contain enzymes adapted for activity in saline environments [135]. In addition to extreme environments, archaea are commonly found in more mesophilic environments such as freshwater lakes [136], sea water [137], wetlands [138], and soil [139].

Archaea in the Gut Microbiome

In addition to residing in most environmental locales, archaea have been documented to live commensally in the microbiome. *Methanobrevibacter smithii* has been long considered to be the dominant archaeon in the human gut [50,140], living in a majority of humans [122,141]. *M. smithii* is believed to complete energy harvesting by absorbing CO_2 produced by other microbes and converting it to methane [142]. Furthermore, the presence of *M. smithii* is enriched in the *Ruminococcus*-containing enterotypes [52] . Other species within the *Methanobrevibacter* genus have been detected in non-human primates [143], ruminants [144], and termites [145]. Additionally, methanogens related to *Methanobrevibacter* have been found in primates [143], swine [146], and cockroaches [147].

Archaea in the gut are much less diverse than the bacterial inhabitants. Often *M. smithii* is the only archaeal species detected [140], and the *M. smithii* population usually exhibits limited genetic diversity. In some cases it appears almost clonal in metagenomic assemblies [50]. Several other archaeal species such as *Methanosphaeara stadtmanae* [122] and *Methanomassiliicoccus luminyensis* [148] have been detected in lower percentages in humans as well.

Classification of Amplicon Based Sequencing Data

The use of amplicon based sequencing creates datasets often containing similar sequences that need to be differentiated and binned by similarity into Operational Taxonomic Units (OTUs). Each OTU represents a species within the sequenced microbial community [149]. From each OTU a representative sequence is chosen. A representative sequence can be chosen by length, identity to other sequences, frequency within the OTU, or at random [150]. After a set of representative sequences is generated over sequencing data, the representative sequences are put through a classifier developed for the sequenced amplicon.

16S Classification

The most popular method to classify 16S bacterial and archaeal sequences is through naïve Bayesian k-mer classifiers. k-mer classifiers work by generating every possible k-length word in a query sequence and then comparing the frequency of each k-mer in the query to k-mer distributions in each of the genera represented in a reference database. Each possible k-mer has its own word specific prior based on its frequency of occurrence in the database, which is in turn used to calculate the probability of observing a k-mer in a given genus. The probability of a query belonging to a genus is calculated by multiplying together all the probabilities of observing each of its k-mers in the genus. This value is computed across all genera, with a naive prior. That is, no genus is deemed more or less likely to contain the query sequence prior to computation. The

genus with the highest probability is assigned to the query sequence [151]. Classifications to lower taxonomic levels are possible, but most current implementations classify to the genus level.

Different k-mer lengths have been used in different implementations. The RDP classifier uses 8-mers, because 8 and 9-mers had superior accuracy in initial tests compared to 6 and 7-mers, and 8-mers are more memory efficient [151]. Greengenes uses a 7-mer classifier, but allows for classifications down to the species level [152,153]. This method has been thoroughly vetted through "leave one out" testing of defined curated references and found to be robust [151]. Other classification methods relying on BLAST [150] and sequence placement within a reference phylogeny [154] have also been used for 16S taxonomic attribution, but k-mer classifiers are preferred because of their accuracy, speed, and lower susceptibility to misclassifications caused by errors within the reference database. Several large software packages have been published incorporating multiple classifiers and other statistical and phylogenetic tools for analysis of 16S sequences [150,154,155].

18S and ITS Classification

Classification of commonly used eukaryotic amplicons is much more difficult than classification of prokaryotic 16S sequences. The eukaryotic taxonomy is currently in a state of flux. In the past decade many new genera have been proposed and rearranged, many of which contain species found in the gut microbiome [156,157,158]. Furthermore, taxonomies have undergone major rearrangements up to the phylum level in fungi [159,160] and up to the kingdom level in protozoa [57,161]. Worse yet, until recently it was commonly accepted practice for a single fungal species to have different names based on its morphology and mating state [15]. Upon recent examination, up to 16 taxonomic names have been found to belong to a single species [162]. This redundant naming process was discontinued in 2011 [163], but many databases still contain outdated names.

As in prokaryotes, ribosomal genes are utilized for amplicon based sequencing, but unlike in prokaryotes, several different genes are commonly used. The 18S ribosomal subunit, the eukaryotic homolog to the prokaryotic 16S rRNA gene, is popular and there are large curated 18S databases available [164], but the 18S gene is too conserved for consistent differentiation below the family level. Like the 18S, well curated resources exist for the 28S rRNA gene. The 28S rRNA is more variable than the 18S gene, but the Internal Transcribed Spacers (ITS) 1 and 2 offer the best species level resolution and identification, particularly for fungi [15]. Several groups have used k-mer classifiers for fungal 18S and 28S rRNA genes [150,165]; however databases are still small compared to bacterial references, and the ITS1 and ITS2 genes are too variable for k-mer classifiers. Also, phylogenetic tree insertion methods using have been used on the 18S and 28S genes [154], but they are vulnerable to misclassified sequences in the reference set and similarly are not feasible for the ITS1 and ITS2 genes. Subsequently, many researchers have turned to using manually curated BLAST results to classify ITS sequences [39,88].

Antibiotics

Human use of antibiotics has a long history. Evidence of antibiotic use dates back to antiquity when ancient Nubians, Greeks, and Chinese used antibiotic laced compounds therapeutically [166,167,168]. More recently, antibiotics were studied in several laboratories in the late 19th century [169] and were characterized and adapted for clinical use by Alexander Fleming in 1929. Antibiotics were immediately hailed as a wonder drug for previously incurable or untreatable bacterial infections [170]. Antibiotics are in use today across the globe as treatments for a wide range of infections [171].

Antibiotic drugs work by exploiting differences in cellular structure or metabolism between humans and an infecting pathogen. Many bactericidal antibiotics such as β-lactams [172] target bacterial cell wall synthesis and maintenance. Others, like aminoglycosides [173] and

tetracyclines [174], disrupt bacterial protein synthesis. Sulfonamides specifically target folate synthesis [175]. Like antibacterial agents, antifungal agents exploit differences in fungal and human cellular biology. However, fungal cells are much more similar to human cells than bacterial cells, and there are fewer workable drug targets. Azoles target lanosterol 14 α-demethylase, which is a necessary enzyme in cholesterol synthesis. Polyene antifungals create pores in fungal cell membranes. Development of new classes of antibacterial and antifungal agents is ongoing [176].

Side effects of antibiotics can range from relatively mild gastrointestinal dysfunction to serious hematological, cardiac, pulmonary, hepatic, nephritic, and neurological problems[177]. Furthermore, repeated exposures to antimicrobial agents have led strains of many important pathogens, such as *Escherichia coli* [178], *Clostridium difficile* [179], and *Mycobacterium tuberculosis* [180], to become drug resistant and difficult to treat.

Ingested and intravenous antibiotics have been demonstrated to dramatically affect the composition of the gut microbiome community [34]. Treatment for a bacterial infection or prophylaxis does not just quash the offending pathogen. Dramatic microbiome and phenotypic changes have been observed in mice [181] and swine [182] treated with even subclinical doses. Further study is required in order to determine the long term effects of antibiotic use on human gut health. Antibiotics have been proposed as a possible mechanism to treat disease related to the gut microbiome, because they are clinically tested and proven to alter the gut community [34]. Several narrow spectrum antibiotics have been proposed as treatments for bacteria like *C. difficile* in order to avoid altering surrounding flora as a means of preserving commensal bacteria and health of the host [183].

Motivation and Thesis Outline

The gut microbiome plays a central role in human health. However, at the start of this thesis project there existed methods mainly focused on characterizing the bacterial gut flora. It is known that the bacteria in the gut interact biochemically with archaea and microeukaryotes; therefore, studying bacteria in isolation provides an incomplete view of the gut community. One of the primary goals of this work was to pioneer methods to characterize these important communities. In Chapter 2, I present a sample-to-analysis pipeline to sequence fungal and microeukaryotic communities from human stool. This work analyzes the effectiveness of established primers and novel primers, developed by Greg Peterfreund in the Bushman Laboratory, and describes a classifier that I developed for classifying reads generated from those primers. In Chapter 3, I introduce a method for successfully amplifying the archaeal 16S gene from stool samples, while avoiding non-target DNA. Finally, in Chapter 4, I study the longitudinal effects of antibiotic treatment in the gut microbiome using a mouse model. This study analyzes both the bacterial and microeukaryotic communities during and after antibiotic treatment. Finally, Chapter 5 describes the impact of this work and the new possibilities that it enables.

References

1. MacNeal WJ, Latzer LL, Kerr JE (1909) The Fecal Bacteria of Healthy Men. Part I. Introduction and Direct Quantitative Observations. The Journal of Infectious Diseases 6: 123-169.

2. Simonds JP (1915) Classification of the Bacillus Welchii Group of Bacteria. The Journal of Infectious Diseases 16: 31-34.

3. Kell DB, Kaprelyants AS, Weichart DH, Harwood CR, Barer MR (1998) Viability and activity in readily culturable bacteria: a review and discussion of the practical issues. Antonie van Leeuwenhoek 73: 169-187.

4. Davis KE, Joseph SJ, Janssen PH (2005) Effects of growth medium, inoculum size, and incubation time on culturability and isolation of soil bacteria. Applied and environmental microbiology 71: 826-834.

5. Hill G, Mitkowski N, Aldrich-Wolfe L, Emele L, Jurkonie D, et al. (2000) Methods for assessing the composition and diversity of soil microbial communities. Applied Soil Ecology 15: 25-36.

6. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74: 5463-5467.

7. Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome research 19: 1141-1152.

8. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

9. Bennett S (2004) Solexa Ltd. Pharmacogenomics 5: 433-438.

10. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, et al. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome research 19: 1527-1541.

11. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009) Metagenomic pyrosequencing and microbial identification. Clin Chem 55: 856-866.

12. Janda JM, Abbott SL (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of clinical microbiology 45: 2761-2764.

13. Garcia-Martinez J, Acinas SG, Anton AI, Rodriguez-Valera F (1999) Use of the 16S--23S ribosomal genes spacer region in studies of prokaryotic diversity. Journal of microbiological methods 36: 55-64.

14. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, et al. (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Applied and environmental microbiology 73: 278-288.

15. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences 109: 6241-6246.

16. Lipscomb DL, Farris JS, Källersjö M, Tehler A (1998) Support, ribosomal sequences and the phylogeny of the eukaryotes. Cladistics 14: 303-338.

17. Krabberød AK, Bråte J, Dolven JK, Ose RF, Klaveness D, et al. (2011) Radiolaria divided into Polycystina and Spasmaria in combined 18S and 28S rDNA phylogeny. PloS one 6: e23526.

18. Felleisen R (1997) Comparative sequence analysis of 5· 8S rRNA genes and internal transcribed spacer (ITS) regions of trichomonadid protozoa. Parasitology 115: 111-119.

19. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.

20. Schloss P, Handelsman J (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome biology 6: 229.

21. Simpson JT, Durbin R (2012) Efficient de novo assembly of large genomes using compressed data structures. Genome research 22: 549-556.

22. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, et al. (2009) Community-wide analysis of microbial genome sequence signatures. Genome biology 10: R85.

23. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, et al. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nature methods 9: 811-814.

24. Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, et al. (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome biology 13: R23.

25. McNulty NP, Yatsunenko T, Hsiao A, Faith JJ, Muegge BD, et al. (2011) The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. Science translational medicine 3: 106ra106.

26. Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, et al. (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proc Natl Acad Sci U S A 107: 7503-7508.

27. Savage DC (1977) Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol 31: 107-133.

28. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, et al. (2010) The human oral microbiome. Journal of bacteriology 192: 5002-5017.

29. Grice EA, Segre JA (2011) The skin microbiome. Nature Reviews Microbiology 9: 244-253.

30. Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, et al. (2010) Disordered microbial communities in the upper respiratory tract of cigarette smokers. PLoS One 5: e15216.

31. Burton JP, Cadieux PA, Reid G (2003) Improved understanding of the bacterial vaginal microbiota of women before and after probiotic instillation. Applied and environmental microbiology 69: 97-101.

32. Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI (2011) Human nutrition, the gut microbiome and the immune system. Nature 474: 327-336.

33. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, et al. (2009) Bacterial Community Variation in Human Body Habitats Across Space and Time. Science (New York, NY).

34. Pflughoeft KJ, Versalovic J (2012) Human microbiome in health and disease. Annual Review of Pathology: Mechanisms of Disease 7: 99-122.

35. Lauber CL, Hamady M, Knight R, Fierer N (2009) Soil pH as a predictor of soil bacterial community structure at the continental scale: a pyrosequencing-based assessment. Applied and environmental microbiology.

36. Costello EK, Stagaman K, Dethlefsen L, Bohannan BJ, Relman DA (2012) The application of ecological theory toward an understanding of the human microbiome. Science 336: 1255-1262.

37. Furness JB, Kunze WA, Clerc N (1999) II. The intestine as a sensory organ: neural, endocrine, and immune responses. American Journal of Physiology-Gastrointestinal and Liver Physiology 277: G922-G928.

38. Quince C, Lundin EE, Andreasson AN, Greco D, Rafter J, et al. (2013) The impact of Crohn's disease genes on healthy human gut microbiota: a pilot study. Gut.

39. Iliev ID, Funari VA, Taylor KD, Nguyen Q, Reyes CN, et al. (2012) Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. Science 336: 1314-1317.

40. Bushman FD, Lewis JD, Wu GD (2013) Diet, the Human Gut Microbiota, and IBD. Anaerobe.

41. Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. Nat Rev Genet 13: 260-270.

42. Khoruts A, Dicksved J, Jansson JK, Sadowsky MJ (2010) Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent Clostridium difficile-associated diarrhea. Journal of clinical gastroenterology 44: 354.

43. Boerner BP, Sarvetnick NE (2011) Type 1 diabetes: role of intestinal microbiome in humans and mice. Annals of the New York Academy of Sciences 1243: 103-118.

44. Scher JU, Abramson SB (2011) The microbiome and rheumatoid arthritis. Nature Reviews Rheumatology 7: 569-578.

45. Sajdel-Sulkowska EM, Zabielski R (2013) Gut Microbiome and Brain-Gut Axis in Autism—Aberrant Development of Gut-Brain Communication and Reward Circuitry.

46. Derrien M, van Passel MW, van de Bovenkamp JH, Schipper R, de Vos W, et al. (2010) Mucin-bacterial interactions in the human oral cavity and digestive tract. Gut Microbes 1: 254-268.

47. Walter J, Ley R (2011) The human gut microbiome: ecology and recent evolutionary changes. Annual review of microbiology 65: 411-429.

48. Albenberg L, Judge C, Esipova T, Grunberg S, Chen J, et al. (2012) An Oxygen Equilibrium at the Host-Microbial Interface Determined by Phosphorescent Nanoprobe Technology: P-257 YI. Inflammatory Bowel Diseases 18: S112-S113.

49. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, et al. (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiol 10: 206.

50. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science (New York, NY) 312: 1355-1359.

51. Baker GC, Smith JJ, Cowan DA (2003) Review and re-analysis of domain-specific 16S primers. Journal of microbiological methods 55: 541-555.

52. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature 473: 174-180.

53. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334: 105-108.

54. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-Bacterial Mutualism in the Human Intestine. Science 307: 1915-1920.

55. Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, et al. (2009) Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. Proc Natl Acad Sci U S A 106: 5859-5864.

56. LabyrinthulidsOpalinids B, Raphidophyceae E (2011) A. Opisthokonts. Evolution of Primary Producers in the Sea 2: 77.

57. Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, et al. (2005) The tree of eukaryotes. Trends Ecol Evol 20: 670-676.

58. Bartnicki-Garcia S (1968) Cell wall chemistry, morphogenesis, and taxonomy of fungi. Annual Reviews in Microbiology 22: 87-108.

59. Latgé JP (2007) The cell wall: a carbohydrate armour for the fungal cell. Molecular microbiology 66: 279-290.

60. Morrow CA, Fraser JA (2013) Ploidy variation as an adaptive mechanism in human pathogenic fungi. Semin Cell Dev Biol.

61. Kronstad J, Staben C (1997) Mating type in filamentous fungi. Annual review of genetics 31: 245-276.

62. Gerstenberger JP, Occhipinti P, Gladfelter AS (2012) Heterogeneity in mitochondrial morphology and membrane potential is independent of the nuclear division cycle in multinucleate fungal cells. Eukaryotic cell 11: 353-367.

63. Markina-Iñarrairaegui A, Etxebeste O, Herrero-García E, Araújo-Bazán L, Fernández-Martínez J, et al. (2011) Nuclear transporters in a multinucleated organism: functional and localization analyses in Aspergillus nidulans. Molecular biology of the cell 22: 3874-3886.

64. Miransari M (2011) Interactions between arbuscular mycorrhizal fungi and soil bacteria. Applied microbiology and biotechnology 89: 917-930.

65. Hageskal G, Lima N, Skaar I (2009) The study of fungi in drinking water. Mycological Research 113: 165-172.

66. Savage AE, Zamudio KR (2011) MHC genotypes associate with resistance to a frog-killing fungus. Proceedings of the National Academy of Sciences 108: 16705-16710.

67. Blum G, Eschertzhuber S, Auberger J, Ulmer H, Geltner C, et al. (2012) Airborne fungus exposure prior to hospitalisation as risk factor for mould infections in immunocompromised patients. Mycoses 55: 237-243.

68. Ampel NM (1996) Emerging disease issues and fungal pathogens associated with HIV infection. Emerging infectious diseases 2: 109-116.

69. Chamilos G, Luna M, Lewis RE, Bodey GP, Chemaly R, et al. (2006) Invasive fungal infections in patients with hematologic malignancies in a tertiary care cancer center: an autopsy study over a 15-year period (1989-2003). Haematologica 91: 986-989.

70. Gallis HA, Berman RA, Cate TR, Hamilton JD, Gunnells JC, et al. (1975) Fungal infection following renal transplantation. Archives of internal medicine 135: 1163-1172.

71. Koo S, Kubiak DW, Issa NC, Dietzek A, Boukedes S, et al. (2012) A targeted peritransplant antifungal strategy for the prevention of invasive fungal disease after lung transplantation: a sequential cohort analysis. Transplantation 94: 281-286.

72. Trenschel R, Peceny R, Runde V, Elmaagacli A, Dermoumi H, et al. (2000) Fungal colonization and invasive fungal infections following allogeneic BMT using metronidazole, ciprofloxacin and fluconazole or ciprofloxacin and fluconazole as intestinal decontamination. Bone marrow transplantation 26: 993-997.

73. Ben-Ami R, Olshtain-Pops K, Krieger M, Oren I, Bishara J, et al. (2012) Antibiotic exposure as a risk factor for fluconazole-resistant Candida bloodstream infection. Antimicrobial agents and chemotherapy 56: 2518-2523.

74. Lionakis MS, Kontoyiannis DP (2003) Glucocorticoids and invasive fungal infections. Lancet 362: 1828-1838.

75. Wajszczuk CP, Dummer JS, Ho M, Van Thiel DH, Starzl TE, et al. (1985) Fungal infections in liver transplant recipients. Transplantation 40: 347.

76. Mandell W, Goldberg DM, Neu HC (1986) Histoplasmosis in patients with the acquired immune deficiency syndrome. The American journal of medicine 81: 974-978.

77. Roberts CJ (1984) Coccidioidomycosis in acquired immune deficiency syndrome: depressed humoral as well as cellular immunity. The American journal of medicine 76: 734-736.

78. Leyden JJ, Marples RR (1973) Ecologic principles and antibiotic therapy in chronic dermatoses. Archives of dermatology 107: 208-211.

79. Libshitz H, Pagani J (1981) Aspergillosis and mucormycosis: two types of opportunistic fungal pneumonia. Radiology 140: 301-306.

80. Gullo A (2009) Invasive fungal infections. Drugs 69: 65-73.

81. Viscoli C, Girmenia C, Marinus A, Collette L, Martino P, et al. (1999) Candidemia in cancer patients: a prospective, multicenter surveillance study by the Invasive Fungal Infection Group (IFIG) of the European Organization for Research and Treatment of Cancer (EORTC). Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 28: 1071-1079.

82. Pagano L, Caira M, Candoni A, Offidani M, Fianchi L, et al. (2006) The epidemiology of fungal infections in patients with hematologic malignancies: the SEIFEM-2004 study. Haematologica 91: 1068-1075.

83. Sasse C, Dunkel N, Schafer T, Schneider S, Dierolf F, et al. (2012) The stepwise acquisition of fluconazole resistance mutations causes a gradual loss of fitness in Candida albicans. Molecular microbiology 86: 539-556.

84. Burgel PR, Baixench MT, Amsellem M, Audureau E, Chapron J, et al. (2012) High prevalence of azole-resistant Aspergillus fumigatus in adults with cystic fibrosis exposed to itraconazole. Antimicrobial agents and chemotherapy 56: 869-874.

85. Perfect JR, Cox GM (1999) Drug resistance in Cryptococcus neoformans. Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy 2: 259-269.

86. Chen KY, Ko SC, Hsueh PR, Luh KT, Yang PC (2001) Pulmonary fungal infection: emphasis on microbiological spectra, patient outcome, and prognostic factors. Chest 120: 177-184.

87. Scanlan PD, Shanahan F, O'Mahony C, Marchesi JR (2006) Culture-independent analyses of temporal variation of the dominant fecal microbiota and targeted bacterial subgroups in Crohn's disease. J Clin Microbiol 44: 3980-3988.

88. Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, et al. (2010) Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. PLoS Pathog 6: e1000713.

89. van der Meer JW, van de Veerdonk FL, Joosten LA, Kullberg B-J, Netea MG (2010) Severe *Candida* spp. infections: new insights into natural immunity. International Journal of Antimicrobial Agents 36: S58-S62.

90. Tovar J, Leon-Avilla G, Sanchez LB, Sutak R, Tachezy J, et al. (2003) Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation. Nature 426: 172-176.

91. Dyall SD, Johnson PJ (2000) Origins of hydrogenosomes and mitochondria: evolution and organelle biogenesis. Cur Op in Microbio 3: 404-411.

92. Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, et al. (2003) Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum. Science 299: 705-708.

93. Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morrison HG, et al. (2010) Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Systematic Biology 59: 518-533.

94. Rodríguez-Ezpeleta N, Brinkmann H, Burger G, Roger AJ, Gray MW, et al. (2007) Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Current Biology 17: 1420-1425.

95. Lonnen J, Kilvington S, Kehoe S, Al-Touati F, McGuigan K (2005) Solar and photocatalytic disinfection of protozoan, fungal and bacterial microbes in drinking water. Water research 39: 877-883.

96. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, Leishmania major. Science 309: 436-442.

97. Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI (2005) The global distribution of clinical episodes of Plasmodium falciparum malaria. Nature 434: 214-217.

98. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, et al. (2005) The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. Science 309: 409-415.

99. Kennedy PG (2008) The continuing problem of human African trypanosomiasis (sleeping sickness). Annals of neurology 64: 116-126.

100. Rafael ME, Taylor T, Magill A, Lim Y-W, Girosi F, et al. (2006) Reducing the burden of childhood malaria in Africa: the role of improved. Nature 444: 39-48.

101. Carmena D, Aguinagalde X, Zigorraga C, Fernández-Crespo J, Ocio J (2007) Presence of Giardia cysts and Cryptosporidium oocysts in drinking water supplies in northern Spain. Journal of applied microbiology 102: 619-629.

102. Benetton M, Gonçalves A, Meneghini M, Silva E, Carneiro M (2005) Risk factors for infection by the Entamoeba histolytica *E. dispar* complex: An epidemiological study conducted in outpatient clinics in the city of Manaus, Amazon Region, Brazil. Transactions of the Royal Society of Tropical Medicine and Hygiene 99: 532-540.

103. Tan KS (2008) New insights on classification, identification, and clinical relevance of Blastocystis spp. Clinical microbiology reviews 21: 639-665.

104. Schuster FL, Ramirez-Avila L (2008) Current world status of Balantidium coli. Clinical microbiology reviews 21: 626-638.

105. Ortega YR (2006) Amoeba and Ciliates. Foodborne Parasites: Springer. pp. 1-14.

106. Prado M, Cairncross S, Strina A, Barreto M, Oliveira-Assis A, et al. (2005) Asymptomatic giardiasis and growth in young children; a longitudinal study in Salvador, Brazil. Parasitology-Cambridge 131: 51-56.

107. Boorom KF, Smith H, Nimri L, Viscogliosi E, Spanakos G, et al. (2008) Oh my aching gut: irritable bowel syndrome, Blastocystis, and asymptomatic infection. Parasit Vectors 1: 40.

108. Parfrey LW, Walters WA, Knight R (2011) Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. Frontiers in microbiology 2.

109. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences of the United States of America 87: 4576-4579.

110. Barker HA Studies upon the methane-producing bacteria.

111. Barker HA, Doudoroff M (1946) Bacterial metabolism. Annual review of biochemistry 15: 475-504.

112. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America 74: 5088-5090.

113. Cavicchioli R (2011) Archaea--timeline of the third domain. Nature reviews Microbiology 9: 51-61.

114. Huber H, Hohn MJ, Stetter KO, Rachel R (2003) The phylum Nanoarchaeota: present knowledge and future perspectives of a unique form of life. Research in microbiology 154: 165-171.

115. Auchtung TA, Takacs-Vesbach CD, Cavanaugh CM (2006) 16S rRNA phylogenetic investigation of the candidate division "Korarchaeota". Applied and environmental microbiology 72: 5077-5082.

116. Barns SM, Delwiche CF, Palmer JD, Pace NR (1996) Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proceedings of the National Academy of Sciences of the United States of America 93: 9188-9193.

117. Elkins JG, Podar M, Graham DE, Makarova KS, Wolf Y, et al. (2008) A korarchaeal genome reveals insights into the evolution of the Archaea. Proceedings of the National Academy of Sciences of the United States of America 105: 8102-8107.

118. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nature reviews Microbiology 6: 245-252.

119. Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, et al. (2011) Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic acids research 39: 3204-3223.

120. Williams TA, Foster PG, Nye TM, Cox CJ, Embley TM (2012) A congruent phylogenomic signal places eukaryotes within the Archaea. Proceedings Biological sciences / The Royal Society 279: 4870-4879.

121. Albers SV, Meyer BH (2011) The archaeal cell envelope. Nature reviews Microbiology 9: 414-426.

122. Dridi B, Henry M, El Khechine A, Raoult D, Drancourt M (2009) High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. PLoS ONE 4: e7063.

123. Ishii K, Mussmann M, MacGregor BJ, Amann R (2004) An improved fluorescence in situ hybridization protocol for the identification of bacteria and archaea in marine sediments. FEMS Microbiology Ecology 50: 203-213.

124. van de Vossenberg JL, Driessen AJ, Konings WN (1998) The essence of being extremophilic: the role of the unique archaeal membrane lipids. Extremophiles : life under extreme conditions 2: 163-170.

125. Bapteste E, Brochier C, Boucher Y (2005) Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. Archaea 1: 353-363.

126. Siebers B, Schonheit P (2005) Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. Current opinion in microbiology 8: 695-705.

127. Brock TD, Brock KM, Belly RT, Weiss RL (1972) Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. Archiv fur Mikrobiologie 84: 54-68.

128. Itoh T, Yoshikawa N, Takashina T (2007) Thermogymnomonas acidicola gen. nov., sp. nov., a novel thermoacidophilic, cell wall-less archaeon in the order Thermoplasmatales, isolated from a solfataric soil in Hakone, Japan. International journal of systematic and evolutionary microbiology 57: 2557-2561.

129. Auernik KS, Cooper CR, Kelly RM (2008) Life in hot acid: pathway analyses in extremely thermoacidophilic archaea. Current opinion in biotechnology 19: 445-453.

130. Gleissner M, Kaiser U, Antonopoulos E, Schafer G (1997) The archaeal SoxABCD complex is a proton pump in Sulfolobus acidocaldarius. The Journal of biological chemistry 272: 8417-8426.

131. Cavicchioli R (2006) Cold-adapted archaea. Nature reviews Microbiology 4: 331-343.

132. Saunders NF, Thomas T, Curmi PM, Mattick JS, Kuczek E, et al. (2003) Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea Methanogenium frigidum and Methanococcoides burtonii. Genome research 13: 1580-1588.

133. Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, et al. (2004) Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea. Genome research 14: 2221-2234.

134. Caton TM, Witte LR, Ngyuen HD, Buchheim JA, Buchheim MA, et al. (2004) Halotolerant aerobic heterotrophic bacteria from the Great Salt Plains of Oklahoma. Microbial ecology 48: 449-462.

135. Madern D, Ebel C, Zaccai G (2000) Halophilic adaptation of enzymes. Extremophiles : life under extreme conditions 4: 91-98.

136. Abreu C, Jurgens G, De Marco P, Saano A, Bordalo AA (2001) Crenarchaeota and Euryarchaeota in temperate estuarine sediments. Journal of applied microbiology 90: 713-718.

137. Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB (2005) Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proceedings of the National Academy of Sciences of the United States of America 102: 14683-14688.

138. Kanokratana P, Uengwetwanit T, Rattanachomsri U, Bunterngsook B, Nimchua T, et al. (2011) Insights into the phylogeny and metabolic potential of a primary tropical peat swamp forest microbial community by metagenomic analysis. Microbial ecology 61: 518-528.

139. Tourna M, Stieglmeier M, Spang A, Konneke M, Schintlmeister A, et al. (2011) Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. Proceedings of the National Academy of Sciences of the United States of America 108: 8420-8425.

140. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635-1638.

141. Million M, Maraninchi M, Henry M, Armougom F, Richet H, et al. (2012) Obesity-associated gut microbiota is enriched in Lactobacillus reuteri and depleted in Bifidobacterium animalis and Methanobrevibacter smithii. International journal of obesity 36: 817-825.

142. Miller TL, Wolin MJ, Conway de Macario E, Macario AJ (1982) Isolation of Methanobrevibacter smithii from human feces. Applied and environmental microbiology 43: 227-232.

143. Facey HV, Northwood KS, Wright AD (2012) Molecular diversity of methanogens in fecal samples from captive Sumatran orangutans (Pongo abelii). American journal of primatology 74: 408-413.

144. Leahy SC, Kelly WJ, Altermann E, Ronimus RS, Yeoman CJ, et al. (2010) The genome sequence of the rumen methanogen Methanobrevibacter ruminantium reveals new possibilities for controlling ruminant methane emissions. PloS one 5: e8926.

145. Leadbetter JR, Breznak JA (1996) Physiological ecology of Methanobrevibacter cuticularis sp. nov. and Methanobrevibacter curvatus sp. nov., isolated from the hindgut of the termite Reticulitermes flavipes. Applied and environmental microbiology 62: 3620-3631.

146. Luo YH, Su Y, Wright AD, Zhang LL, Smidt H, et al. (2012) Lean breed Landrace pigs harbor fecal methanogens at higher diversity and density than obese breed Erhualian pigs. Archaea 2012: 605289.

147. Sprenger WW, van Belzen MC, Rosenberg J, Hackstein JH, Keltjens JT (2000) Methanomicrococcus blatticola gen. nov., sp. nov., a methanol- and methylamine-reducing methanogen from the hindgut of the cockroach Periplaneta americana. International journal of systematic and evolutionary microbiology 50 Pt 6: 1989-1999.

148. Dridi B, Henry M, Richet H, Raoult D, Drancourt M (2012) Age-related prevalence of Methanomassiliicoccus luminyensis in the human gut microbiome. APMIS : acta pathologica, microbiologica, et immunologica Scandinavica 120: 773-777.

149. Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6: e1000667.

150. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335-336.

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

152. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, et al. (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME journal 6: 610-618.

153. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology 72: 5069-5072.

154. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: a software environment for sequence data. Nucleic acids research 32: 1363-1371.

155. Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, et al. (2011) CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC bioinformatics 12: 356.

156. Kurtzman CP, Robnett CJ, Basehoar-Powers E (2008) Phylogenetic relationships among species of Pichia, Issatchenkia and Williopsis determined from multigene sequence analysis, and the proposal of Barnettozyma gen. nov., Lindnera gen. nov. and Wickerhamomyces gen. nov. FEMS yeast research 8: 939-954.

157. Kurtzman CP (2003) Phylogenetic circumscription of Saccharomyces, Kluyveromyces and other members of the Saccharomycetaceae, and the proposal of the new genera Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma and Zygotorulaspora. FEMS yeast research 4: 233-245.

158. Minter D (2009) Cyberlindnera, a replacement name for Lindnera Kurtzman et al., nom. illegit. Mycotaxon 110: 473-476.

159. Schüβler A, Schwarzott D, Walker C (2001) A new fungal phylum, the *Glomeromycota*: phylogeny and evolution. Mycological Research 105: 1413-1421.

160. James TY, Letcher PM, Longcore JE, Mozley-Standridge SE, Porter D, et al. (2006) A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota). Mycologia 98: 860-871.

161. Cavalier-Smith T (2004) Only six kingdoms of life. Proceedings Biological sciences / The Royal Society 271: 1251-1262.

162. Martini AV, Kurtzman CP (1985) Deoxyribonucleic acid relatedness among species of the genus Saccharomyces sensu stricto. International journal of systematic bacteriology 35: 508-511.

163. Hawksworth DL (2011) A new dawn for the naming of fungi: impacts of decisions made in Melbourne in July 2011 on the future publication and regulation of fungal names. IMA Fungus: The Global Mycological Journal 2: 155.

164. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic acids research 35: 7188-7196.

165. Liu K-L, Porras-Alfaro A, Kuske CR, Eichorst SA, Xie G (2012) Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. Applied and environmental microbiology 78: 1523-1533.

166. Bassett EJ, Keith MS, Armelagos GJ, Martin DL, Villanueva AR. Tetracycline-labeled human bone from ancient Sudanese Nubia (AD 350); 1980. AAAS.

167. Conly JM, Johnston LB (2004) Coming full circle: From antibiotics to probiotics and prebiotics. Can J Infect Dis Med Microbiol 15: 161-163.

168. Nelson ML, Dinardo A, Hochberg J, Armelagos GJ (2010) Brief communication: Mass spectroscopic characterization of tetracycline in the skeletal remains of an ancient population from Sudanese Nubia 350-550 CE. Am J Phys Anthropol 143: 151-154.

169. Coppola A, Viggiani E (1980) Selected items from the history of pathology -- Vincenzo Tiberio (1896-1915). Am J Pathol 101: 30.

170. Sykes R (2001) Penicillin: from discovery to product. Bull World Health Organ 79: 778-779.

171. Rossignoli A, Clavenna A, Bonati M (2007) Antibiotic prescription and prevalence rate in the outpatient paediatric population: analysis of surveys published during 2000–2005. European journal of clinical pharmacology 63: 1099-1106.

172. Fisher JF, Meroueh SO, Mobashery S (2005) Bacterial resistance to beta-lactam antibiotics: compelling opportunism, compelling opportunity. Chem Rev 105: 395-424.

173. Mingeot-Leclercq M-P, Glupczynski Y, Tulkens PM (1999) Aminoglycosides: activity and resistance. Antimicrobial agents and chemotherapy 43: 727-737.

174. Schnappinger D, Hillen W (1996) Tetracyclines: antibiotic action, uptake, and resistance mechanisms. Archives of microbiology 165: 359-369.

175. Tenover FC (2006) Mechanisms of antimicrobial resistance in bacteria. The American journal of medicine 119: S3-S10.

176. Groll AH, De Lucca AJ, Walsh TJ (1998) Emerging targets for the development of novel antifungal therapeutics. Trends in microbiology 6: 117-124.

177. Cunha BA (2001) Antibiotic side effects. Med Clin North Am 85: 149-185.

178. Johnson JR, Sannes MR, Croy C, Johnston B, Clabots C, et al. (2007) Antimicrobial drug–resistant Escherichia coli from humans and poultry products, Minnesota and Wisconsin, 2002–2004. Emerging infectious diseases 13: 838.

179. Sebaihia M (2006) The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nature Genetics 38: 779-786.

180. Amaral L, Boeree MJ, Gillespie SH, Udwadia ZF, van Soolingen D (2010) Thioridazine cures extensively drug-resistant tuberculosis (XDR-TB) and the need for global trials is now! International Journal of Antimicrobial Agents 35: 524-526.

181. Cho I, Yamanishi S, Cox L, Methe BA, Zavadil J, et al. (2012) Antibiotics in early life alter the murine colonic microbiome and adiposity. Nature 488: 621-626.

182. Looft T, Johnson TA, Allen HK, Bayles DO, Alt DP, et al. (2012) In-feed antibiotic effects on the swine intestinal microbiome. Proceedings of the National Academy of Sciences 109: 1691-1696.

183. Lancaster JW, Matthews SJ (2012) Fidaxomicin: the newest addition to the armamentarium against Clostridium difficile infections. Clin Ther 34: 1-13.

CHAPTER 2: A Sample-to-Analysis Pipeline for Analysis of Microeukaryotes in Human Stool

The contents of this chapter have been published in:

> Dollive S, Peterfreund GL, Sherrill-Mix S, Bittinger K, Sinha R, Hoffmann C, Nabel CS, Hill DA, Artis D, Bachman MA, Custers-Allen R, Grunberg S, Wu GD, Lewis JD, Bushman FD. (2012). A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. Genome Biol 13: R60.

2.1 Abstract

Here we present a pipeline for analysis of deep sequencing data on single cell eukaryotes. We designed a new 18S rRNA gene specific PCR primer set and compared a published rRNA gene internal transcribed spacer (ITS) gene primer set. Amplicons were tested against 24 specimens from defined eukaryotes and eight well-characterized human stool samples. A software pipeline (https://sourceforge.net/projects/brocc/) was developed for taxonomic attribution, validated against simulated data, and tested on pyrosequence data. This study provides a well-characterized tool kit for sequence-based enumeration of eukaryotic organisms in human microbiome samples.

2.2 Introduction

Many microbiome studies are limited to characterizing the bacterial components, because bacteria are the largest microbial component of the microbiome, changes in the bacterial community have been associated with many diseases [1], and robust methods to characterize bacteria have been vetted and standardized [2]. Nevertheless, single cell eukaryotes form an important part of microbiome communities, but enumerating community membership and proportions in complex mixtures remains challenging. Advances in sequencing technology and bioinformatics have made possible several strategies. Shotgun metagenomics, in which all DNA from a sample is sequenced, can yield data on the types of organisms and genes present in a

mixed community. However, in many types of microbiome samples, eukaryotic microbes are a minor component, so shotgun metagenomics can be inefficient and expensive for their identification. Target gene sequencing can yield detailed information on community membership efficiently, as with the 16S rRNA gene amplicons widely used for profiling bacterial communities. However, there are no universally conserved regions in eukaryotic genomes analogous to those in the 16S rRNA locus of bacteria that yield similarly low level classifications. For microbiome samples from the digestive system, the potential masking effects of food DNA provides another complication, and for many sample types host DNA can also interfere.

Many diseases are mediated by infections of single cell eukaryotes [3,4,5], including infections of the gut [6], skin [7], urogenital tract [8], and pulmonary system [9]. In some cases infections have been associated with alteration of the normal microbiome [10], as in oral thrush [11] and aspergillosis [12], while others are apparently caused by invasion by a single eukaryotic pathogen such as *Mucor* [12] or *Giardia* [13]. Thus better understanding of the dynamics of eukaryotic components of microbiome communities will help in understanding and treating many of these infections.

Eukaryotic rRNA genes and their associated transcribed spacers have been used as marker genes [14,15,16,17], though target amplicons are not fully universal. In eukaryotes, the 18S, 5.8S, and 28S ribosomal subunits are encoded in a single locus separated by the first and second internal transcribed spacers (ITS). The ITS RNAs are degraded shortly after transcription and are not incorporated into the ribosome [18], thus ITS RNAs are less conserved than the 18S and 28S RNAs. Previously developed eukaryotic rRNA gene amplicons can query these regions, but most have not been designed or vetted for use specifically in human microbiome studies.

Here we describe a pipeline based on rRNA gene amplicons for analysis of eukaryotes of the human microbiome by deep sequencing. Sequencing 18S rRNA genes could be confounded by the potentially more abundant rRNA gene sequences from the mammalian host or, in samples from the gastrointestinal tract, from food. We thus designed an 18S rRNA gene amplicon that avoids mammalian and plant sequences, and also compared a published ITS1 amplicon targeting fungi [16]. We developed a flexible software pipeline (BROCC) for attributing sequences that was tailored for use with the complex and sometimes inconsistent taxonomic assignments characteristic of single cell eukaryotes. Because some fungi can be hard to lyse, we compared four methods for lysis and DNA purification. Performance was tested over 24 DNA samples from known eukaryotes and eight human stool samples. No single marker gene strategy can quantify all eukaryotic sequences in a sample, but the methods described here allow characterization of a large and well-characterized subset.

2.3 Materials and Methods

Sample Collection

Isolates of *Aspergillus*, *Candida*, *Penicillium*, *Cryptococcus*, and Dematiaceous mold were obtained from the Clinical Microbiology Laboratory at the Hospital of the University of Pennsylvania. Cultures were treated at 95 ^{o}C for five min to sterilize before removal from the laboratory. The *Pneumocystis*, *Coccidioides*, *Leishmania*, *Toxoplasma*, *Plasmodium*, *Arabidopsis*, *Saccharomyces* and human samples were from lab strains at the University of Pennsylvania. The samples were bead-beaten for 1 min, heat inactivated for 5 min at 95^{o}C and then DNA was extracted with the Qiagen Stool DNA Kit using the manufacturer's protocol. DNA extraction for these isolates was performed by Rebecca Custers-Allen. In subsequent studies we have found that the Qiagen Stool Kit is not DNA free (data not shown), explaining the origin of

some of the background sequences. The human stool samples were from healthy adults described in [2,19].

Primer design

Greg Peterfreund designed the 18S_0067a_deg primer by screening a set of aligned eukaryotic 18S rRNA gene sequences downloaded from the Silva database [20] and searching for mammal-specific polymorphisms in the 5' conserved regions that flank the hypervariable regions. Three bases at 65-67 were conserved in nearly all 18S rRNA gene but were absent in mammalian 18S rRNA genes, providing the basis for designing selective primers. The NSR399 primer was obtained from the European Ribosomal RNA Database. The ITS amplicons were amplified with the ITS1F/ITS2 primers as in Ghannoum et al [16].

DNA purification

DNA was purified from human stool (stored frozen at -80°C) using four different methods as specified by the manufacturer except where noted. Approximately 220mg of stool was used for each extraction. Human stool samples were extracted by Stephanie Grunberg.

The FastDNA extractions were done with the FastDNA kit as described by Ghannoum et al[16], except the FastPrep Instrument was replaced by a BioSpec Mini-Beadbeater-16. The archaeal extractions were preformed according to the methods of Dridi et al.[21]. The PowerSoil extractions were bead beaten for 1.5 min in MoBio garnet tubes and centrifuged at 1500 rcf for 5 min. 1mL of supernatant was transferred to a PowerBead Tube and heated at 65°C for 10min and then 95°C for 10min. We then used the manufacturer's protocol, skipping the first sample vortex (steps 1 and 2) and spun for 2 min instead of 1 at the spin filter loading step (step 15). The samples that were purified with the PSP extraction method were placed in Lysing Matrix E tubes (MP Biomedical) with 1400 µl of stool stabilizer from the PSP kit and were bead beaten in a

Mini-BeadBeater-16 (BioSpec). Samples were then heated at 95°C for 15 minutes, placed on ice for 1 minute, and spun down at 13400g for 1 minute. The supernatant was then transferred to the PSP InviAdsorb tubes and the rest of the protocol for the PSP Spin Stool DNA Plus was followed according to the manufacturer's instructions. As controls, DNA free water was passed through each DNA extraction procedure, amplified, and samples were sequenced even in cases where no DNA was detectable after amplification ("water controls" Figure 2-5).

Sequence acquisition

Primers with 12 base barcodes were used for 454 FLX sequencing. DNA was initially amplified with AccuPrime DNA polymerase and buffer 2 (Invitrogen). The PCR was carried out with a 5 min denaturing step at 95°C, followed by 35 cycles of a 45 sec denaturing step at 95°C, a 45 sec annealing step at 56°C, and a 1.5 min extension step at 72°C. Finally, there was a 10 min extension step at 72°C and samples were held at 4°C. The resulting amplicons were then sequenced on a Roche 454 Junior instrument using the FLX Titanium chemistry according the manufacturer's instructions.

Bioinformatic analysis

Raw sequence data was denoised and analyzed using the QIIME pipeline [22]. OTUs were formed by CD-HIT [23]at 99% convergence for the 18S rRNA gene amplicon and 95.2% convergence for the ITS1 amplicon. The last 20 bases in reads from the 18S rRNA gene amplicons were trimmed due to low overall quality. Homopolymer limits in the read quality filtering were disabled for the ITS1 amplicon.

The BROCC classifier

BROCC classifies query sequences by voting on BLAST hits scored by identity. All hits are filtered for identity and coverage. Classifications are voted on in a bottom up fashion, starting at the species level. Specific identity filters are specified by the user for the genus and species level in addition to the main identity filter used for all other levels. Once a classification is made at a given level, all the higher levels are called automatically. If a consensus is not reached at a given level, that level and lower levels are left blank in the final classification. Genus and species identity filters were set at 83.05% and 95.2% for the ITS1 amplicon and 96% and 99% for the 18S rRNA gene amplicon. All other levels were filtered at 80%. The minimum coverage and generic classification filters were set at 70% for all amplicons. Classifications at the species through family level required a 60% majority to be accepted. Classifications at the order level and above required a 90% majority to be accepted. The BROCC program is implemented in Python version 2.7. It queries the NCBI taxonomy and requires local installations of MySQL and BLAST. The online BLAST user interface was used in error checking.

2.4 Results

DNA from food is detectable in fecal material

Humans consume other eukaryotes as food, so in order to design maximally useful amplicons for the detection of eukaryotic rRNA gene sequences in gut microbiome samples, we first investigated the survival of DNA during passage through the gut. In an early study of this issue, plasmid DNA was fed to mice and low molecular weight DNA from pellets was found to contain apparent plasmid-derived DNA, which was detected as smears on Southern blots [24]. Another study showed that 16S rRNA gene sequences in pellets of gnotobiotic (germ-free) mice resembled 16S sequences in mouse food [25]. Our own evidence from shotgun metagenomic

studies also suggested that DNA from food may be detectable in human stool [19], though this has not been studied in detail. In a further study (data not shown), we gavaged mice with purified bacterial plasmid DNA and showed that plasmid DNA could be detected in fecal pellets six hours but not 60 hours after feeding using Taqman Q-PCR. Based on these observations, we sought to identify eukaryotic rRNA gene amplicons that could detect single cell eukaryotes of the human microbiome while selectively avoiding amplifying rRNA genes from food organisms and host.

Design of amplicons

We targeted the 18S rRNA gene (Figure 2-1A) due to its high conservation among eukaryotes [26] and the substantial bioinformatic resources available for 18S rRNA gene analysis [20,27]. We analyzed 18S rRNA gene sequences from the Silva database [20] and manually scanned alignments for mammalian and plant specific polymorphisms. A primer was designed by Greg Peterfreund and analyzed *in silico* for specificity by Scott Sherrill-Mix (18S_0067a_deg; Figure 2-1B and C) that showed low edit distance (high identity) to 18S rRNA genes of Fungi, Amoebozoa, Chromalveolates, Rhizaria, and most excavates, but showed lower identity to human 18S rRNA genes due to mismatches at the 3' end. In addition, some though not all plants showed relatively high edit distance to 18S_0067a_deg (Figure 2-1B and C) . It was paired with the universal NSR399 18S rRNA gene primer, which is complementary to all eukaryotic clades [28].

The 18S rRNA gene is not sufficiently polymorphic for classification of some groups at a low taxonomic level [17], so we also tested an ITS1 primer set, which queries a less-conserved region and targets fungi selectively. We used a version of the ITS1F/ITS2 primer set previously reported to show discrimination at low levels of the fungal taxonomy [16].

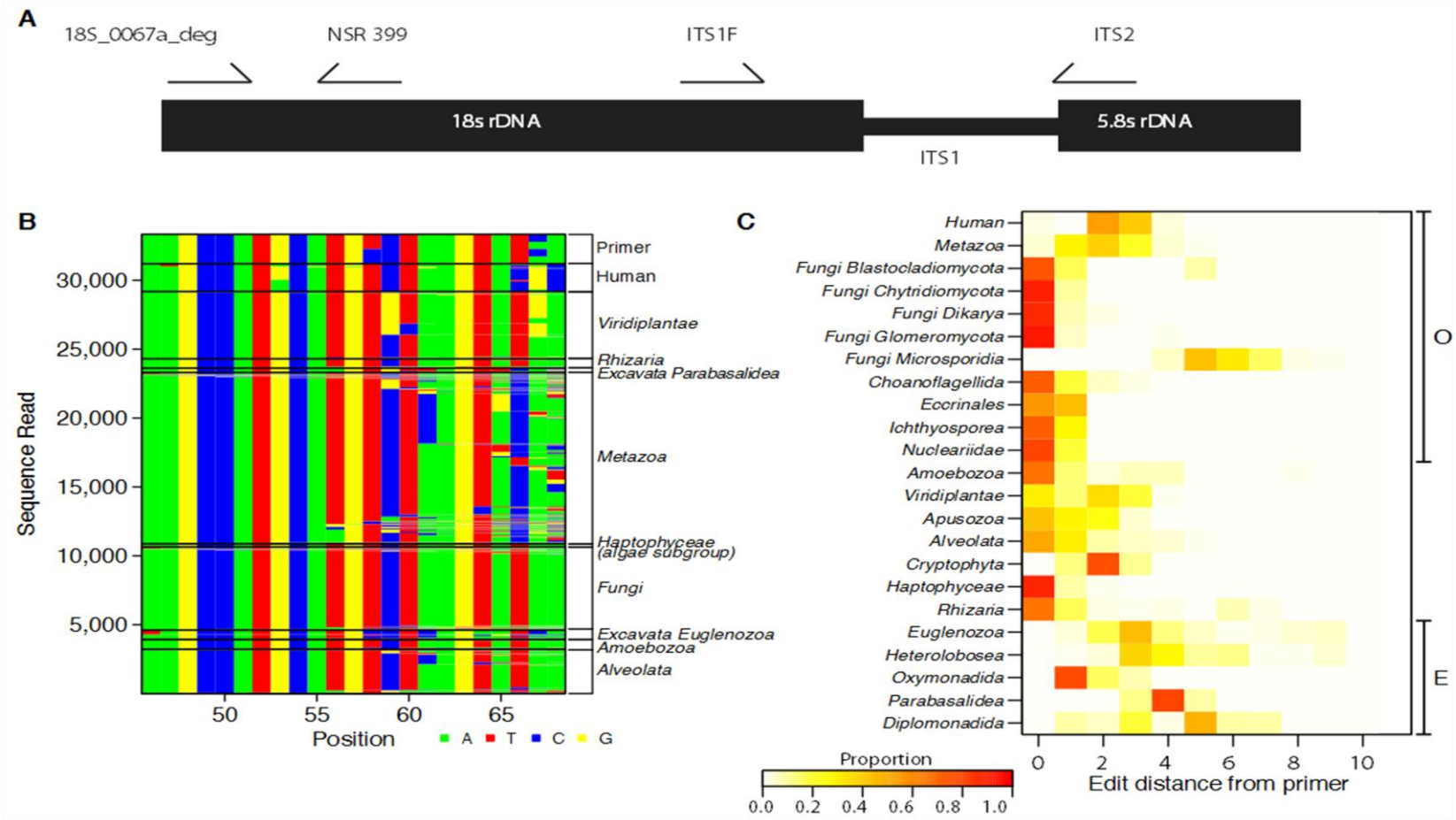


Figure 2-1. The eukaryotic ribosomal DNA locus and the targets of amplicons studied here. A) Part of the rRNA gene locus showing primer binding sites (not to scale). B) Comparison of sequence complimentarily for the 18S-0067a-deg primer against various eukaryotic groups. Clades belonging to Excavate (E) or Opisthokont (O) groups [31] [42] are marked on the left. *Apusozoa* consist of ciliated protozoans[43]. *Cryptophyta* and *Haptophyceae* are a subgroups of algae [31]. C) Heat map indicating the edit distance (numbers of mismatches) between the 18S-0067a-deg primer and target sites in the indicated organisms. Colors are scaled to the relative proportion of each taxa represented at each edit distance. Panels 2-1B and 2-1C were generated by Scott Sherrill-Mix.

All primers used for amplification also contained a DNA bar code, which consisted of 12 bases which indexed the DNA specimen studied. Sequence reads could then be separated by bar code during bioinformatic analysis, allowing many amplicons to be sequenced in pools.

Classification of amplicon sequences using BROCC (BLAST Read and OTU Consensus Classifier)

Classifying sequences from microeukaryotes presents special challenges in automated assignment. 1) There are large numbers of accepted synonyms for many taxonomic groups. 2) Databases contain an unusually high level of misclassifications. 3) Sexual and asexual forms (anamorphs and telomorphs) of a single fungal species can be in different taxa, even up to the family level. 4) Databases contain large numbers of environmental sequences with minimal or no classification that nevertheless are returned as hits from database searches. For these reasons, we designed BROCC to classify single cell eukaryotes while respecting these limitations. BROCC also facilitates interfacing with the popular QIIME pipeline[22], which was originally developed for use with bacterial 16S rRNA gene tags.

We chose to use a BLAST-based method, rather than a kmer-based classifier such as RDP [29], because the high level of variation between closely related ITS sequences could result in misplaced assignments. Phylogenetic-based methods such as ARB [27]have difficulties with ITS sequences because of rapid divergence and common indels.

BROCC classifies amplicons using BLAST searches against large and relatively uncurated databases. There are curated databases for several eukaryotic amplicons that can be used for phylogenetic assignment [17,20], but large curated databases do not exist for ITS1, which is used here. It is widely speculated that the great majority of fungi have not been studied, motivating use of the broadest possible databases for human microbiome studies. BROCC uses

blastn, but output from other versions of BLAST such as blastx can be substituted. Parameters are user-adjustable. BROCC first filters input BLAST hits for sufficient coverage and identity to the query sequence. If a query sequence has too many hits that are below the preset coverage threshold (70% default), or BLAST did not return a hit, it is not classified, and a message is written to the output file. BROCC then determines the identity and taxonomic hierarchy of each high quality hit using a local user installed sql database and NCBI's e-fetch tool.

BROCC then votes on the quality filtered BLAST hits, starting at the species level. At each level of the taxonomy BROCC requires the taxon with the most votes to surpass a user specified threshold for that level in order to accept it as a valid classification. If a sufficient majority is not reached, BROCC will not make a classification for that level and iterate to the next higher taxonomic level for another round of voting. BROCC filters are independently configurable at the genus and species levels, and another filter can be assigned for the remaining taxonomic levels. Here different defaults were used for ITS and 18S rRNA gene amplicons. Species and genus defaults for ITS rRNA gene amplicons were chosen on the basis of [30], and are 95.2% and 83.05%; 80% was used for higher taxa. For 18S rRNA gene amplicons, experience (data not shown) indicated that 99% was suitable for species attribution, 96% for genus, and 80% for higher levels.

BROCC also contains a user modifiable list of high level and partial assignments in its configuration file. These assignments are ignored at lower taxonomic levels where they are uninformative and can distort voting, but included in higher levels. For example, a sequence read with a kingdom level assignment only is excluded up to the kingdom level, at which point the vote is counted in the kingdom assignment. In cases where the proportion of high level and partial assignments exceeds a given threshold (default 0.70), the query sequence is unassigned and marked accordingly.

BROCC output includes both files containing classifications with standardized taxonomy (domain, kingdom, phylum, class, order, family, genus, species) and a second with the complete NCBI taxonomy[31], which includes subtaxa, supertaxa, and unranked intermediate taxonomic levels. The third file contains a log of the voting record, including how many votes were cast, how many votes the winning taxon received, and how many generic classifications were ignored for each query sequence. This file also indicates those queries that were unclassified. Both taxonomy files are suitable for use in the QIIME pipeline (i. e. they are in the same format as the output classifications as the QIIME assign_taxonomy.py script).

Testing BROCC performance on an in silico-constructed community of known membership

Kyle Bittinger verified performance of BROCC by testing assignments over an *in silico* generated mixed community of known membership (Figure 2-2). We selected six eukaryotic microbial organisms, and extracted sequences corresponding to our 18S and ITS rRNA gene amplicon regions. To simulate the characteristics of pyrosequencing data, we added base substitution errors at a rate of 1% and truncated each sequence by a length selected randomly from an exponential distribution, such that the average trim value was 5 bases. For each strain, 32 different reads were generated, and then classified by BROCC.

For the 18S rRNA gene sequences, the majority of reads for each organism were classified to at least the genus level for 4/6. One of the remaining two was classified at the family level, and another was classified at only the phylum level (*Rhodotorula bacarum*). For *Rhodotorula*, the NCBI taxonomy jumps from phylum to genus, disrupting attribution. For the ITS amplicon, 4/6 were classified to the species level and one was classified at the genus level (*Penicillium*). *Dendryphion* was unclassified, due to an abundance of short sequence matches in the database that covered less than 70% of the ITS query and thereby disrupted assignment. We

conclude from this that 1) BROCC works well for attribution even in the presence of sequence errors and truncations, 2) the ITS amplicon yields lower level assignments than the 18S rRNA gene amplicon for those sequences accessible with the ITS primers used, and 3) failed assignments were mainly attributable to problems in the underlying database.

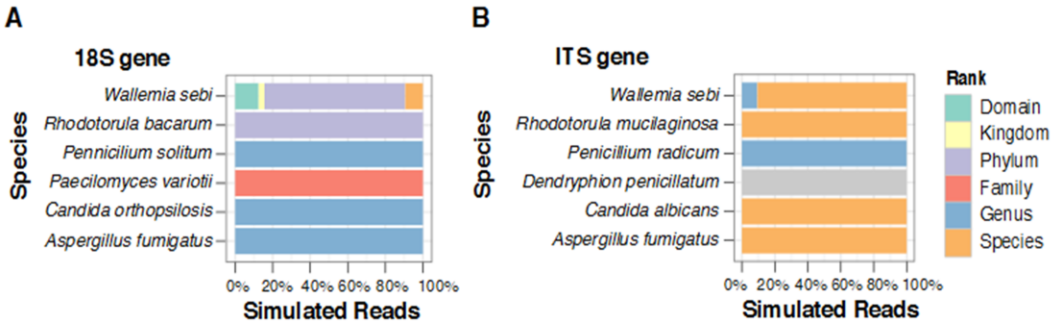


Figure 2-2. Classification of *in silico* generated mixed communities of eukaryotic rRNA gene sequences using BROCC. For each community, sequences from six different organisms were pooled, with 32 sequences per organism. Sequences contained 1% added error and variable length truncations. The organism chosen is shown to the left of each panel, assignments are shown to the right of each figure panel by the color code. A) 18S rRNA gene assignments. B) ITS rRNA gene assignments. Figures were generated by Kyle Bittinger.

Testing the pipeline using a collection of DNAs from microeukaryotes of clinical interest

In order to test the performance of our pipeline, we tested DNA extracted from clinical isolates of fungi and molds, as well as selected laboratory strains of model eukaryotes (Supplementary Table 2-1). We also tested DNA from humans and *Arabidopsis thaliana*, which are selectively non-targeted organisms. DNA samples were amplified with our 18S and ITS rRNA gene primer pairs and sequenced using the 454/Roche platform. The raw sequences (54,698 for 18S rRNA genes, 35,259 for ITS genes) were processed and denoised in the QIIME pipeline [22]. OTUs were formed with percent identity values used for species level attribution above. Taxa were assigned using BROCC. We scored a BROCC classification as correct if it

returned an accepted synonym, anamorph, or teleomorph from the Mycobank database [32] or the NCBI taxonomy database matching the known assignment (Figure 2-3A and B).

For the 18S rRNA gene amplicon, of the 23 classified samples tested, the major OTU was annotated as the correct organism at the family level or lower for 18 specimens and at the genus level or lower for 16 specimens (Figure 2-3A). Taxa called correctly at the genus level or lower included *Aspergillus*, most *Saccharomycetaceae* yeasts (*Candida* and *Saccharomyces*), *Penicillium*, *Pneumocystis*, and *Toxoplasma gondii*. A low number of reads were detected for *Arabidopsis thaliana* despite the effort to avoid amplifying plant DNA, and these reads were also correctly placed. Taxa called correctly to the family level included *Leishmania* and *Candida glabrata*. *Coccidioides* was called correctly at the Class level. The human DNA sample yielded only low numbers of reads, and the most abundant OTU matched *Saccharomycetaceae* yeasts, consistent with the idea that only low level contaminating environmental DNA amplified from these samples. *Plasmodium* did not amplify with this primer pair, consistent with the large edit distance between the primer sequences and the rRNA gene target.

Cryptococcus neoformans classified correctly to the species level, but *Cryptococcus laurentii* initially only classified correctly to the phylum level. Analysis showed this was due to a large number of database entries for closely related sequences annotated as 'Uncultured soil basidiomycete'. We thus added this term to our list in BROCC of unhelpful classifications to be excluded, after which *C. laurentii* was correctly classified to the Class level.

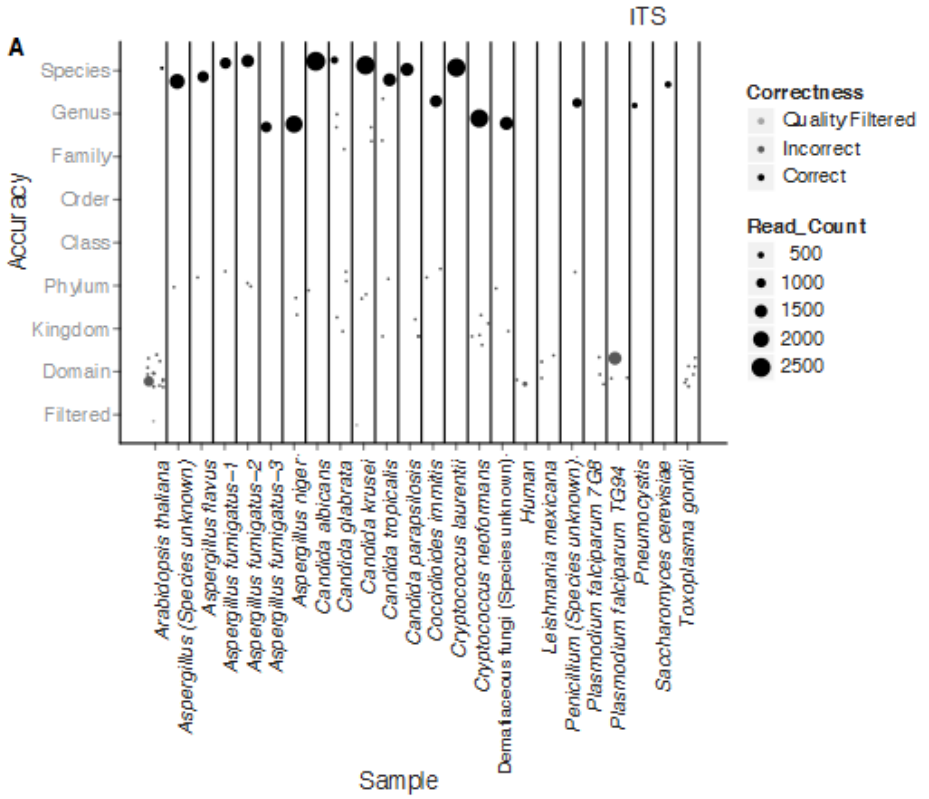
For the ITS gene amplicon, of the 23 samples tested, the major OTU was annotated as the correct organism at the genus level or lower for 18 specimens (Figure 2-3B). Taxa called correctly included *Aspergillus*, *Cryptococcus*, *Penicillium*, *Pneumocystis* and *Saccharomycetaceae* yeasts (*Candida* and *Saccharomyces*). Human and *Arabidopsis* were not

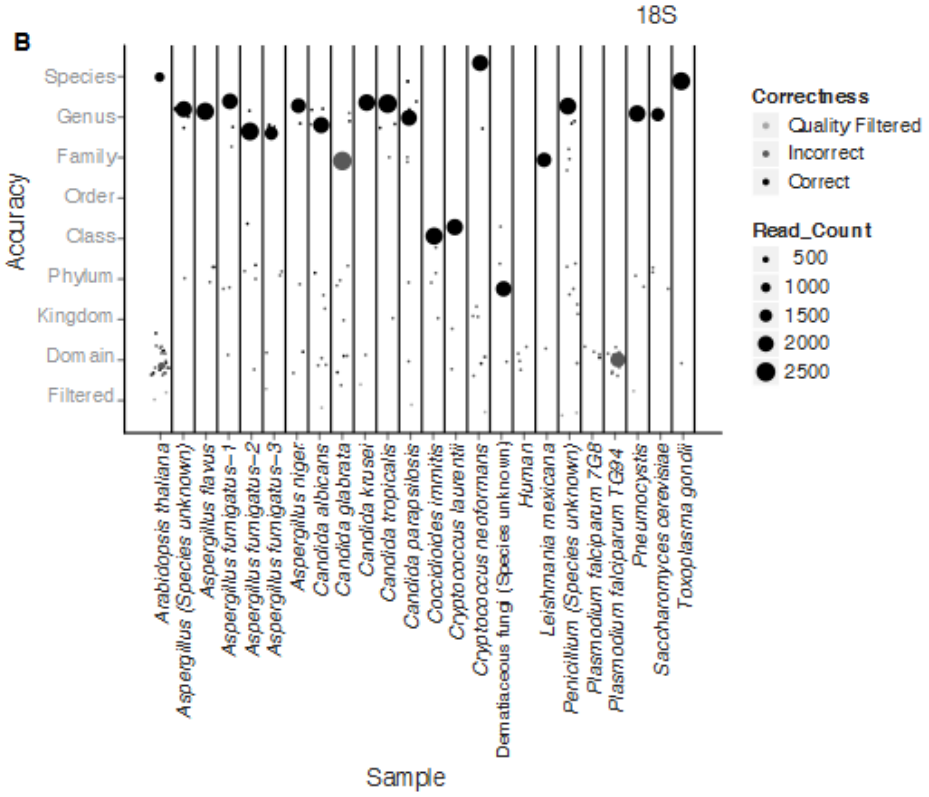
correctly called and the major OTU corresponded to a *Saccharomyces* yeast, consistent with low level contamination. We also failed to correctly call the Apicomplexan samples (*Plasmodium* and *Toxoplasma*), consistent with the presence of several mismatched basepairs in the forward primer, and *Leshmania*. Again, for most of these the numbers of reads were low and corresponded to abundant environmental Fungi which were probable contaminants.

One clinical strain was dubbed a Dematiaceous mold, which is not a taxonomic identifier. Analysis of the 18S rRNA gene amplicon data called it only as *Ascomycota*, because divergent annotation a lower levels obstructed deeper classification by BROCC. However analysis of the ITS amplicon data called it as genus *Exophalia*, which fits with the clinical profile.

Most samples also showed additional low level OTUs, usually represented by less than 5 sequence reads unrelated to the correct call. In some cases these were identifiable as common environmental Fungi that likely contaminated either the original DNA samples or reagents used for DNA purification. Extensive amplification of extraction negative controls occasionally yielded such OTUs (data shown below). Other low level OTUs in Figure 2-3 were not identified and may be products of mispriming, chimera formation, or pyrosequencing error.

Figure 2-3. Analysis of DNA samples from known eukaryotes. A) 18S and B) ITS rRNA gene amplicons. The sample tested is listed along the x-axis. The y-axis shows the level of taxonomic placement of each OTU in each sample relative to the correct taxon indicated on the x-axis. The numbers of sequence reads are shown by the size of the point. Thus large circles high up on the y-axis indicated correct placement of the major taxa.





Comparison of DNA purification methods

Choice of cell lysis and DNA extraction methods influences both the DNA yield and proportions of taxa for bacterial 16S rRNA gene analysis [2], and the known difficulties of lysing yeasts suggest the issue may be even more pronounced here. We thus compared four different extraction methods for preparing samples for analysis of eukaryotic rRNA gene sequences: PSP Spin Stool DNA Plus Kit, MoBio PowerSoil kit, FastDNA with Fungal Protocol [16], and an archaeal specific extraction method [21]. For some, harsher lysis steps were used than in the original protocols (see methods). Eight stool samples from healthy adults were subjected to separate extractions with each of the four kits. The PSP kit yielded the most DNA on average for the same weight of starting material. Output DNA from each method was then tested using both the ITS1 and 18S rRNA gene amplicons.

Amplification products were separated by agarose gel electrophoresis and visualized by staining with ethidium bromide (Supplementary Figure 2-1). The genomic DNA from the FastDNA protocol produced no detectable amplification. The PSP and PowerSoil extractions produced similar banding patterns on ethidium bromide stained agarose gels, though the PSP extractions produced brighter bands overall. The archaeal extraction method produced sporadic bands that were generally less bright than the PSP and PowerSoil samples. Based on these findings, the PSP kit seems superior. The archaeal, PSP, and PowerSoil samples were then compared after deep sequencing by the 454/Roche method.

Comparison of taxa reported with the 18S and the ITS rRNA gene amplicons for human stool samples

We acquired 54,411 sequence reads for the 18S rRNA gene amplicon and 39,827 sequence reads for the ITS1 amplicon from the 8 stool samples (Supplementary Table 2-2). The

sequence reads were clustered into OTUs and assigned to eukaryotic taxa using BROCC. The relative abundance of community members was assessed by plotting OTUs ranked by abundance versus their within sample abundance for samples extracted with the PSP method (Figure 2-4A and B). The 18S rRNA gene amplicon yielded 93 OTUs and the ITS amplicon yielded 215 OTUs. For both the 18S and ITS rRNA gene amplicons, a few OTUs contained most reads, and this was more pronounced for the 18S rRNA gene amplicon data. The majority of OTUs assigned by BROCC from both amplicons belonged to fungal phyla (62.4% in 18S and 90.5% in ITS1 rRNA gene amplicons), mainly *Ascomycota* (81.0% in 18S and 57.4% in ITS1 rRNA gene amplicons) and *Basidiomycota* (17.2% in 18S and 25.7% in ITS1 rRNA gene amplicons). Recovery of plant and animal DNA from the 18S and ITS rRNA gene amplicons was suppressed effectively. Only two OTUs in the 18S rRNA gene amplicon totaling 35 reads and 5 OTUs in the ITS amplicon totaling 5 reads were classified as plant. No OTUs were classified as vertebrate, though in other experiments with these primers small numbers of host and vertebrate sequences have been detected (data not shown).

The numbers of reads returned for each OTU can be used as a surrogate for relative abundance, though this measure must be used with caution due to unequal amplification due to internal secondary structure, differential complementarity of target sequences and primers, and different amplicon lengths. The proportions of sequences are shown as stacked bar graphs in Figure 2-5 for the PSP and PowerSoil extraction methods. Yields from the Archaeal extraction were lowest of the three, and showed multiple samples with few or no reads, and so were not studied further. Sequence reads were detected in 6 of 8 negative controls (Figure 2-5 B and D), in which DNA-free water was subjected to the purification, amplification and sequencing procedures, but the read numbers were typically much lower than for the stool samples (Supplementary Table 2-2).

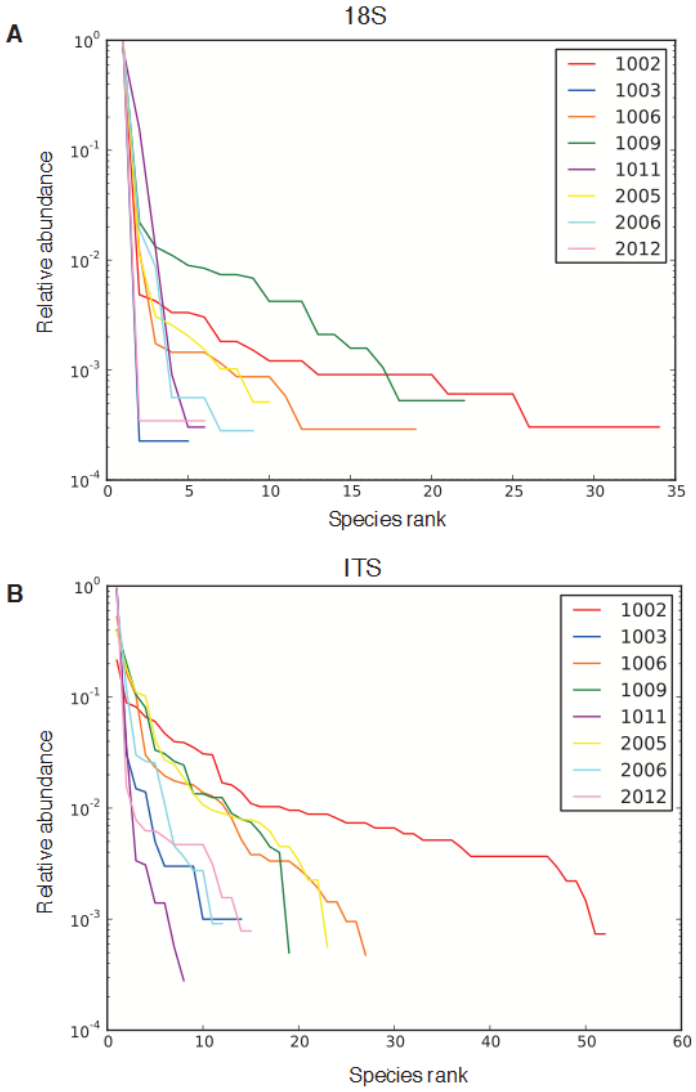


Figure 2-4. Rank-abundance plots for OTUs from stool samples. A) 18S and B) ITS rRNA gene amplicons. The rank (relative abundance) of each OTU is shown on the x-axis, with the most abundant on the left. The proportion contributed by that OTU is shown on the y-axis. The key in the upper right shows the color code for the different human subjects studied.

For the 18S rRNA gene amplicon, 99.6% of fungal reads were assigned to *Ascomycota*, while the rest were assigned to *Basidiomycota*, except for a 3 read OTU assigned to *Entomophthora*. For the ITS amplicon, 83.7% of fungal reads were assigned to *Ascomycota*,

9.79% were assigned to *Basidiomycota*, and 6.4% were only classified to the kingdom level. Twelve reads from PowerSoil extraction of subject 1006 were assigned to *Mucoromycotina*.

The 18S rRNA gene amplicon also detected two gut parasites, *Blastocystis* and *Endolimax*. These were not detected using the ITS amplicon, which is specific for Fungi. The DNA extraction method used affected the results--*Blastocystis* was detected in both the PSP and PowerSoil extractions from subject 2006 and *Endolimax* in the PSP extraction in subject 2006, but not in samples extracted by other methods. It is unclear whether this divergence is due to bias in the extraction methods or uneven distributions of organisms in stool samples.

The *Saccharomycetaceae* proved to be the dominant lineage in the eight stool samples for both the 18S and ITS1 rRNA gene amplicons. Both amplicons were dominated by *Saccharomyces* and *Candida* genera (Figure 2-5A and C). The majority of *Saccharomycetaceae* reads recovered with the 18S rRNA gene amplicon were classified as *Saccharomyces* in all samples. However, for the ITS1 rRNA gene amplicon, reads were classified as a mixture of *Candida* and *Saccharomyces*. Analysis of the 18S rRNA gene sequence over the window queried by our amplicon revealed that *Saccharomyces* and *Candida* are poorly distinguished over this region, which was corroborated by a multilocus phylogeny over the *Saccharomycetaceae* family [33].

Aside from the typical gut inhabitants, our study yielded several examples of fungal rRNA genes potentially derived from food. In subject 1006, *Agaricus bisporus*, the common button mushroom, was detected as a high count OTU using all extraction methods for the ITS1 amplicon samples. *Claviceps purpurea*, which grows on rye and other cereals and is a causative agent of ergot [34], was detected as a rare OTUs in subjects 1002, 1006, and 2006. *Wallemia sebi*, often found in food[35], was detected in 1002, 1006, 1009, and 2005 for multiple extraction

methods. The substantial amount of *Saccharomyces* that appeared in all subjects, may be derived from bread, beer, or other leavened and fermented foodstuffs. Distinguishing fungal sequences derived from food presents an ongoing challenge in gut microbiome studies.

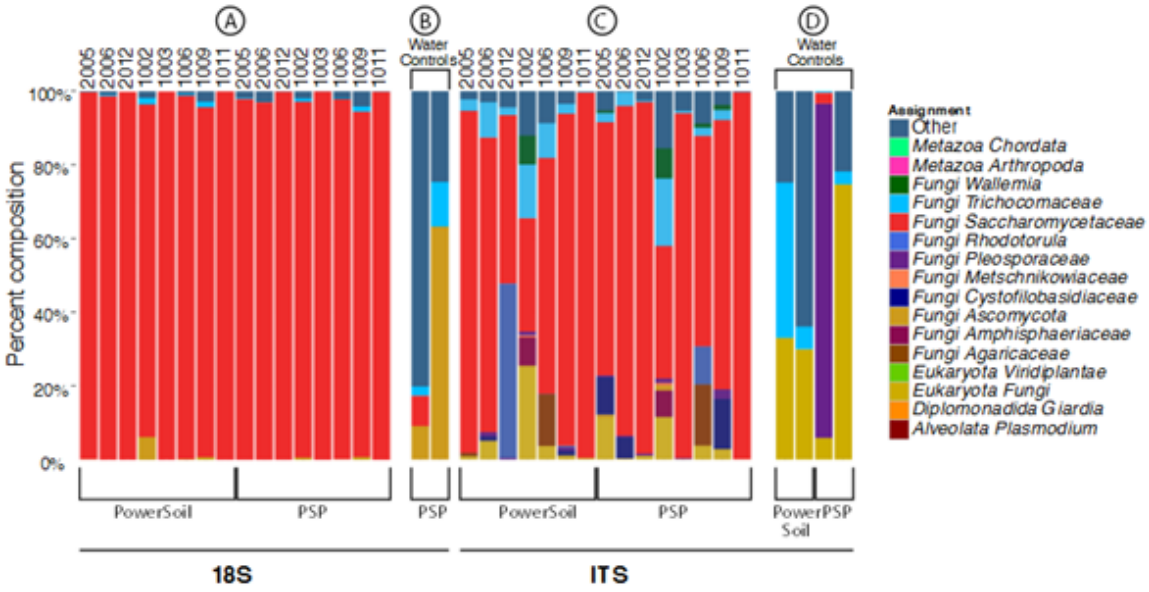


Figure 2-5. Comparison of major eukaryotic microbes detected in human stool. Samples were assayed with the 18S rRNA gene amplicon, the ITS1 rRNA gene amplicon, and the shotgun genomic data in human stool. Human subjects and DNA purification methods are as indicated on the x-axis. Taxa are shown at the Family level or as indicated. A) 18S rRNA gene amplicon used to analyze stool samples. B) 18S rRNA gene amplicon contamination controls. C) ITS amplicon stool samples. D) ITS rRNA gene amplicon contamination controls. The contamination controls in B) and D) consisted of DNA-free water passed through the full DNA purification, sequencing and analytical pipeline--6 of 8 samples yielded pyrosequence data, though with low read numbers.

Comparison of the performance of BROCC to other classifiers over the experimental data sets.

Taking advantage of these data, we next compared BROCC to two other classifiers, MEGAN and MARTA, which were not specifically designed for use with single cell eukaryotes. Supplementary Table 2-4 summarizes the differences among the programs. For more discussion of the assignment problem see [36,37,38,39] and references therein.

The three programs were first tested by comparing for the number of correct assignments for the known samples in Figure 2-3. The number of samples with correct assignments for the 18S rRNA gene amplicon at the Genus level or lower were 17 for BROCC, 19 for MARTA, and 3 for MEGAN out of 20 possible. For the ITS amplicon, the numbers were 18 for BROCC, 11 for MARTA, and 6 for MEGAN out of 18 possible. Thus BROCC and MARTA were comparable, with BROCC performing somewhat better for the ITS gene amplicon. MEGAN was more conservative and made fewer low level assignments for ITS, because it was more strongly influenced by database errors or alignments with only high level taxonomic placements.

In some comparisons, MARTA yielded more low level classifications due to accepting single high quality matches for assignment, which can be an advantage or disadvantage depending on the quality of the underlying database. MARTA classified *Candida krusei* as *Pichia fermentans* in the 18S rRNA gene amplicon and *Coccidioides immitis* as *Coccidioides posadasii* in the ITS amplicon. MARTA considered 4 database hits for *C. krusei* and 6 for *C. immitis*, while BROCC considered 98 for *C. krusei* and 27 for *C. immitis*. In both cases BROCC made a correct genus level assignment only and not the erroneous species level assignment. In four cases in the ITS amplicon assignments, MARTA failed to make an assignment due to interference from multiple aligning database sequences assigned as "unidentified" or "uncultured", which were correctly classified to low taxonomic levels by BROCC.

We then compared the assignments for BROCC and MARTA against the human stool samples, for which the composition is not known. MEGAN was not considered further due to inferior performance on the known samples. We assigned each classification level a score. Species level assignments received value 1, genus value 2, and so on up to unclassified, which received value 9. Scores were compared between BROCC and MARTA. This showed that BROCC consistently yielded lower level classifications (Wilcoxon signed rank test; p=0.014 for

the 18S rRNA gene amplicon, and $p=4.1x10^{-15}$ for the ITS amplicon). Inspection of the data showed the numbers of unclassified OTUs generated by MARTA was largely responsible for the inferior score.

BROCC also contains functionality assisting in implementation that is lacking in the other packages (Supplemental Table 2-4). BROCC can extract useful information from partial assignments -- for example, a database hit assigned only at the kingdom level, is not tallied during the process of assignment at lower ranks, but considered in the case of a kingdom assignment. BROCC reports the reason for excluding database hits in the output file. BROCC also outputs file types that are easily integrated into the QIIME pipeline [22] for evaluation of microbial community structure, accelerating downstream steps in a typical analysis.

2.5 Discussion

Here we present a pipeline for characterization of eukaryotic taxa in microbiome samples. For many types of samples, single cell eukaryotes are a minority component, so that shotgun metagenomic analysis is inefficient and expensive. Thus, despite the rapid advance of methods, marker gene analysis remains the method of choice for many applications.

We describe experiments to characterize the performance of two primer sets querying the eukaryotic ribosomal rRNA genes. Data from us and others shows that interfering DNA from food or host cells must be considered in designing the amplification strategy. We thus devised an 18S rRNA gene amplicon that selectively avoids plant and animal 18S rRNA gene sequences. We also studied a second amplicon that targets ITS sequences from Fungi, which also minimizes contamination with plant and animal DNA but queries a narrower group of eukaryotes. The ITS rRNA gene region studied is more diverse than the 18S rRNA gene region, allowing lower level phylogenetic placement of some fungal groups. Both amplicons were effective in detecting

Aspergillus, *Saccharomycetaceae*, *Penicillium*, and *Pneumocystis*. The 18S rRNA gene amplicon selectively detected *Leishmania* and *Toxoplasma*. In stool, the 18S rRNA gene amplicon but not ITS detected *Blastocystis* and *Endolimax*. The ITS amplicon selectively classified *Cryptococcus* and the Dematiaceous mold. Neither primer set detected *Plasmodium*. Both amplicons detected *Saccharomycetaceae* yeast as the major group in stool samples. In unpublished work, the ITS amplicon has also been used to characterize bronchoalveolar lavage samples that were also typed in clinical culture-based assays, producing nearly identical assignments (E. Charlson, R. Collman, and F. D. B., unpublished data).

The present state of fungal taxonomy creates challenges in data analysis. Most fungi have not yet been formally described by taxonomists [40], so many sequence reads will be from unknown groups. Names differ for anamorphs (asexual forms) and teleomorphs (sexual reproductive forms) of what are apparently the same species, either of which may occur in the microbiome [41]. Consequently, several OTUs were classified with different names, but belonged to the same holomorph (pool of anamorphs and teleomorphs). Even though they are the same holomorph, *Candida* is taxonomically placed in the family *Saccharomycetaceae*, but *Clavispora* is placed in the family *Metschnikowiaceae*. Efforts to improve databases by eliminating the dual naming system and creating accurate phylogenies for fungi should help in this regard [17].

We demonstrated that DNA can survive passage through the GI tract of a mouse, albeit inefficiently, and our rRNA gene amplicon assays of human stool did detect some OTUs that likely came from food. For some of the fungal groups, it is difficult to know whether they are true gut residents or transients from food. Perhaps the development of detailed databases of eukaryotic rRNA gene sequences common in human food can assist in distinguishing true gut residents from transients.

2.6 Acknowledgements

This work was supported by Project UH2DK083981, the Penn Genome Frontiers Institute; NIH AI087990, AI083480, U01AI095608 to (DA); NIH AI39368 (GDW); T32-AI060516 (DAH); Penn Digestive Disease Center (P30 DK050306); The Joint Penn-CHOP Center for Digestive, Liver, and Pancreatic Medicine; S10RR024525; UL1RR024134, K24-DK078228; and the Howard Hughes Medical Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources, National Institutes of Health, or Pennsylvania Department of Health.

2.7 References

1. Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. Nat Rev Genet 13: 260-270.

2. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, et al. (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiol 10: 206.

3. Agarwal R (2011) Severe asthma with fungal sensitization. Curr Allergy Asthma Rep 11: 403-413.

4. Banuls AL, Bastien P, Pomares C, Arevalo J, Fisa R, et al. (2011) Clinical pleiomorphism in human leishmaniases, with special mention of asymptomatic infection. Clin Microbiol Infect 17: 1451-1461.

5. Madigan MT, Martinko, John M. (2006) Brock Biology of Microorganisms. Upper Saddle River: Pearson Prentice Hall.

6. Ott SJ, Kuhbacher T, Musfeldt M, Rosenstiel P, Hellmig S, et al. (2008) Fungi and inflammatory bowel diseases: Alterations of composition and diversity. Scand J Gastroenterol 43: 831-841.

7. Grice EA, Segre JA (2011) The skin microbiome. Nat Rev Micro 9: 244-253.

8. Harp DF, Chowdhury I (2011) Trichomoniasis: evaluation to execution. Eur J Obstet Gynecol Reprod Biol 157: 3-9.

9. Chaturvedi V, Chaturvedi S (2011) Cryptococcus gattii: a resurgent fungal pathogen. Trends Microbiol 19: 564-571.

10. Caggiano G, Puntillo F, Coretti C, Giglio M, Alicino I, et al. (2011) Candida colonization index in patients admitted to an ICU. Int J Mol Sci 12: 7038-7047.

11. Martin R, Wachtler B, Schaller M, Wilson D, Hube B (2011) Host-pathogen interactions and virulence-associated genes during Candida albicans oral infections. Int J Med Microbiol 301: 417-422.

12. Knutsen AP, Slavin RG (2011) Allergic bronchopulmonary aspergillosis in asthma and cystic fibrosis. Clin Dev Immunol 2011: 843763.

13. Cotton JA, Beatty JK, Buret AG (2011) Host parasite interactions and pathophysiology in Giardia infections. Int J Parasitol 41: 925-933.

14. Scanlan PD, Marchesi JR (2008) Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. ISME J 2: 1183-1193.

15. Gharizadeh B, Norberg E, Loffler J, Jalal S, Tollemar J, et al. (2004) Identification of medically important fungi by the Pyrosequencing technology. Mycoses 47: 29-33.

16. Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, et al. (2010) Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. PLoS Pathog 6: e1000713.

17. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A.

18. Allmang C, Mitchell P, Petfalski E, Tollervey D (2000) Degradation of ribosomal RNA precursors by the exosome. Nucleic Acids Res 28: 1684-1691.

19. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334: 105-108.

20. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188-7196.

21. Dridi B, Henry M, El Khechine A, Raoult D, Drancourt M (2009) High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. PLoS ONE 4: e7063.

22. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335-336.

23. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.

24. Palka-Santini M, Schwarz-Herzke B, Hosel M, Renz D, Auerochs S, et al. (2003) The gastrointestinal tract as the portal of entry for foreign macromolecules: fate of DNA and proteins. Mol Genet Genomics 270: 201-215.

25. Hill DA, Hoffmann C, Abt MC, Du Y, Kobuley D, et al. (2009) Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis. Mucosal immunology.

26. Hagenbuchle O, Santer M, Steitz JA, Mans RJ (1978) Conservation of the primary structure at the 3' end of 18S rRNA from eucaryotic cells. Cell 13: 551-563.

27. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: a software environment for sequence data. Nucleic acids research 32: 1363-1371.

28. Wuyts J, Perriere G, Van De Peer Y (2004) The European ribosomal RNA database. Nucleic Acids Res 32: D101-103.

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

30. Liggenstoffer AS, Youssef NH, Couger MB, Elshahed MS (2010) Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores. ISME J 4: 1225-1235.

31. Federhen S (2012) The NCBI Taxonomy database. Nucleic Acids Res 40: D136-143.

32. Crous PW, Gams, W., Stalpers, J.A., Robert, V., and Stegehuis, G. (2004) MycoBank: an online initiative to launch mycology into the 21st century. Studies in Mycology: 19–22.

33. Kurtzman CP (2003) Phylogenetic circumscription of Saccharomyces, Kluyveromyces and other members of the Saccharomycetaceae, and the proposal of the new genera Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma and Zygotorulaspora. FEMS Yeast Res 4: 233-245.

34. Strickland JR, Looper ML, Matthews JC, Rosenkrans CF, Jr., Flythe MD, et al. (2011) Board-invited review: St. Anthony's Fire in livestock: causes, mechanisms, and potential solutions. J Anim Sci 89: 1603-1626.

35. Zeng QY, Westermark SO, Rasmuson-Lestander A, Wang XR (2004) Detection and quantification of Wallemia sebi in aerosols by real-time PCR, conventional PCR, and cultivation. Appl Environ Microbiol 70: 7295-7302.

36. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M (2011) Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics 12 Suppl 2: S4.

37. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res 40: e9.

38. Liu Z, DeSantis TZ, Andersen GL, Knight R (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res 36: e120.

39. Schloss PD, Westcott SL (2011) Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol 77: 3219-3226.

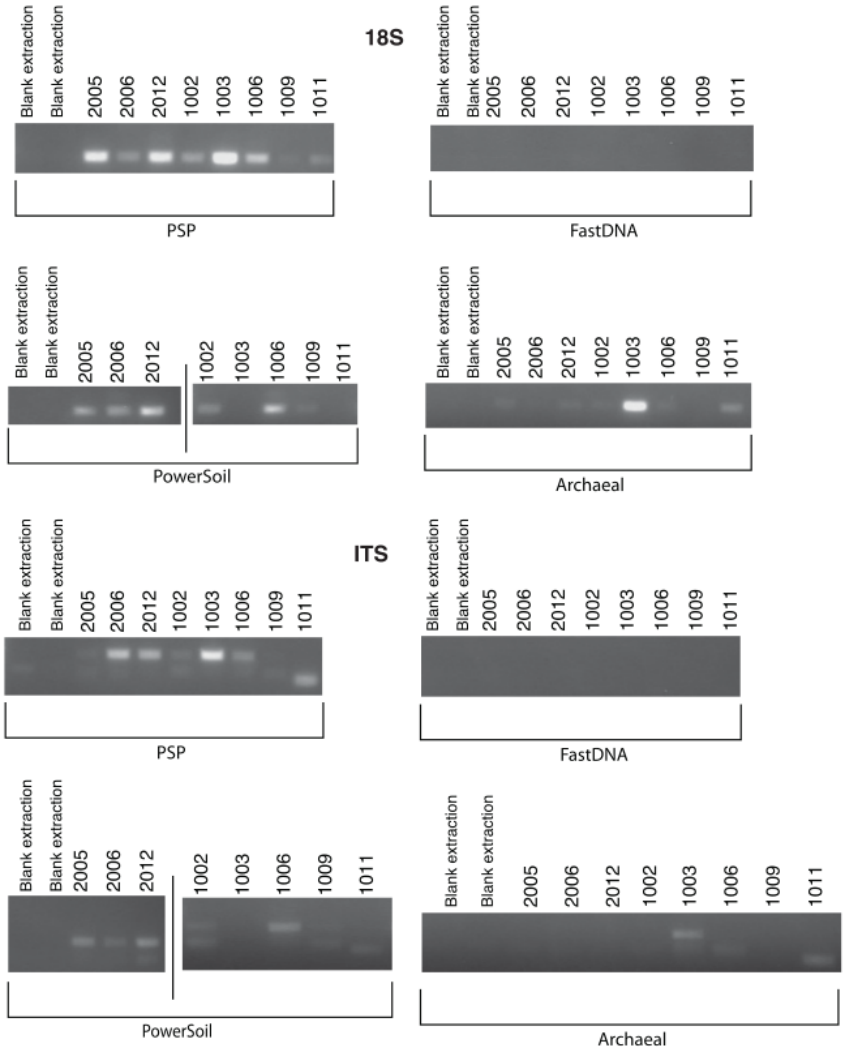
40. Hawksworth DL (2006) Pandora's mycological box: molecular sequences vs. morphology in understanding fungal relationships and biodiversity. Rev Iberoam Micol 23: 127-133.

41. Gargeya IB, Pruitt WR, Simmons RB, Meyer SA, Ahearn DG (1990) Occurrence of Clavispora lusitaniae, the teleomorph of Candida lusitaniae, among clinical isolates. J Clin Microbiol 28: 2224-2227.

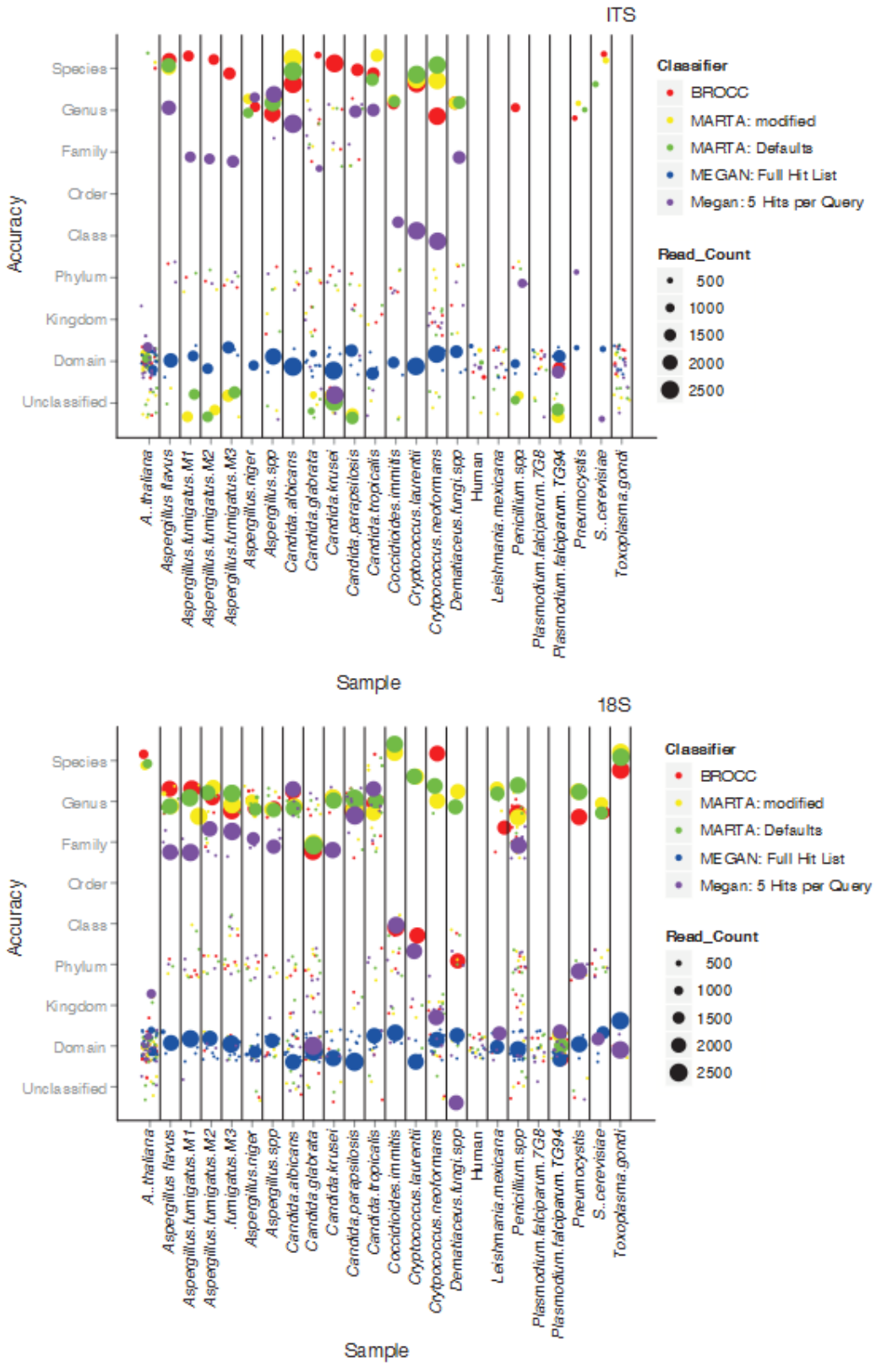
42. Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, et al. (2005) The tree of eukaryotes. Trends Ecol Evol 20: 670-676.

43. Cavalier-Smith T, Chao EE (2010) Phylogeny and evolution of apusomonadida (protozoa: apusozoa): new genera and species. Protist 161: 549-576.

Supplementary Information



Supplementary Figure 2-1. Comparison of PCR amplification reactions for DNA purified from stool using different methods. Average DNA yields were: PSP 59.6 ng/µL, PowerSoil 30.4 ng/µL, FastDNA extraction 15.8 ng/ µL, and the Archaeal method 12.7 ng/µL. PCR products were separated on an 0.8% agarose gel and stained with ethidium bromide. Top: amplification products generated using the 18S primer pair. Bottom: amplification products generated using the ITS1F-ITS2 primer pair.

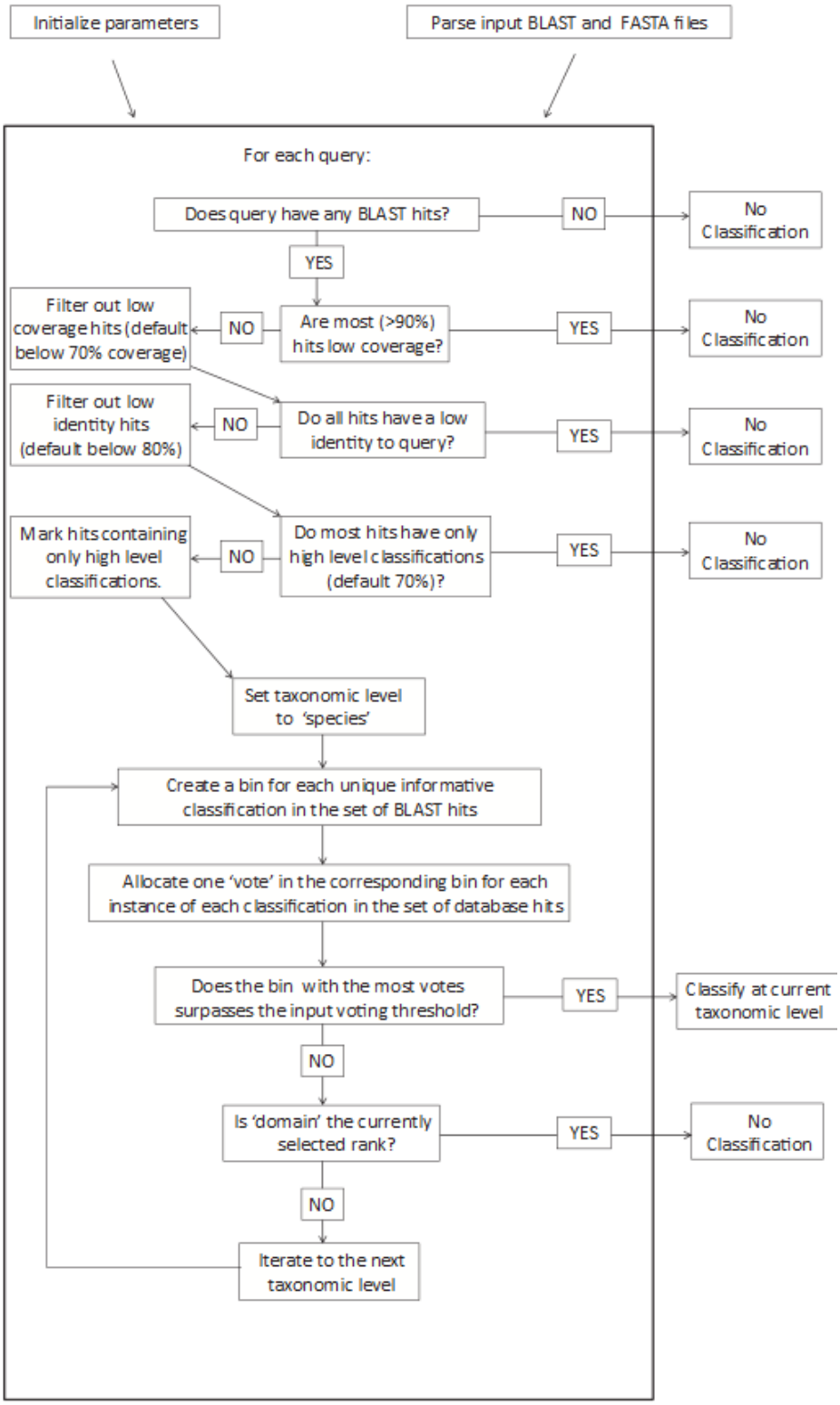


Supplementary Figure 2-2. Analysis of DNA samples from known eukaryotes using BROCC, MARTA, and MEGAN. (a) 18S and (b) ITS rRNA gene amplicons classified by all three classifiers. The sample tested is listed along the x-axis. Individual OTUs in each sample are shown by the points, which are sized in proportion to their read counts. A point is colored by the program and configuration used to classify that point. These data were classified by BROCC using default settings, MARTA using default settings, MARTA using a BLAST word size and voting thresholds to match the BROCC default settings, MEGAN using default settings and the same blastn output used by BROCC, and MEGAN using an abbreviated blastn output with a maximum of five hits per query sequence. The lowest level of correct classification for each OTU is listed on the y-axis.

```
INPUT:
        H, array of BLAST-like database hits
        c, minimum coverage threshold
        I, array of minimum percent identity thresholds
        P, array of minimum consensus thresholds

FOR each taxonomic rank, r, beginning with species:
        LET  V be an array of votes
        FOR each taxon, i, in rank r:
            V_i = 0
    FOR each BLAST-like database hit:
        NEXT hit if coverage < c
        NEXT hit if percent identity < I_r
        m <- taxon assignment for hit at rank r
        INCREMENT  V_m
    k <- argmax(V)
    IF (V_k / sum(V)) > P_r
        OUTPUT  r, k
```

Supplementary Figure 2-3. Pseudocode of the BROCC program.



Supplementary Figure 2-4. Flow chart of BROCC implementation.

			18s			ITS1		
Species	Lineage	Source	Amplicon Target	Level of Classification Correct (Major OTU)	Sequence Count	Amplicon Target	Level of Classification Correct (Major OTU)	Sequence Count
Arabidopsis thaliana	Streptophyta	Lab Strain	No	Phylum	2485	No	Domain (Chaetosphaeriaceae)	1708
Aspergillus flavus	Ascomycota	Clinical Lab	Yes	Genus	2815	Yes	Genus	2079
Aspergillus fumigatus-1	Ascomycota	Clinical Lab	Yes	Genus	2681	Yes	Species	1465
Aspergillus fumigatus-2	Ascomycota	Clinical Lab	Yes	Genus	2324	Yes	Species	1444
Aspergillus fumigatus-3	Ascomycota	Clinical Lab	Yes	Genus	2741	Yes	Species	1697
Aspergillus niger	Ascomycota	Clinical Lab	Yes	Genus	2248	Yes	Genus	1357
Aspergillus (species unknown)	Ascomycota	Clinical Lab	Yes	Genus	2233	Yes	Genus	2451
Candida albicans	Ascomycota	Clinical Lab	Yes	Genus	2452	Yes	Species	2779
Candida glabrata	Ascomycota	Clinical Lab	Yes	Family (Saccharomyces)	2877	Yes	Species	816
Candida krusei	Ascomycota	Clinical Lab	Yes	Genus	2503	Yes	Species	2710
Candida parapsilosis	Ascomycota	Clinical Lab	Yes	Genus	2879	Yes	Species	1749
Candida tropicalis	Ascomycota	Clinical Lab	Yes	Genus	2450	Yes	Species	1779
Coccidioides immitis	Ascomycota	Lab Strain	Yes	Class	2614	Yes	Genus	1585
Cryptococcus laurentii	Basidiomycota	Clinical Lab	Yes	Class	2435	Yes	Species	2669
Crytpococcus neoformans	Basidiomycota	Clinical Lab	Yes	Species	2525	Yes	Species	2630
Dematiaceus fungi spp.	Ascomycota	Clinical Lab	Yes	Phylum	2288	Yes	Genus	1789
Human	Chordata	Cell Line	No	Domain (Penicillium)	13	No	Domain (Phaeosphaeria)	347
Leishmania mexicana	Euglenozoa	Lab Strain	Yes	Family	2117	No	Domain (Pichia)	48
Penicillium spp.	Ascomycota	Clinical Lab	Yes	Genus	2657	Yes	Genus	1146
Plasmodium falciparum 7G8	Apicomplexa	Lab Strain	No	Domain	41	No	Domain (Penicillium)	28
Plasmodium falciparum TG94	Apicomplexa	Lab Strain	No	Domain (Candida)	2209	No	Domain (C. parapsilosis)	1763
Pneumocystis	Ascomycota	Mouse Lung Isolate	Yes	Genus	2495	Yes	Genus	495
Saccharomyces cerevisiae	Ascomycota	Lab Strain	Yes	Genus	1891	Yes	Species	641
Toxoplasma gondi	Apicomplexa	Lab Strain	Yes	Family	2725	No	Domain (Tremellomycetes)	84

Supplementary Table 2-1. Samples studied from known eukaryotic organisms. Species that were classified incorrectly have their assigned taxonomic attribution in parentheses. All other species were classified correctly.

			18s						ITS1					
			Archaeal Extraction		PowerSoil		PSP		Archaeal Extraction		PowerSoil		PSP	
Sample Name	Species	Disease State	Sequence Count	Barcode	Sequence Count	Barcode	Sequence Count	Barcode	Sequence Count	Barcode	Sequence Count	Barcode	Sequence Count	Barcode
1002	Human	Healthy	384	ATGAGACTCCAC	3506	TAATCCACAGCG	3309	TCGAATCACAGC	480	GGCTTGTACTAG	2649	AGCCATACTGAC	1358	CGAGTCTAGTTG
1003	Human	Healthy	3920	GTGTTGCAGCAT	94	GTAGAGCTGTTC	4432	AGTGTGGACTTC	2030	CGGACTACAACT	0	CACGGACTATAC	1000	TGACAATCCAGC
1006	Human	Healthy	565	TGCGTCAGTTAG	2879	TAGTTGCGAGTC	3454	CTCCACATGAGA	3064	TATGGATTCCGG	1217	TGTCAAGTCGTG	2098	ACCGTAATCCAG
1009	Human	Healthy	0	ATGTTGGCTACG	2139	TATCAGGTGTGC	1899	TGTGGTACTACG	1	CACCAGATTCAG	2581	TCGACCAGCAAT	2005	GCTGGTATCTGA
1011	Human	Healthy	3441	TGCTCTAGTGGA	124	ATCGATCTGTGG	3314	CACATGCCTAAG	4349	AGTTAGTGCGTC	3630	ACGTTAGCACAC	3568	CGATAACATGCC
2005	Human	Healthy	709	GTCAAGAACCTC	3473	GTCTCATGTAGG	1955	CTTCGTGGTAGA	433	CTATACCACGGA	1189	GACTAACGTCAC	1775	CTTGACTGAGGT
2006	Human	Healthy	174	TTGCACGATTGG	3821	GTTCGCGTATAG	3574	TAGGATTGCTCG	510	GTATAGGTTCGC	1120	CTAGCGAACATC	1096	AGCTATCCACGA
2012	Human	Healthy	725	GACCACTACGAT	3833	GTATGACTGGCT	2899	AGCGCAACATTC	129	AGTGTTCGATCG	1076	CTGGCTGTATGA	1278	GAACTAGTCACC
Water Control			3	GTGACCTGATGT	15	TCAGGACTGTGT	243	GCACACACGTTA	168	CTGTGTTCAGGA	121	CGTCTCAGAACA	1551	GTCACCGAACTA
Water Control			50	TCAGCCATGACA	15	TTAGGTGCAGCT	272	TTCCTAGGTGAG	394	ACGCAACTGCTA	180	TCACAGATCCGA	170	TGCGCTTGGATA

Supplementary Table 2-2. Samples studied from human stool.

Primer	Amplicon	Sequence
18S_0067a_deg	18S	AAGCCATGCATGYCTAAGTATMA
NSR 399	18S	TCTCAGGCTCCYTCTCCGG
ITS1F	ITS1	CTTGGTCATTTAGAGGAAGTAA
ITS2	ITS1	GCTGCGTTCTTCATCGATGC

Supplementary Table 2-3. Oligonucleotides used in this chapter.

BROCC	MARTA	MEGAN
Votes after % id filter, can consider only 1 hit	Votes sometimes, or picks top hit	Lowest Common Ancestor
Uses percent identity	Uses bitscore	Considers hits by bitscore filter and top % of bit score filter
Filters for idenity, coverage, and classification level	Filters for coverage and idenity	Filters for bitscore
Different identity thresholds at different levels	Iterates through hits by bitscore	Needs full consensus
Adjustable coverage cutoff	Hardcoded 80% coverage filter	n/a
Excludes generic db classifications	Excludes all "uncultured" or "unidentified"	Includes generic db hits in decision
Begins voting at bottom of taxonomy	Begins voting at bottom of taxonomy	n/a
Can use any blast implementation	Megablast is Hardcoded	Can use any blast implementation
More informative unclassified messages	Cryptic unclassified messages	n/a
Queries NCBI online for taxonomy	Queries SQL database for taxonomy	Uses NCBI taxonomy in program
Implemented in Python	Implemented in Java	Implemented in Java
Can integrate into Qiime	Requires intervention for use with Qiime	Output usable by Qiime if names match, runs in GUI

Supplementary Table 2-4. Comparison of BROCC, MARTA, and MEGAN programs.

CHAPTER 3: Protocol to Characterize Archaea in the Microbiome through Sequencing

3.1 Abstract

Prior to this work, efficient methods to characterize archaeal communities in the gut microbiome by deep sequencing were not available. Here we surveyed primers previously used for archaeal community characterization in environmental samples to identify primers that both exclude host DNA and detect archaea. Additionally we refined wet side DNA extraction methods and thermocycler conditions to improve detection with successfully validated primers. Ultimately, we present a working protocol for characterizing the archaeal community composition in the gut microbiome.

3.2 Introduction

The archaea consist of a diverse group of prokaryotes that are phylogenetically closer to eukaryotes than bacteria[1]. The archaea inhabit most settings on Earth including environments with extreme heat[2], cold[3], salinity[4], and high[5] and low[6] pH, in addition to more temperate locales[7]. The archaeal ability to live in extreme environments and phylogenetic distinctiveness has lead to the evolution of characteristic cell wall, cell membrane, and surface structures[8], along with metabolic capabilities that are unique to the archaea, such as methanogenesis[9] and synthesis of distinctive lipids[10].

Due to the archaea's genetic and biochemical distinctiveness, much interest has been generated around detecting and characterizing archaea in the human microbiome, since archaea would be potentially able to fill a unique niche in the microbiome community[11] [12]. Prior to this dissertation, *Methanobrevibacter smithii* was acknowledged as the dominant commensal archaeon in the gut microbiome[13,14] and present in most individuals[15]. *Methanospheara*

stadtmanae has also been documented to occur in some individuals [16]. Here we develop methods to characterize the archaeal community through sequencing.

3.3 Materials and Methods

Samples and DNA Extraction

In the preliminary studies the initial macaque stool samples were extracted with the Qiagen Stool kit as per the manufacturer's instructions and were described in detail previously [17]. 4 of the samples (AM40, AV86, MIT4, and C2T64) came from SIV positive macaques. The remaining 4 samples (DE20, C2T3, C2T10, and C4T1) came from uninfected macaques. All the murine samples came from healthy mice. DNA from these samples was extracted with the same method used for the macaques. 4 of the samples (11F, 12B, 13B, and 15D) were described previously [18], and 3 (IKK2, IKK4, and IKK5) were produced for another unpublished study.

The human samples used for later study were derived from self reported healthy adults and extracted with the PSP kit modified for difficult to lyse cell wall. The stool samples[19,20] and their extraction[21] were described previously.

Primer Selection and Modification

We formed the initial set of tested primers through a search of the literature [22,23,24,25]. The 958af primer was modified after a visual inspection of alignment to several *Solfulobus* and related 16S sequences from the GenBank[26] database and the UCSC Genome Browser[27]. The 17th base pair made degenerate from G to K (G or T).

Amplification Protocol

Initial test amplifications were preformed with the GreenTaq system. Reactions contained 10µL genomic DNA diluted to 5 ng/µL, 2 µL of each primers diluted to 20 pmol/ µL,

4.3 µL Greentaq Mix, 5.7 µL water, .5 µL BSA (10mg/mL), and 2.5 µL Triton (1%). We used the following thermocycles protocol: denaturing at 95°C for 5 minutes, 25 cycles of amplification with 30 seconds for denaturing at 95°C, 30 seconds of annealing at 50°C or 55°C, and 72°C for 90 seconds. At the end of the program, there was an 8 minute extension at 72°C.

Later amplifications used the Accuprime system according the manufacturer's instructions. Thermocycler conditions were as follows: initial denaturing for 5 minutes at 95°C, then 47 cycles of amplification with 30 seconds of denaturing at 95, 30 seconds of annealing at variable temperatures, and 90 seconds of extension at 72°C. The annealing temperature started at 67°C and decreased by 1°C each cycle until it reached 55°C, where it remained for 10 cycles. Next, it decreased to 53°C for 5 cycles, then decreased to 50°C for 20 more cycles. Afterwards there was an 8 minute extension at 72°C.

Sequencing and Bioinformatics

Amplified DNA was cloned with the Invitrogen TOPO4 and TOPOXL cloning kits according the manufacturer's instructions. Samples were sequenced on an ABI 3730 DNA analyzer. Sequences were classified the web based RDP classifier tool [28]. Classifications were checked with NCBI's BLAST and the RDP SeqMatch tool.

3.4 Results

Primer Selection

Several studies have used the archaeal 16S rRNA gene to characterize and analyze archaeal communities in environmental samples from seawater[29,30], soil[31,32], and others[33], prior to this work. Through a literature search, we collected 6 noted archaeal specific 16S rRNA primers (Table 3-1) [22,23,24] and paired each one with the universal 1378ar reverse

primer[25], which amplifies both archaea and eukaryotes. We used this scheme to test the specificity of individual forward primers.

Name	Sequence 5'-3'	Source
4aF	TCCGGTTGATCCTGCCRG	Grobowski 2005
398aF	CCRGGCCCTACGGGG	Baker 2003
571aF	GCYTAAAGWRDCCGTAGC	Baker 2003
1040aF	GAGAGGWGGTGCATGGCC	Baker 2003
958aF	AATTGGAKTCAACGCCGGR	Collins 2005
1204aF	AGGTMBGYATGCCCCKAA	Baker 2003
1378aR	TGTGTGCAAGGAGCAGGGAC	Lepp 2004

Table 3-1. Primer sequences used in this chapter.

Primers pairs were queried against a set of extracted stool samples from healthy mice and healthy and SIV infected macaques extracted with the Qiagen Stool Kit. Samples were initially amplified using 25 PCR cycles at annealing temperatures of 50°C and 55°C, based on a previously described procedure for amplifying 16S[17]. Amplifications with annealing temperatures at 50°C generally produced brighter bands than the 55°C amplification but consisted of nonspecific amplification, so amplifications at 55°C were considered.

In initial tests, two of the forward primers, 571af and 1204af, produced no bands and were not considered. Two more forward primers, 398af and 1040af, produced visible bands, but bands were determined to be off target eukaryotic host sequences upon Sanger sequencing. The remaining forward primers, 4af and 958af, produced bands in a limited number of samples (Fig 3-1), which were verified with Sanger sequencing. All reads produced from macaque samples were classified as *Methanobrevibacter* with the RDP classifier. All but one read produced from the murine samples were classified as *Methanobrevibacter*. The remaining murine derived sequence was classified with 79% confidence as belonging to the extremophile genus *Thermogymnomonas*, which prior to this work has been observed in hot springs[34]. The murine samples only

produced visible amplification with the 4af forward primer, and the macaque samples only produced visible amplification with the 958af forward primer. The reason for this is unclear, since k-nearest neighbor analysis with the RDP classifier tool made taxonomic assignments for reads from both the macaque and mouse samples to archaeal species that have homology to both the 4af and 958af primers.

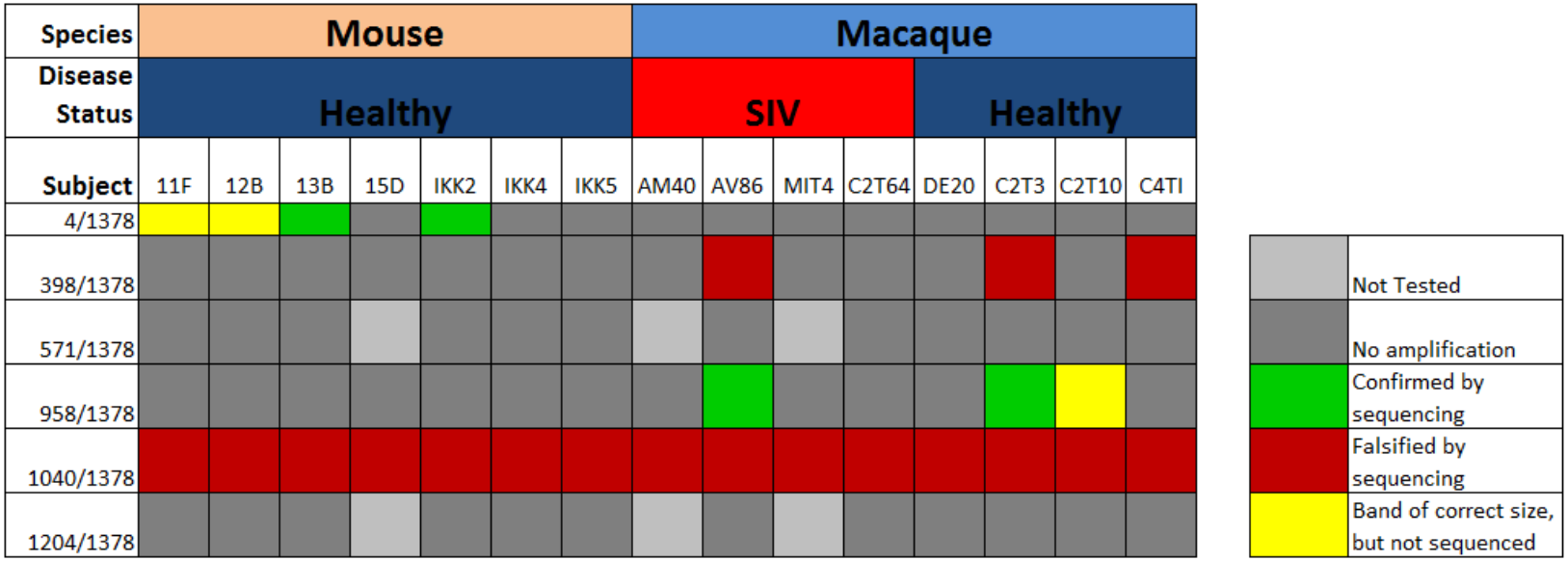


Figure 3- 1. Amplification of different primers across macaque and murine stool samples. Species is indicated by the top row. Disease state is indicated by the second row. Identity of the individual from which the stool sample was obtained is indicated by the third row. Primer pairs are indicated in the leftmost column. Results from testing each primer pair over each sample is indicated by the colors of chart boxes. Amplifications were preformed with Greentaq and a standard 25 cycle thermocycler program on murine and macaque samples extracted with the Qiagen Stool kit.

Assay Refinement

Upon assessment of sequencing results, we selected the 958aF/1378aR for further development, due its small amplicon size (~450 base pairs), which was suitable for adaption for 454 sequencing. Through endpoint PCR testing of genomic DNA derived from pure culture *Sulfolobus solfataricus*, we observed that the 958aF/1378aR primer pair does not produce visible amplification on an ethidium bromide stained gel. Bioinformatic analysis of database sequences revealed that this effect was caused by a point mutation 2 base pairs from the 3' end of the 958aF primer. This mutation occurs in species spanning the entire *Sulfolobales* order. Because we were unsure if any members *Sulfolobales* reside in the gut microbiome, we decided to modify the 958aF primer by making the 17th base pair degenerate from G to K (G or T). This new primer was renamed 958aF-deg.

Despite amplifying target archaeon sequences with the previously described methods, some of the sequenced data from the 958aF/1378aR primer pair resulted from nonspecific amplification. We implemented a complex touchdown thermocycler protocol, which reduced the presence of off target sequences (see methods). Also, after seeing an increase in amplification performance in fungal detection assays after surveying different extraction methods[21], we used to the PSP extraction method modified for difficult-to-lyse cells (see methods) and saw a corresponding increase in performance in endpoint PCR, which was unsurprising because archaeal cell walls have been documented as relatively hard to lyse [15]. Also we changed our PCR cocktail to use Accuprime taq, which has been documented to perform well with difficult-to-amplify samples (unpublished observations), and saw a dramatic increase in amplification in endpoint PCR. With this improved extraction and amplification procedure, we were able to detect archaea in a higher proportion of human samples. Limited sequencing was preformed with the Sanger method on 8 human samples extracted with the PSP kit (Fig 3-2). All reads from human

samples were attributed to *Methanobrevibacter*. Off target amplification of the host genome was not observed for these samples.

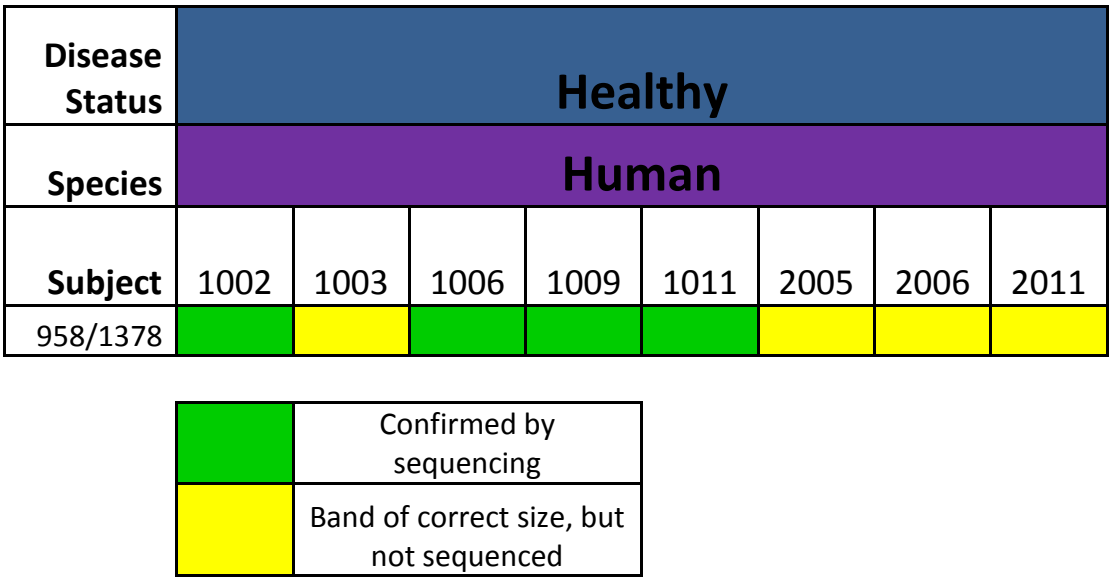


Figure 3-2 Amplification of human stool samples over the 958aF/1378aR primer pair. All samples were obtained from healthy adults and extracted with the PSP kit with bead beating. Designation of the individual from which the stool sample was obtained is indicated by the second row. Amplification results are indicated by the colors of chart boxes. Amplifications were preformed with Accuprime and a double touchdown 47 cycle thermocycler program.

3.5 Discussion

Prior to this work, *M. smithii* had been documented to live in the human gut, but genomic assays for detection of archaea more broadly in the human microbiome were not developed. This work demonstrates the wide spread existence of archaea in the gut and lays a foundation for characterization of the archaeal microbiome community. While *Methanobrevibactor* was the predominant species detected through sequencing, *Thermogymnomonas* was detected as well, suggesting multiple archaeal species inhabit the gut. The method described here was further

developed for 454 sequencing by Christian Hoffmann and was successfully applied to a large set of human stool samples, where several divergent lineages were detected, including *Methanobrevibactor*, *Nitrososphaera*, *Methanospheara*, *Thermoplasmata*, and *Thermogymnomonas*. While the archaea compose a minority component of the microbiome, further studies are needed to determine how their unique metabolic capabilities change community function and dynamics.

3.6 Acknowledgements

I would like to thank members of the Bushman lab for their help suggestion and especially acknowledge Christian Hoffmann for his invaluable instruction and advice, which made this project successful.

3.7 References

1. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences of the United States of America 87: 4576-4579.

2. Takai K, Sako Y (1999) A molecular view of archaeal diversity in marine and terrestrial hot water environments. FEMS Microbiology Ecology 28: 177-188.

3. Cavicchioli R (2006) Cold-adapted archaea. Nature reviews Microbiology 4: 331-343.

4. Dennis PP, Shimmin LC (1997) Evolutionary divergence and salinity-mediated selection in halophilic archaea. Microbiology and molecular biology reviews : MMBR 61: 90-104.

5. Rees HC, Grant WD, Jones BE, Heaphy S (2004) Diversity of Kenyan soda lake alkaliphiles assessed by molecular methods. Extremophiles : life under extreme conditions 8: 63-71.

6. Schleper C, Puehler G, Holz I, Gambacorta A, Janekovic D, et al. (1995) Picrophilus gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0. Journal of bacteriology 177: 7050-7059.

7. Abreu C, Jurgens G, De Marco P, Saano A, Bordalo AA (2001) Crenarchaeota and Euryarchaeota in temperate estuarine sediments. Journal of applied microbiology 90: 713-718.

8. Albers SV, Meyer BH (2011) The archaeal cell envelope. Nature reviews Microbiology 9: 414-426.

9. Bapteste E, Brochier C, Boucher Y (2005) Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. Archaea 1: 353-363.

10. Koga Y, Morii H (2006) Special methods for the analysis of ether lipid structure and metabolism in archaea. Analytical biochemistry 348: 1-14.

11. Dridi B, Raoult D, Drancourt M (2011) Archaea as emerging organisms in complex human microbiomes. Anaerobe 17: 56-63.

12. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-1031.

13. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science (New York, NY) 312: 1355-1359.

14. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635-1638.

15. Dridi B, Henry M, El Khechine A, Raoult D, Drancourt M (2009) High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. PLoS ONE 4: e7063.

16. Jones BV, Begley M, Hill C, Gahan CG, Marchesi JR (2008) Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proceedings of the National Academy of Sciences of the United States of America 105: 13580-13585.

17. McKenna P, Hoffmann C, Minkah N, Aye PP, Lackner A, et al. (2008) The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. PLoS pathogens 4: e20.

18. Hoffmann C, Hill DA, Minkah N, Kirn T, Troy A, et al. (2009) Community-wide response of gut microbiota to enteropathogenic Citrobacter infection revealed by deep sequencing. Infection and immunity.

19. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, et al. (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiol 10: 206.

20. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334: 105-108.

21. Dollive S, Peterfreund GL, Sherrill-Mix S, Bittinger K, Sinha R, et al. (2012) A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. Genome biology 13: R60.

22. Grabowski A, Nercessian O, Fayolle F, Blanchet D, Jeanthon C (2005) Microbial diversity in production waters of a low-temperature biodegraded oil reservoir. FEMS Microbiology Ecology 54: 427-443.

23. Baker GC, Smith JJ, Cowan DA (2003) Review and re-analysis of domain-specific 16S primers. Journal of microbiological methods 55: 541-555.

24. Collins G, O'Connor L, Mahony T, Gieseke A, de Beer D, et al. (2005) Distribution, localization, and phylogeny of abundant populations of Crenarchaeota in anaerobic granular sludge. Applied and environmental microbiology 71: 7523-7527.

25. Lepp PW, Brinig MM, Ouverney CC, Palm K, Armitage GC, et al. (2004) Methanogenic Archaea and human periodontal disease. Proc Natl Acad Sci U S A 101: 6176-6181.

26. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) GenBank. Nucleic acids research 33: D34-38.

27. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome research 12: 996-1006.

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

29. van der Maarel MJ, Sprenger W, Haanstra R, Forney LJ (1999) Detection of methanogenic archaea in seawater particles and the digestive tract of a marine fish species. FEMS microbiology letters 173: 189-194.

30. Teske A, Sorensen KB (2008) Uncultured archaea in deep marine subsurface sediments: have we caught them all? The ISME journal 2: 3-18.

31. Bintrim SB, Donohue TJ, Handelsman J, Roberts GP, Goodman RM (1997) Molecular phylogeny of Archaea from soil. Proceedings of the National Academy of Sciences of the United States of America 94: 277-282.

32. Nicol GW, Leininger S, Schleper C, Prosser JI (2008) The influence of soil pH on the diversity, abundance and transcriptional activity of ammonia oxidizing archaea and bacteria. Environmental microbiology 10: 2966-2978.

33. Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, et al. (2006) Lineages of acidophilic archaea revealed by community genomic analysis. Science 314: 1933-1935.

34. Itoh T, Yoshikawa N, Takashina T (2007) Thermogymnomonas acidicola gen. nov., sp. nov., a novel thermoacidophilic, cell wall-less archaeon in the order Thermoplasmatales, isolated from a solfataric soil in Hakone, Japan. International journal of systematic and evolutionary microbiology 57: 2557-2561.

CHAPTER 4: Perturbation of Fungi and Bacteria through Antibiotics in the Murine Gut

The contents of this chapter have been submitted for publication in:

> Dollive S, Chen YY, Grunberg S, Bittinger K, Vandiver L, Cuff C, Lewis JD, Wu GD, Bushman FD. (2013). Fungi of the murine gut: episodic variation and proliferation during antibiotic treatment.

4.1 Abstract

Antibiotic use in humans has been associated with outgrowth of fungi. Here we investigated the gut microbiome over 76 days of antibiotic treatment and recovery using a controlled mouse model. Mouse stool was studied a surrogate for the microbiota of the lower gastrointestinal tract. The abundance of fungi and bacteria was measured using quantitative PCR, and the proportional composition of the communities quantified using 454/Roche pyrosequencing of rRNA gene tags. Bacteria dropped >3 orders of magnitude after initiating treatment, so that the predominant 16S sequences detectable during treatment were transients derived from food. Upon cessation of treatment, bacteria mostly returned to their previous numbers and types after 8 weeks, though communities remained detectably different from untreated controls. Fungal communities varied substantially over time, even in the untreated controls. Separate cages within the same treatment group showed radical differences, but mice within a cage generally behaved similarly. Fungi increased 40-fold in abundance upon antibiotic treatment but declined back to their original abundance after cessation of treatment. At the last time point, *Candida* remained more abundant than prior to treatment. These data show that 1) gut fungal populations change radically during normal mouse husbandry, 2) fungi grow out in the gut upon suppression of bacterial communities with antibiotics, and 3) perturbations due to antibiotics persist long term in both the fungal and bacterial microbiota.

4.2 Introduction

The gut microbiome is generally stable but can be changed through exterior perturbation. Perturbation can be introduced through changes in diet [1], changes in the immune system [2], presence of a pathogen [3], and use of probiotics [4] or antibiotics [5], and others factors. Antibiotic use is perhaps the most common and dramatic source of change in the microbiome community and is a subject of clinical interest.

The effects of antibiotic use on the human microbiome can be challenging to clarify fully--confounding factors include complications of the underlying diseases states and concomitant use of additional forms of therapy[6]. Despite these difficulties, outgrowth of fungi has been repeatedly linked to antibiotic treatment at body sites including the gut[7,8], vagina[9], mouth[10], skin[11] and others [12,13]. Fungal infection associated with antibiotic use is of particular concern in immunocompromised states such as HIV/AIDS[14,15,16], some cancers[13,17,18], and transplantation[19,20,21,22] [23]. Many of these conditions necessitate the use of corticosteroids, which further predisposes the host to fungal infection[24]. Invasive fungal infections have been increasing in recent decades[6,17], and the rise of azole-resistant species of *Candida*[12,25], *Aspergillus*[26,27], and *Cryptococcus*[14,28] brings further urgency to understanding the interaction between commensal fungi and bacteria under antibiotic treatment.

Rodent models have been used to study the effects of antibiotics on the mammalian gut, using culture based[29,30], metagenomic[31], and immunologic[31,32] methods. Antibiotic treatment can predispose the host to infection by pathogens[33,34] and alter microbial communities long term[34]. Induced exposure to *Candida albicans* shapes the bacterial composition of the murine gut during antibiotic recovery[30] and can cause gastritis[35], while *Candida tropicalis* has been associated with increased severity in ulcerative colitis[36].

Phenotypic effects have been found even after treatment with subclinical doses of antibiotics [37]. In studies of the role of the vertebrate microbiome in mice, antibiotic treatment is often used to suppress the host bacteria, but the effect of this intervention on fungi is not commonly considered [33,38,39,40].

Here we characterize the bacterial and fungal microbiota of mice during antibiotic treatment and subsequent recovery after cessation. We analyzed abundance using quantitative PCR (henceforth "QPCR"), and analyzed the types present using 454/Roche pyrosequencing of rRNA gene tags. We found that fungi indeed grew out upon antibiotic treatment. After cessation of antibiotic treatment, fungal and bacterial communities approached their pre-antibiotic states, but increased abundance of *Candida* persisted in the gut at the last time point studied eight weeks later. To our surprise, we also found that the fungal communities changed radically over time in both control and treated mice. For each condition, specific fungi colonized multiple mice in the same cage, then gave way to subsequent fungal colonists over time, and different patterns were seen in different cages.

4.3 Material and Methods

Ethics Statement

The Institutional Review Board of the University of Pennsylvania approved all IACUC protocols (protocol #803408). The animal care facility is operated by the University Laboratory Animal Resources, which is fully accredited by the American Association for Accreditation of Laboratory Animal Care. Laboratory animals are maintained in accordance with the applicable portions of the Animal Welfare Act and their guidelines prescribed in the DHHS publication, "Guide for the Care and Use of Laboratory Animals". Mice are euthanized using $C0_2$ inhalation. Methods for

euthanasia are consistent with the recommendations of the Panel on Euthanasia of the American Veterinary Medical Association.

Mouse Husbandry

Thirty C57B6 eight week old female mice were purchased from Jackson Laboratory and placed on a standardized diet for two weeks prior to the study to stabilize their flora. Mice were housed by treatment group with five mice per cage. Mice treated with antibiotics received ampicillin, neomycin, vancomycin, and metronidazole in water. Water was spiked with aspartame in both the treated groups and controls. Mice were fed AIN-76A Rodent Diet from Research Diet (D10001) for the course of the study, which includes 15% casein lactic.

DNA Extraction

DNA was extracted from 1-2 mouse pellets per mouse per time point. Samples were homogenized for 80 seconds on a Mini-BeadBeater-16 (BioSpec) in Lysing Matrix E tubes (MP Biomedical). Samples were then incubated at 95°C for 15 minutes and then cooled on ice for 60 seconds. Then samples were extracted with the PSP DNA extraction kit and using a protocol described previously[41]. Sequences of oligonucleotides used in this study are presented in Supplementary Table 4-1.

DNA Sequencing and QPCR

Extracted DNA was quantified with the Picogreen system. Typical volumes produced, used in the calculations in Supplementary Table 4-2, were 250 microliters. DNA was amplified using primers annealing to the V1V2 region of the 16S bacterial gene or the ITS1 fungal rRNA gene spacer, and amplified with AccuPrime taq with Buffer 2 (Invitrogen). Thermocycler protocols for 16S[42] and ITS and 18S[41] amplicons were described previously. PCR amplicons were

purified with Agencourt AMPure XP beads. Sequencing was performed on a 454 Junior using Titanium chemistry. For both 16S and ITS amplicons, DNA free water was subjected to the same purification procedure and analyzed by 454/Roche pyrosequencing (Supplementary Figure 4-6). A subset of samples showed recoverable sequences, but stool samples showed distinct community composition, so conclude that environmental contamination made a minimal contribution to the samples analyzed.

16S qPCR was performed using the Taqman method as described previously[31]. 18S qPCR was performed with the Applied Biosystems SYBRGreen Fast chemistry. Ten µL SYBRGreen FAST 2X master mix, 1 µL of each primer diluted to 20 pM, and 8 µL DNA were added to each reaction. Thermocycling was performed as follows: samples were initially denatured once for 2 minutes at 50°C then 10 minutes at 95°C. Then samples were cycled 40 times with a dissociation step at 95°C for 15 seconds and an annealing and extension step at 60°C for 1 minute. Primer sequences for these assays can be found in Supplementary Table 4-1. All DNA sequences generated in this study have been submitted to the Sequence Read Archive.

Bioinformatic Analysis

Bioinformatic analysis was performed with the QIIME software package[43] using default parameters except where indicated, and using R. Fungal reads were queried against the nt database using NCBI's blastn tool and then classified with the BROCC classifier[41].

Eukaryotic PCoA analysis was performed using taxonomic relationship corresponding to the NCBI Taxonomy[44]. Because de novo tree construction using the ITS1 region is not feasible due to length variation inherent in the ITS gene[41,45,46,47], we chose to asses Unifrac distances between eukaryotic communities using the NCBI Taxonomy to generate taxonomic trees. To transform the taxonomy into a phylogenetic tree, all edges between taxa were assigned equal

weight. Classifications were curated manually for parsimony. While the fungal taxonomy is imperfect and in a state flux[48,49,50,51], we note that in practice the Unifrac metric is relatively robust to the method used in creating phylogenetic trees[52]. Statistical significance for treatment groups was determined using the nonparametric Mann-Whitney test in R. Permanova tests and Procrustes analysis were performed in QIIME.

4.4 Results

Longitudinal analysis of the murine gut during antibiotic treatment

To assess the relationship between bacterial and fungal lineages during antibiotic treatment, an antibiotic cocktail containing vancomycin, ampicillin, neomycin, and metronidazole was given to twenty C57B6 mice in water. After 2 weeks, antibiotic treatment was stopped for ten of the mice. These mice did not receive any antibiotics during the remaining nine weeks of the study ("AbxShortTerm" mice). The remaining ten mice under antibiotic treatment continued to receive antibiotics for the duration of the study ("AbxContinuous" mice). In parallel, ten control mice received no antibiotics over the course of the study ("Control" mice). Fecal samples were collected over one week prior to initiating the study, then at the indicated time points during the study (Figure 4-1). Mouse husbandry and sample collection was performed by Ying-Yu Chen. DNA was purified from stool pellets using a procedure that included bead beating and a high temperature incubation to facilitate lysis of fungal cells[42] by Stephanie Grunberg and Serena Dollive.

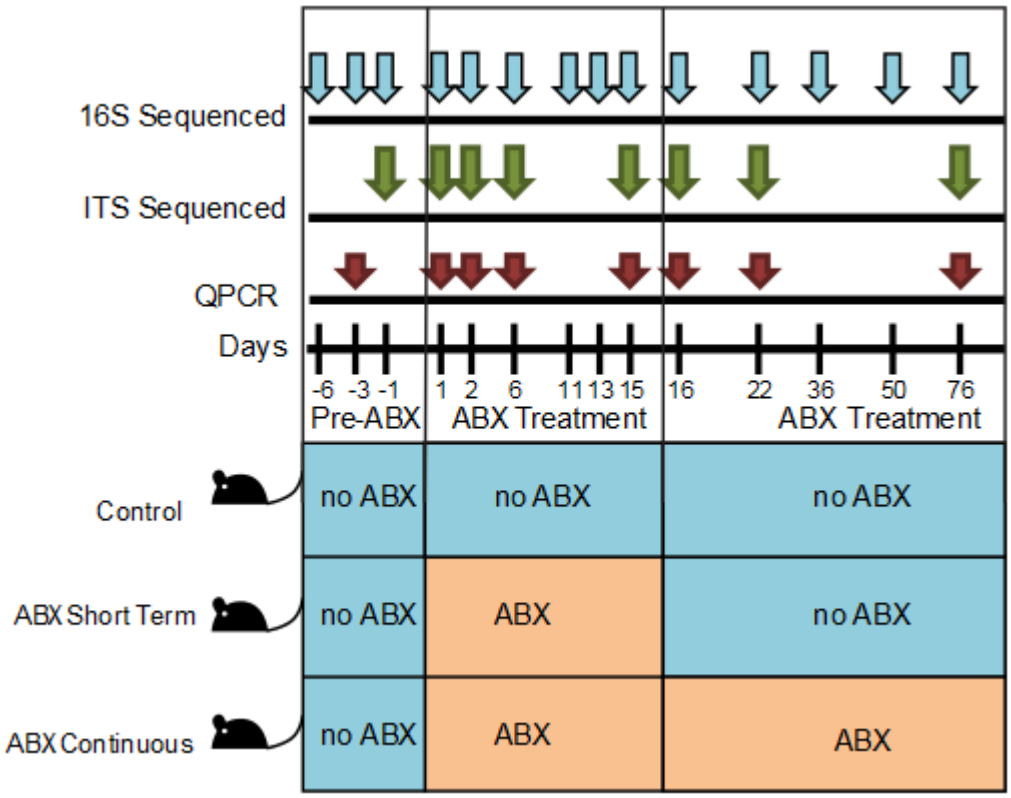


Figure 4-1. Experimental Diagram. The time line for the 76 days of sample collection is shown along the top, and the periods of antibiotic treatment are shown at the bottom. Antibiotic treatment was initiated at time zero.

Analysis of the numbers of bacterial 16S and eukaryotic 18S gene copies present after antibiotic treatment

We first investigated the changes in abundance of bacteria and fungi, using stool specimens as a proxy for the lower intestinal microbiome. To assess changes in abundance, we first quantified the abundance of bacterial and fungal genomes in the samples per ng of DNA using quantitative PCR. For bacteria, a QPCR assay was used that detected the bacterial 16S rRNA gene, and for fungi, an assay was used detecting the 18S rRNA gene. The primers for the fungal assay were designed to suppress amplification of metazoan DNA originating from the host

or food materials [53]. The specificity was confirmed by pyrosequencing products of amplification with these primers (described below and Supplementary Figure 4-1).

At the start of the study, fecal pellets contained $5X10^5$ to 10^6 copies of bacterial 16S rRNA genes per ng DNA (Figure 4-2). After initiation of antibiotic treatment, this fell as low as 100 copies per ng DNA, or a drop of >3 orders of magnitude. Upon cessation of antibiotic treatment the community recovered to its former high numbers. Thus we conclude that the antibiotic treatment was highly effective at reducing the numbers of bacteria present in gut, as has been seen in many studies (e. g. [34] [54] [55]), and that the community was sufficiently resilient to return to its former size after cessation of treatment.

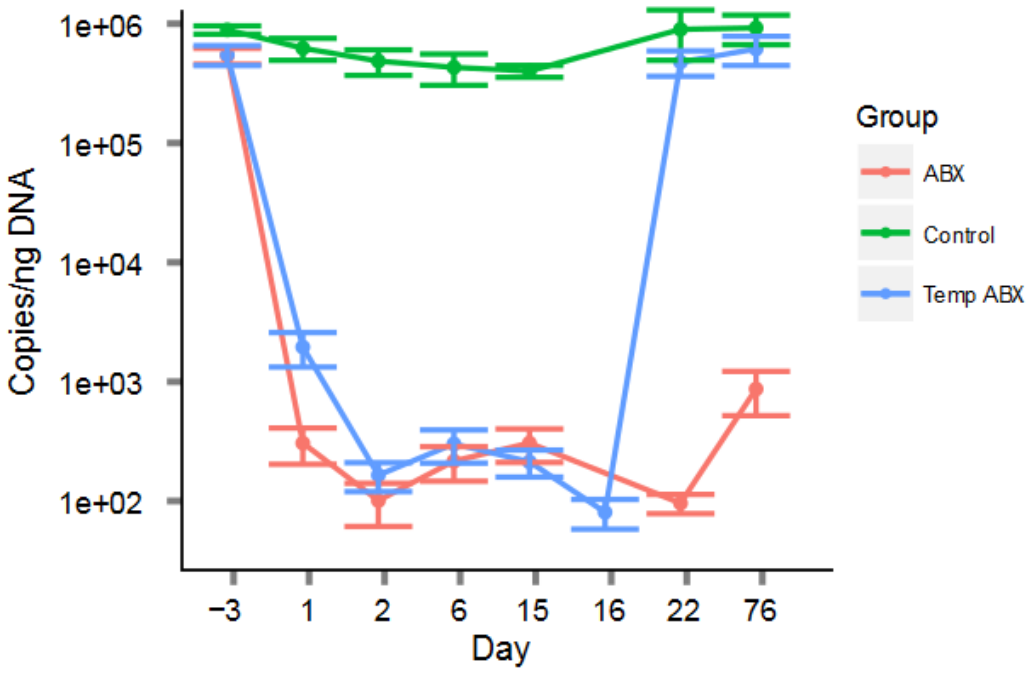


Figure 4-2. Longitudinal analysis of 16S rRNA gene copies per ng of stool DNA. The groups of mice tested are shown by the color code (key at right). Error bars indicate standard error.

Prior to initiation of antibiotic treatment, 1-3X10^5 18S rRNA gene copies were detected per ng DNA (Figure 4-3). Upon initiation of antibiotic treatment the number climbed two orders of magnitude, ranging between 3-6X10^8 copies per ng DNA. Upon cessation of antibiotic treatment, the numbers dropped back to roughly their former levels, thus displaying a favoring of smaller community size in the absence of antibiotic pressure.

The abundance of fungi in the Control group showed an unexpected increase at day 22. Further analysis showed that the increase was in only one of the two cages housing the control animals, and correlated with the appearance of a new fungal lineage at high levels in all animals in that cage (described below).

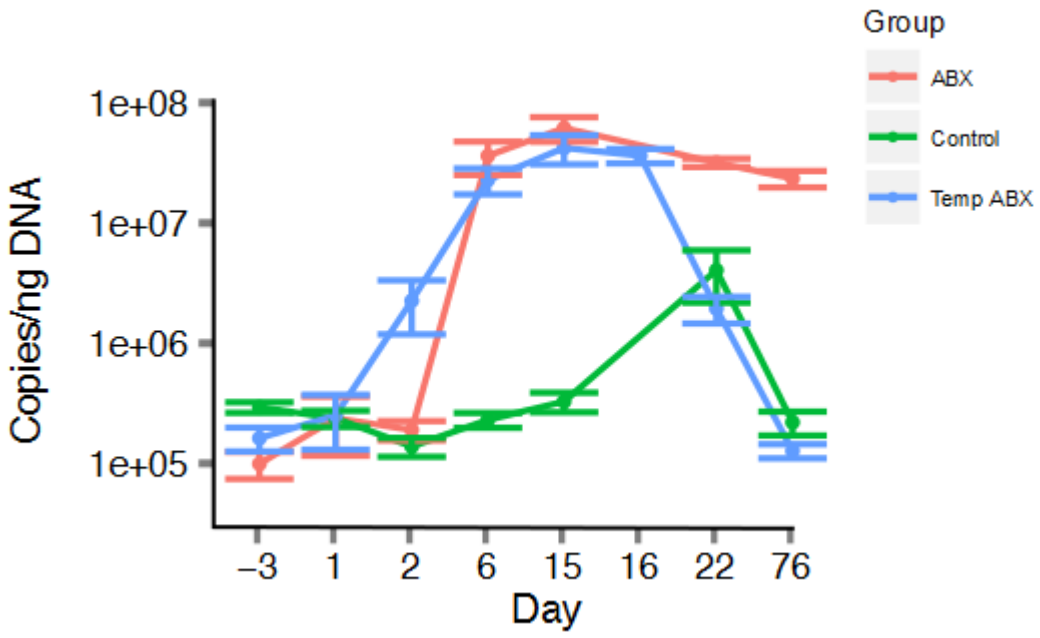


Figure 4-3. Longitudinal analysis of 18S rRNA gene copies per ng of stool DNA. The groups of mice tested are shown by the color code (key at right). Error bars indicate standard error. The amplicon used was designed to suppress amplification of DNA from mouse or food materials.

Additional information is required to relate these numbers of rRNA gene copies to the numbers of organisms present. This issue is addressed in the next section.

Assessing the absolute abundance of bacteria and fungi

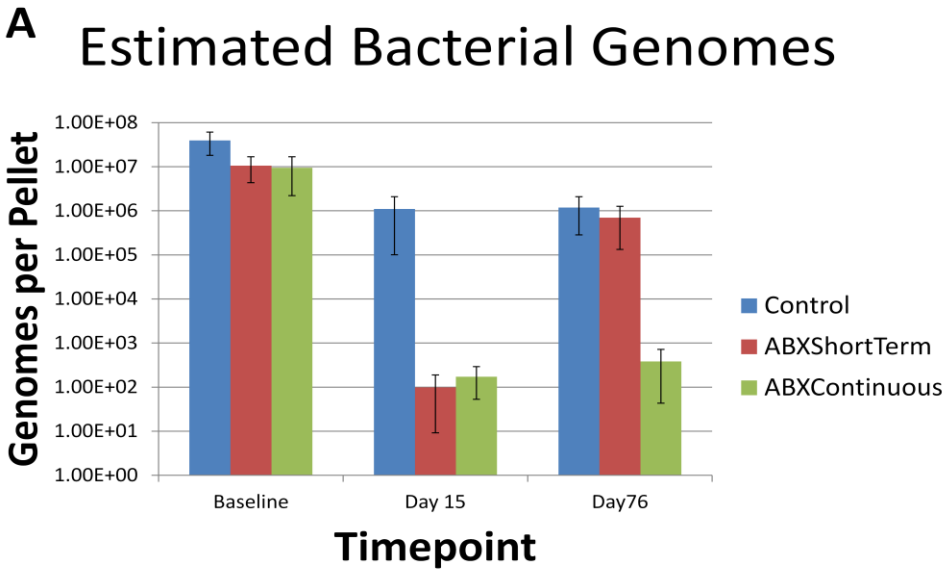
Several corrections are required to link the QPCR data to the total number of organisms per stool pellet. One consideration is that bacterial [56,57] and fungal [58,59] genomes typically contain multiple rRNA gene copies. From published data on complete genome sequences, we estimated the mean number of 16S rRNA gene copies per bacteria at 5 [56], and 18S copies per fungal genome at 100 [59], though the number for fungi is tentative due to the difficulty of accurately sequencing tandem direct repeats and variability in copy number.

Another concern in assessing possible fungal outgrowth during antibiotic treatment is that the total number of microbes in pellets, and thus total DNA, may go down with treatment, so that fungi could falsely appear to proliferate only because total DNA content went down as bacterial numbers fell. Thus we sought to correct the above assays, which were normalized to weight of DNA, to better reflect the counts of individual organisms by putting the final analysis on a per pellet basis. Values for inferred microbial genomes are shown in Supplementary Table 4-2, and a few ratios of interest are presented in Supplementary Table 4-3. Average pellet weights were 16.08 mg (SD=3.329) in the presence of antibiotic (n=20) and 18.64 (SD=2.685) in the control mice (n=19). Thus the difference in mean weight was small (14%), so we treated starting weights as equal below, though the difference did achieve significance (p=0.0129, Mann-Whitney U test).

DNA yields differed substantially (Supplementary Table 4-2). Quantification of yields after 15 or 76 days of antibiotic treatment showed drops of 4.7 and 5.7 fold (Supplementary Table 4-3). Thus the analysis of the numbers of microbial genomes needs to take into account the drop in total DNA. After withdrawal of antibiotic treatment (AbxShortTerm, Day 76), the total DNA

yield returned to within a factor of two of the starting value. Evidently bacterial DNA is the predominant source of DNA in mouse pellets, and the community mostly returned to its former size after cessation of antibiotic treatment.

Taking these factors into account, we find that in the absence of antibiotic treatment, a typical stool pellet contained ~10^7 bacteria, and this dropped to ~10^2 bacteria after 15 days of antibiotic treatment (Figure 4-4). Fungal genomes were much less abundant initially, only in the range of $6X10^3$ per pellet. After 15 days of treatment with antibiotics, the numbers increased to $2X10^5$, or an increase of ~30 fold. Fungal genome numbers remained high for the period of antibiotic treatment. Eight weeks after cessation of antibiotic treatment, counts in the ABXShortTerm groups returned to pretreatment level. Thus changes in fungal cell abundance were substantial, though less than suggested by the analysis in Figure 4-3, which was normalized to the total weight of DNA, because total DNA went down with antibiotic treatment. Microscopic inspection of stool specimens also suggested an increase in numbers of large cells, consistent with an increase in absolute fungal numbers (data not shown).



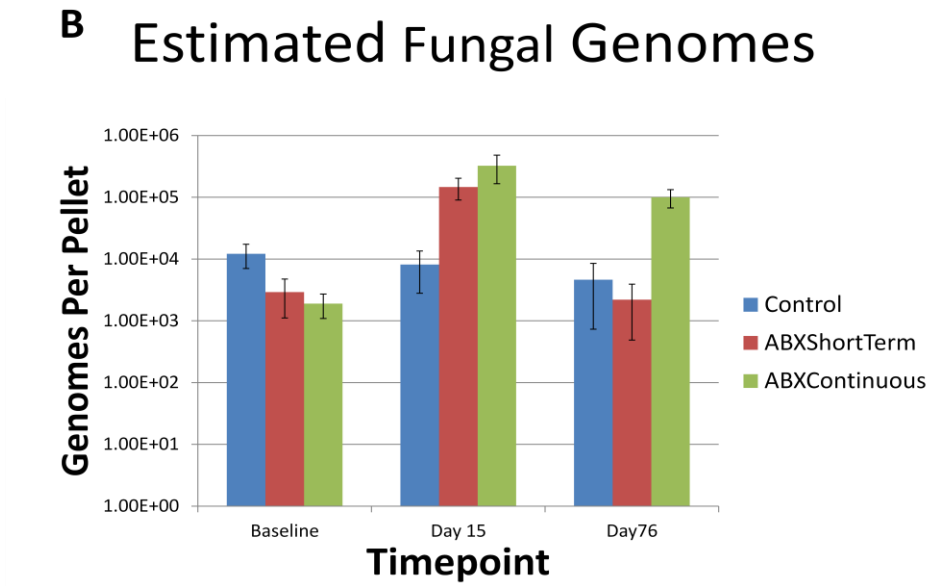


Figure 4-4. Diagrams of estimated A) Bacterial and B) Fungal genomes per mouse pellet.

Analysis of bacterial lineages using 454/Roche deep sequencing

To probe microbial dynamics under antibiotic therapy, we analyzed the longitudinal DNA samples using 454/Roche pyrosequencing. DNA was purified from stool from 13 time points (Figure 4-1). Bacterial sequences were amplified using primers matching the 16S rRNA gene V1V2 region [42,60]. Sequencing yielded 239,867 reads, which were condensed into OTUs at 97% similarity and taxonomy assigned using the RDP classifier[61].

Prior to antibiotic treatment, communities were dominated by the Firmicute lineage *Lachnospiraceae* and the Bacteriodetes lineage *Bacteroidales,* along with a substantial number of less abundant lineages (Figure 4-5; Supplementary Figure 4-2 A-I presents time points with each mouse shown individually). After one day of antibiotic treatment, the previously dominant lineages decreased sharply in abundance, and *Lactococcus* became the dominant community member. At later times under antibiotic treatment *Lactococcus* was the predominant or sole

lineage detectable. Five aliquots of sterile mouse chow were analyzed by amplification with the V1V2 primers and 454/Roche pyrosequencing, revealing that *Lactococcus* 16S rRNA genes were the predominant phylotype in chow (Supplementary Figure 4-3). We thus conclude that *Lactococcus* DNA is present in sterile mouse food, and that the antibiotic treatment eliminated the great majority of live bacteria, i. e. the 10^2 bacteria detected per pellet in Supplementary Table 4-2 represents an upper bound.

After antibiotic treatment was stopped for the ABXShortTerm group, major groups that were predominant before antibiotic treatment returned to their former levels, but at different rates. An OTU classified as *Lachnospiraceae* and several OTUs classified as *Clostridium* returned within one week. Several other clades, including *Ruminococcaceae* and other Firmicutes increased in proportion by two weeks after cessation of treatment. *Bacteroidales* did not fully return until the end of the experiment at eight weeks. *Enteroccocus*, *Escherichia*, and *Paenibacillus*, which were not dominant members of the communities in the Control or antibiotic treated groups, had elevated proportions over the recovery period but decreased in relative abundance after eight weeks off antibiotics.

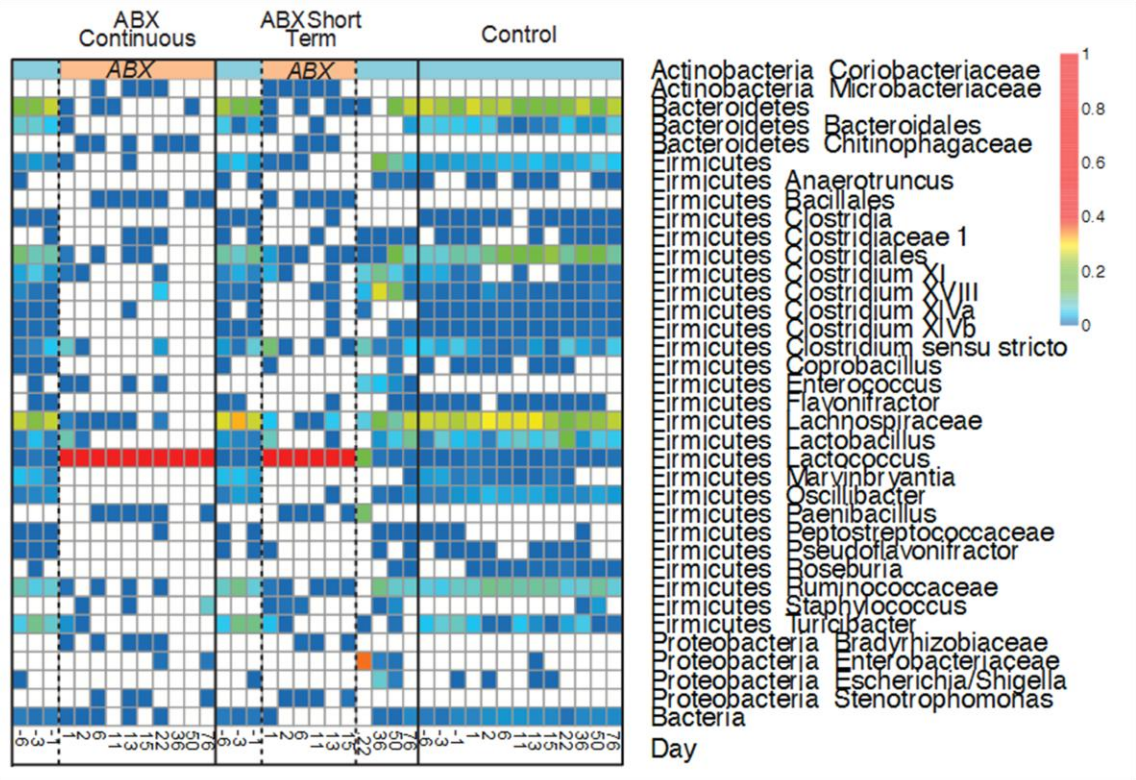


Figure 4-5. Longitudinal analysis of bacterial lineages inferred from 16S rRNA gene sequencing. Bacterial lineages detected are summarized in heat map format. Each column shows the average for the ten mice in each group and at the time point indicated. Sequence samples were rarefied to a standard 200 reads per individual before averaging. The periods of antibiotic treatment are indicated at the top in salmon color, the periods off antibiotic by light blue. The day of treatment is indicated at the bottom. The color code to the right indicates the proportions.

Changes in the types of bacterial lineages were paralleled by changes in the species richness (Figure 4-6). Prior to antibiotic treatment, 54.6 (SD=6.9) phylotypes were detected after data from each mouse was normalized to 200 reads. After 2 days of antibiotic treatment, this fell to 7 (SD=2.0) and persisted for the remainder of the antibiotic treatment. Upon cessation of antibiotic treatment, the community slowly returned to its former richness reaching 49.4 (SD=5.6) lineages over 61 days, still less than the corresponding Control group which averaged 57.2 (SD=17) lineages on the same day (p=0.02 Mann-Whitney U).

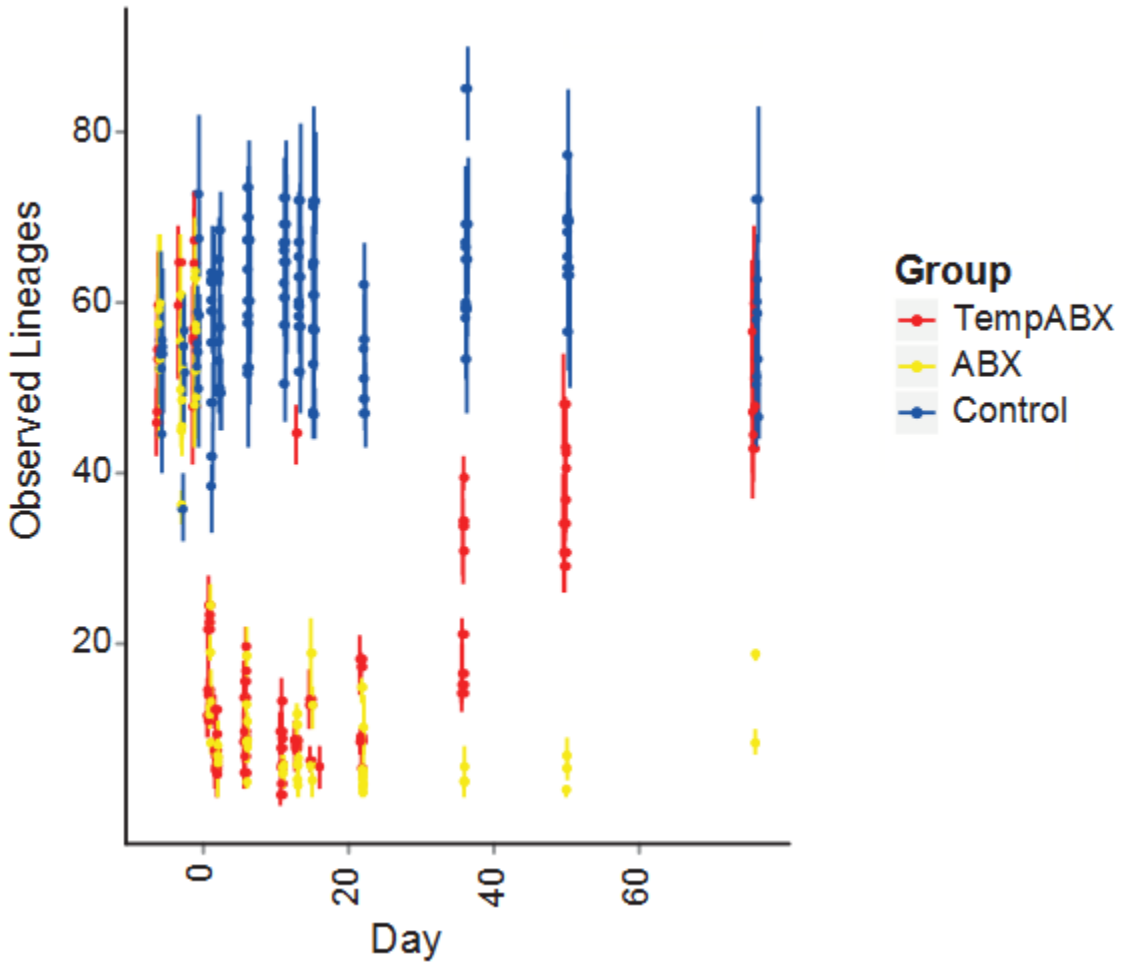


Figure 4-6. Abundance analysis of observed bacterial lineages. Each sequence set for each animal was rarefied to 200 sequences per sample 10 times, and the number of different OTUs assessed. Means are indicated by points, error bars indicate the range observed. The groups studied are indicated by the key at the right.

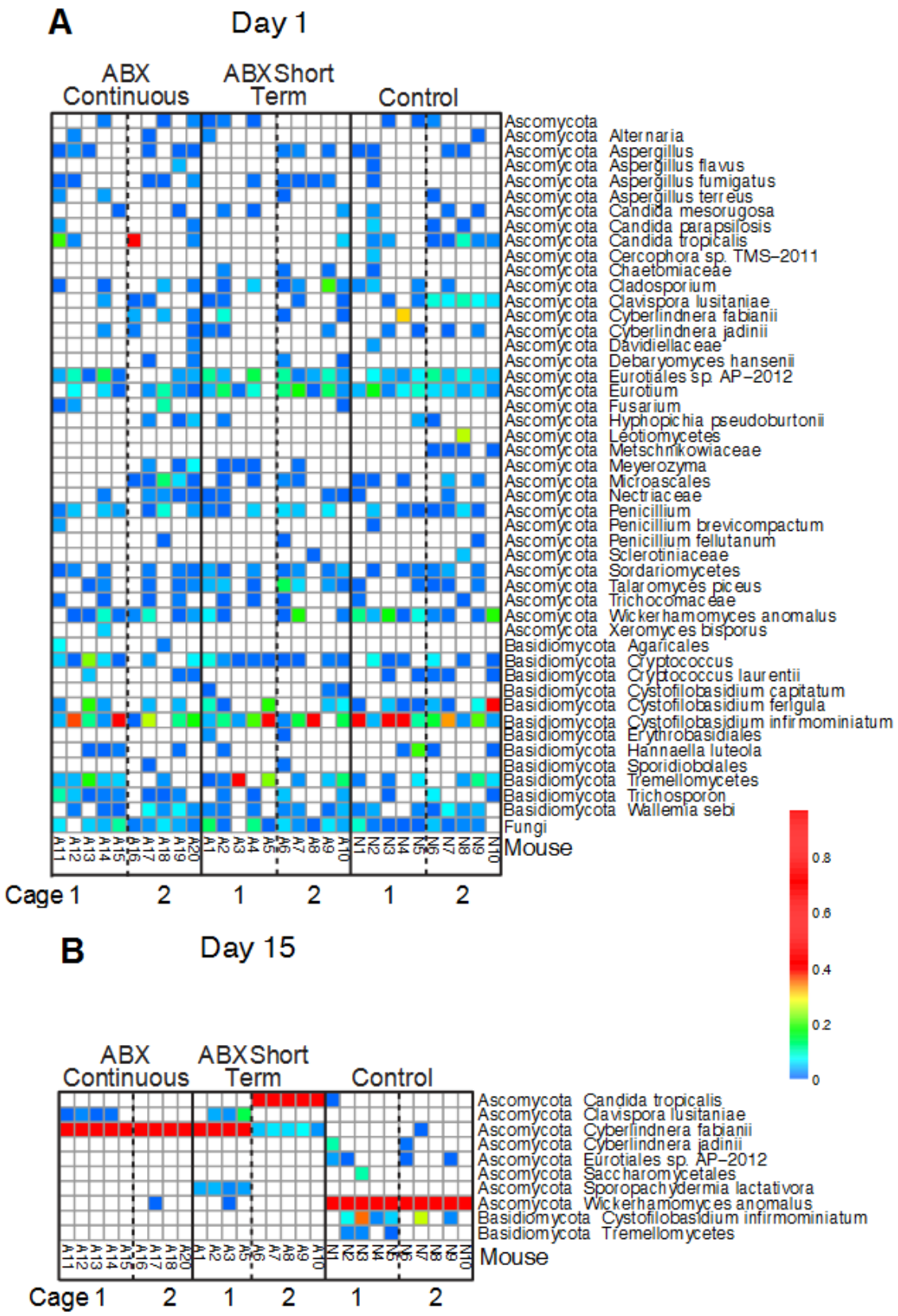
Analysis of microeukaryotes using 454/Roche deep sequencing

To characterize microeukaryotes, we sequenced selected samples using 18S and ITS amplicons. 18S sequences were prepared by Lee Vandivier, and ITS sequences were prepared by Serena Dollive. To compare samples from the different treatment groups, 134,677 ITS sequences and 26,355 18S sequences were generated, OTUs were formed, and taxonomic attribution was preformed with BROCC [41]. The 18S amplicon is more universal than the ITS amplicon [41],

while the ITS amplicon provides greater resolution for some fungal lineages [48], so both were used [53]. To check that the two amplicons were yielding consistent information, we compared sequence samples from 15 mice amplified using both amplicons. Sequence samples were characterized by generating pairwise UniFrac distances, then the distance matrices for each were compared using Procrustes analysis. This showed high correlation between the two ($p<0.0001$, no better fits after 10^4 permutations) and compositional comparison also showed similar profiles (Supplementary Figure 4-1).

The longitudinal behavior of fungal communities was explored in detail using the ITS amplicon, which revealed strong effects of both antibiotic treatment and housing history of the animals (Figure 4-7; the full set of time points, with each animal shown individually is in Supplementary Figure 4 A-I). In the control animals (five mice in each of two cages), although the composition of the bacterial community remained relatively stable, the fungal community changed dramatically. For four samples taken over the first eight days, most of the ten mice in the two cages showed colonization by diverse fungal lineages, and no lineage predominated. By day 15, however, the situation had changed radically, with both cages dominated by a phylotype annotated as *Wickerhamomyces*. This changed by day 22, with cage 1 dominated by *Debaryomyces*, and the second cage showing more diverse colonization, where *Debaryomyces* was present but not predominant. The outgrowth of *Debaryomyces* in cage 1 correlated with the increase in abundance of total DNA in the 18S QPCR analysis in the control animals at day 22 (Figure 4-3), which also occurred only in cage 1. By day 76 the controls had changed again. At this time both cages were dominated by *Eurotiales*, though the abundance was greater in cage 1 than cage 2. These findings document radical changes in murine gut fungi 1) in a single mouse facility 2) for mice on a homogeneous diet, 3) in the absence of any intervention, and 4) differing between cages.

For the antibiotic treated animals, the communities were diverse prior to initiation of treatment, then under treatment showed cage-specific patterns of divergence. By day 2, *Clavispora* and *Cyberlindnera* were predominant in both cages in the ABXContinuous group and cage 1 in the ABXShortTerm group. In cage 2 in the short term group the mice were more heterogenous, with *Candida*, *Cyberlindnera*, and *Pichia* dominating in different mice. By day 6 this had changed, with *Cyberlinderna* dominating in three cages (both ABXContinuous and ABXShortTerm cage 1), and *Candida* dominating in all mice in ABXShortTerm cage 2. This pattern persisted at day 15, but by day 22 one ABXContinuous cage was dominated by *Candida*, and by day 76 *Candida* was the only fungus detectable in both ABXContinuous cages. For the ABXShortTerm mice, *Sporopachydermis* dominated in cage 1 and *Candida* in cage 2 on day 22, then the fungal populations returned to a more diverse mixture by day 76, but *Candida* was relatively abundant in most of the short term treated mice at the last time point studied. Thus these data emphasize the heterogeneity of the community responses in individual cages, but also the robust persistence of the *Candida* community. A few samples were also analyzed with the 18S rRNA gene amplicon and generally yielded similar results (Supplementary Figure 4-1). An analysis of mouse chow DNA using the ITS amplicon showed no obvious relationship to the major lineages detected in pellets (Supplementary Figure 4-3B).



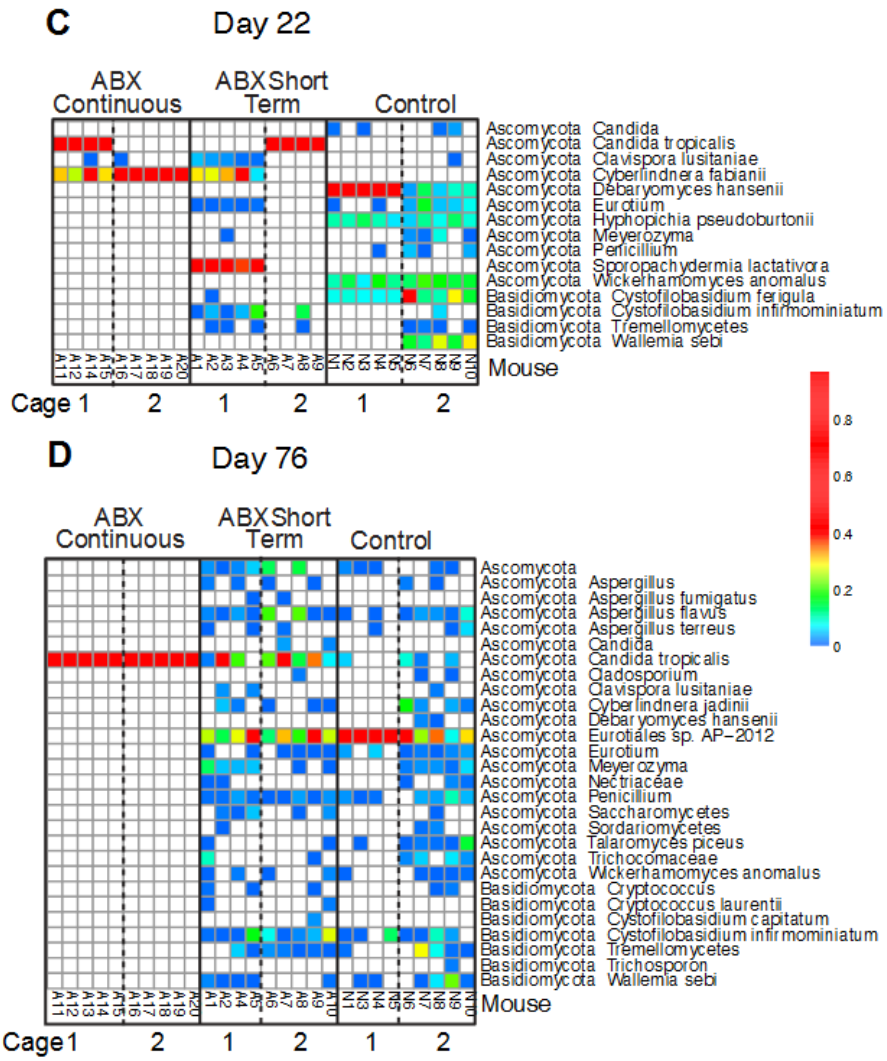


Figure 4-7. Longitudinal analysis of fungal lineages inferred from ITS rRNA gene sequencing. Fungal lineages detected are summarized in heat map format. Data for four days are shown (days 1, 15, 22, and 76). Each column indicates a single mouse. The groups tested are indicated at the top of the columns. The ten mice in each of the three treatment groups were each housed in two cages of five each. The distribution of mice in cages is indicated at the bottom of the columns. The day of treatment is indicated at the top. The color code to the right indicates the proportions.

Initially an average of 30.7 (SD=9.9) phylotypes were detected in the ITS data per animal (Figure 4-8), but these numbers fell to 5.2 (SD=2.9) per animal during antibiotic treatment as the *Candida* overgrew the community. Upon cessation of antibiotic treatment the number of

phylotypes returned to their former levels (31.1, SD=8.0). Thus the fungal community returned to its original richness (p=0.673, Mann-Whintey U), but despite this, the contribution of *Candida* was higher than before treatment.

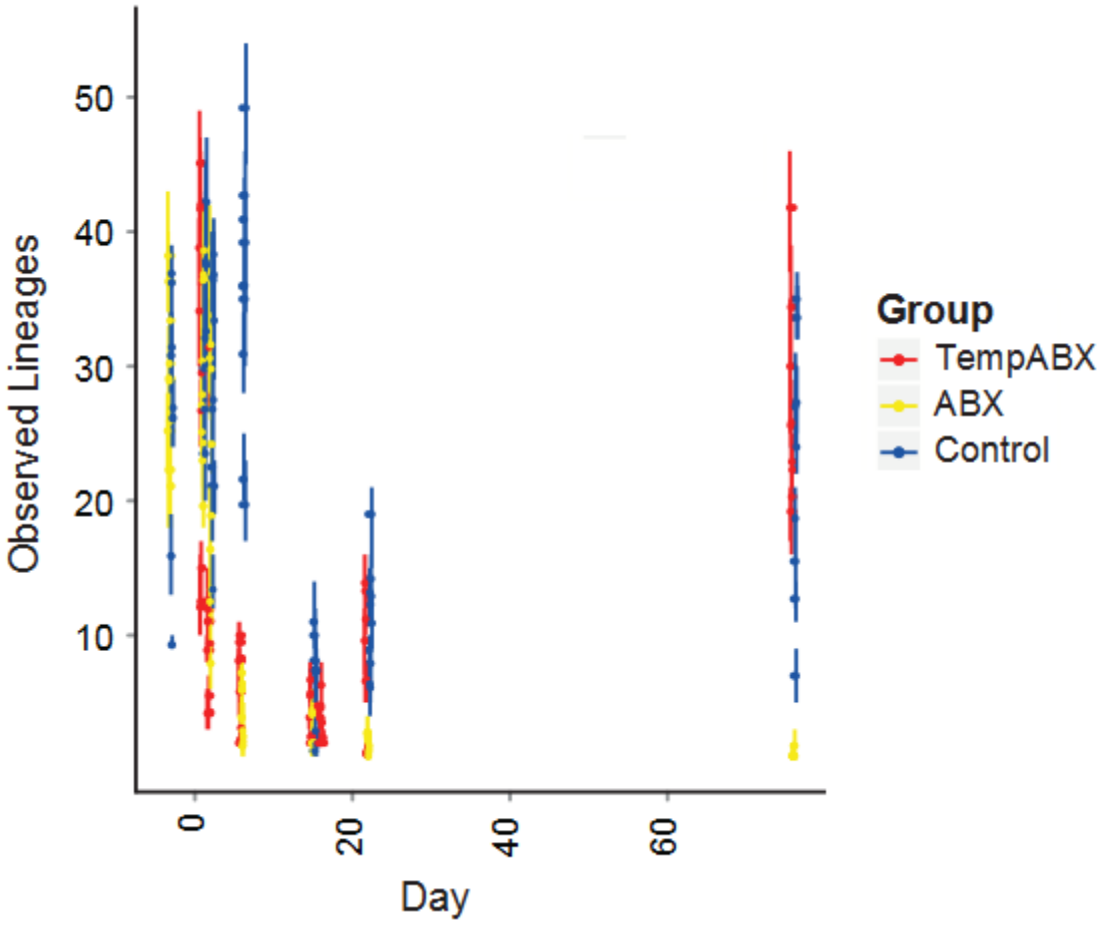


Figure 4-8. Abundance analysis of observed fungal lineages. Each sequence set was rarefied to 200 sequences per sample, and the number of different OTUs assessed. Means are indicated by points, error bars indicate the range observed. The groups studied are indicated by the key at the right.

Community comparisons using Unifrac

To compare community structures over the course of antibiotic treatment, the 454 data for the 16S and ITS tags were analyzed using weighted[62] and unweighted[63] Unifrac

(Supplementary Figure 4-5). We tested for differenced in community composition between the control group and the treatment groups at each time point by comparing distances within the control group to distances between control and treatment groups. Significance was determined using the Mann-Whitney test.

The ABXShortTerm and ABXContinuous groups were not significantly different from the Control group before antibiotic treatment for either the bacterial or fungal communities (Supplementary Figure 4-5). After one day of treatment, both groups were significantly different from the Control in both weighted and unweighted Unifrac and remained distinct throughout the antibiotic treatment (Supplementary Figure 4-5). One week after the ShortTermABX group stopped receiving antibiotics, both the bacterial and fungal communities remained significantly different from the other two (Supplementary Figure 4-5). On day 76, 61 days after antibiotic cessation, differences between the bacterial communities were slight but still achieved significance in the weighted and unweighted analysis, and some but not all comparisons were significant for the ITS analysis. Thus by the last time point the communities had approached but not completely returned to their pre-treatment states ($p<0.005$).

4.5 Discussion

We report a study of the response of bacterial and fungal communities in the mouse gut to antibiotic treatment. Animals were tracked over 76 days, and sampled densely over this period. Use of both QPCR to assess total abundance and 454/Roche sequencing to quantify community structure provided a detailed picture of the effects of antibiotic treatment and subsequent recovery on the gut microbiome.

The bacterial communities initially were dominated by *Bacteriodetes* and *Lachnospiraceae*, then changed quickly during the antibiotic treatment, reaching a stable state by

day two. At this time, the number of inferred organisms had dropped >3 orders of magnitude, to the point that the 16S rRNA gene sequences were mostly transients from food, documenting near complete clearance of bacteria from the gut. Upon cessation of antibiotic treatment, the bacteria returned to nearly their original state, though perturbations remained. The control animals remained relatively stable over the time course, dominated by *Bacteriodetes* and *Lachnospiraceae* throughout.

Fungal communities changed radically over the time course studied, in association with the caging history of the mice. Even the untreated Control group showed waves of succession that differed in each cage. Cage 1 was first heterogenous, then dominated by *Wickerhamomyces*, then *Debaryomyces*, then *Eurotiales*. Cage 2 was heterogenous through the first six days, then dominated by *Wickerhamomyces,* then again heterogeneous. These data document a quite surprising degree of fungal variation in healthy laboratory mice.

Previous studies have shown that colonization by specific bacteria in mice can strongly influence the outcome of immunological assays [64]. Given the recently reported importance of fungi in mouse models of IBD [36], our data suggest that researchers will need to take care to monitor and control fungal populations to obtain meaningful data. The mice in this study were housed on a conventional SPF environment. It would be useful to compare housing in a barrier facility where food, water, and bedding have all been sterilized. Colitis phenotypes in genetically-determined mouse models are known to vary in different facilities, suggesting that it will be useful to assess the role of variability in fungal colonization.

In the presence of antibiotics, the fungal community showed several waves of succession, which again differed between cages. Depending on the cage and time point, the communities could be heterogeneous, or dominated by *Clavispora*, *Cyberlindnera*, *Sporopachydermia* or

Candida. In cases where communities in a cage were dominated by a single fungal lineage, this was true of all mice in the cage at that time point. One likely explanation is that coprophagia resulted in all mice in a single cage acquiring the same fungal colonists. Some of these fungal lineages were seen sporadically in contamination controls (Supplementary Figure 4-6), suggesting that gut fungi may have been acquired episodically from the environment.

Candida was a particularly robust colonist in the presence of antibiotics. By day 76, all the ABXContinuous mice were colonized at a high level exclusively by *Candida*. Further supporting the robustness of *Candida*, results from cage two in the ABXShortTerm group by chance provides a competition experiment. Of the five mice analyzed on day two, two were colonized with *Candida*, one with *Cyberlindnera*, one with *Pichia*, and one with both *Candida* and *Cyberlindnera* (Supplementary Figure 4-4, part C). Coprophagia would presumably allow the three species to compete for colonization opportunities. By day six, all mice were colonized with *Candida*, and this persisted through the cessation of antibiotic treatment by day 22. At the end of the experiment on day 76, *Candida* was still more abundant in the ABXShortTerm group than prior to treatment, all emphasizing that *Candida* was favored by the antibiotic treatment and persisted subsequently.

4.6 Acknowledgements

We are grateful to members of the Bushman laboratory for help and suggestions. This work was supported by Project UH2DK083981, the Penn Genome Frontiers Institute; NIH AI39368 (GDW); Penn Digestive Disease Center (P30 DK050306); The Joint Penn-CHOP Center for Digestive, Liver, and Pancreatic Medicine; S10RR024525; UL1RR024134, and K24-DK078228; and the University of Pennsylvania Center for AIDS Research (CFAR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources, National Institutes of Health, or Pennsylvania Department of

Health.

4.7 References

1. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334: 105-108.

2. McKenna P, Hoffmann C, Minkah N, Aye PP, Lackner A, et al. (2008) The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. PLoS pathogens 4: e20.

3. Adachi JA, Jiang Z-D, Mathewson JJ, Verenkar MP, Thompson S, et al. (2001) Enteroaggregative Escherichia coli as a major etiologic agent in traveler's diarrhea in 3 regions of the world. Clinical infectious diseases 32: 1706-1709.

4. Martin F-PJ, Wang Y, Sprenger N, Yap IK, Lundstedt T, et al. (2008) Probiotic modulation of symbiotic gut microbial–host metabolic interactions in a humanized microbiome mouse model. Molecular systems biology 4.

5. Jakobsson HE, Jernberg C, Andersson AF, Sjölund-Karlsson M, Jansson JK, et al. (2010) Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. PloS one 5: e9836.

6. Chen KY, Ko SC, Hsueh PR, Luh KT, Yang PC (2001) Pulmonary fungal infection: emphasis on microbiological spectra, patient outcome, and prognostic factors. Chest 120: 177-184.

7. Samonis G, Gikas A, Anaissie EJ, Vrenzos G, Maraki S, et al. (1993) Prospective evaluation of effects of broad-spectrum antibiotics on gastrointestinal yeast colonization of humans. Antimicrobial agents and chemotherapy 37: 51-53.

8. Mulligan ME, Citron DM, McNamara BT, Finegold SM (1982) Impact of cefoperazone therapy on fecal flora. Antimicrobial agents and chemotherapy 22: 226-230.

9. Abbott J (1995) Clinical and microscopic diagnosis of vaginal yeast infection: a prospective analysis. Annals of emergency medicine 25: 587-591.

10. Oksala E (1990) Factors predisposing to oral yeast infections. Acta odontologica Scandinavica 48: 71-74.

11. Leyden JJ, Marples RR (1973) Ecologic principles and antibiotic therapy in chronic dermatoses. Archives of dermatology 107: 208-211.

12. Ben-Ami R, Olshtain-Pops K, Krieger M, Oren I, Bishara J, et al. (2012) Antibiotic exposure as a risk factor for fluconazole-resistant Candida bloodstream infection. Antimicrobial agents and chemotherapy 56: 2518-2523.

13. Bow EJ, Louie TJ (1987) Changes in endogenous microflora among febrile granulocytopenic patients receiving empiric antibiotic therapy: implications for fungal superinfection. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 137: 397-403.

14. Cheong JW, McCormack J (2012) Fluconazole resistance in cryptococcal disease: emerging or intrinsic? Medical mycology : official publication of the International Society for Human and Animal Mycology.

15. Moris DV, Melhem MS, Martins MA, Souza LR, Kacew S, et al. (2012) Prevalence and antifungal susceptibility of Candida parapsilosis complex isolates collected from oral cavities of HIV-infected individuals. Journal of medical microbiology 61: 1758-1765.

16. Ampel NM (1996) Emerging disease issues and fungal pathogens associated with HIV infection. Emerging infectious diseases 2: 109-116.

17. Chamilos G, Luna M, Lewis RE, Bodey GP, Chemaly R, et al. (2006) Invasive fungal infections in patients with hematologic malignancies in a tertiary care cancer center: an autopsy study over a 15-year period (1989-2003). Haematologica 91: 986-989.

18. Viscoli C, Girmenia C, Marinus A, Collette L, Martino P, et al. (1999) Candidemia in cancer patients: a prospective, multicenter surveillance study by the Invasive Fungal Infection Group (IFIG) of the European Organization for Research and Treatment of Cancer (EORTC). Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 28: 1071-1079.

19. Gallis HA, Berman RA, Cate TR, Hamilton JD, Gunnells JC, et al. (1975) Fungal infection following renal transplantation. Archives of internal medicine 135: 1163-1172.

20. Trenschel R, Peceny R, Runde V, Elmaagacli A, Dermoumi H, et al. (2000) Fungal colonization and invasive fungal infections following allogeneic BMT using metronidazole, ciprofloxacin and fluconazole or ciprofloxacin and fluconazole as intestinal decontamination. Bone marrow transplantation 26: 993-997.

21. Koo S, Kubiak DW, Issa NC, Dietzek A, Boukedes S, et al. (2012) A targeted peritransplant antifungal strategy for the prevention of invasive fungal disease after lung transplantation: a sequential cohort analysis. Transplantation 94: 281-286.

22. Karchmer AW, Samore MH, Hadley S, Collins LA, Jenkins RL, et al. (1995) Fungal infections complicating orthotopic liver transplantation. Transactions of the American Clinical and Climatological Association 106: 38-47; discussion 47-38.

23. Charlson ES, Diamond JM, Bittinger K, Fitzgerald AS, Yadav A, et al. (2012) Lung-enriched Organisms and Aberrant Bacterial and Fungal Respiratory Microbiota after Lung Transplant. Am J Respir Crit Care Med 186: 536-545.

24. Lionakis MS, Kontoyiannis DP (2003) Glucocorticoids and invasive fungal infections. Lancet 362: 1828-1838.

25. Sasse C, Dunkel N, Schafer T, Schneider S, Dierolf F, et al. (2012) The stepwise acquisition of fluconazole resistance mutations causes a gradual loss of fitness in Candida albicans. Molecular microbiology 86: 539-556.

26. Burgel PR, Baixench MT, Amsellem M, Audureau E, Chapron J, et al. (2012) High prevalence of azole-resistant Aspergillus fumigatus in adults with cystic fibrosis exposed to itraconazole. Antimicrobial agents and chemotherapy 56: 869-874.

27. Arikan-Akdagli S (2012) Azole resistance in Aspergillus: global status in Europe and Asia. Annals of the New York Academy of Sciences 1272: 9-14.

28. Perfect JR, Cox GM (1999) Drug resistance in Cryptococcus neoformans. Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy 2: 259-269.

29. Samonis G, Anastassiadou H, Dassiou M, Tselentis Y, Bodey GP (1994) Effects of broad-spectrum antibiotics on colonization of gastrointestinal tracts of mice by Candida albicans. Antimicrobial agents and chemotherapy 38: 602-603.

30. Mason KL, Erb Downward JR, Mason KD, Falkowski NR, Eaton KA, et al. (2012) Candida albicans and bacterial microbiota interactions in the cecum during recolonization following broad-spectrum antibiotic therapy. Infection and immunity 80: 3371-3380.

31. Hill DA, Hoffmann C, Abt MC, Du Y, Kobuley D, et al. (2010) Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis. Mucosal Immunol 3: 148-158.

32. Noverr MC, Falkowski NR, McDonald RA, McKenzie AN, Huffnagle GB (2005) Development of allergic airway disease in mice following antibiotic therapy and fungal microbiota increase: role of host genetics, antigen, and interleukin-13. Infection and immunity 73: 30-38.

33. Kamada N, Kim YG, Sham HP, Vallance BA, Puente JL, et al. (2012) Regulated virulence controls the ability of a pathogen to compete with the gut microbiota. Science 336: 1325-1329.

34. Peterfreund GL, Vandivier LE, Sinha R, Marozsan AJ, Olson WC, et al. (2012) Succession in the gut microbiome following antibiotic and antibody therapies for Clostridium difficile. PloS one 7: e46966.

35. Mason KL, Erb Downward JR, Falkowski NR, Young VB, Kao JY, et al. (2012) Interplay between the gastric bacterial microbiota and Candida albicans during postantibiotic recolonization and gastritis. Infection and immunity 80: 150-158.

36. Iliev ID, Funari VA, Taylor KD, Nguyen Q, Reyes CN, et al. (2012) Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. Science 336: 1314-1317.

37. Cho I, Yamanishi S, Cox L, Methe BA, Zavadil J, et al. (2012) Antibiotics in early life alter the murine colonic microbiome and adiposity. Nature 488: 621-626.

38. Samuel BS, Hansen EE, Manchester JK, Coutinho PM, Henrissat B, et al. (2007) Genomic and metabolic adaptations of Methanobrevibacter smithii to the human gut. Proceedings of the National Academy of Sciences of the United States of America 104: 10643-10648.

39. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, et al. (2009) The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med 1: 6ra14.

40. Chung H, Pamp SJ, Hill JA, Surana NK, Edelman SM, et al. (2012) Gut immune maturation depends on colonization with a host-specific microbiota. Cell 149: 1578-1593.

41. Dollive S, Peterfreund GL, Sherrill-Mix S, Bittinger K, Sinha R, et al. (2012) A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. Genome biology 13: R60.

42. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, et al. (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiol 10: 206.

43. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Meth 7: 335-336.

44. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic acids research 37: D5-15.

45. Ko KS, Jung HS (2002) Three nonorthologous ITS1 types are present in a polypore fungus Trichaptum abietinum. Molecular phylogenetics and evolution 23: 112-122.

46. Rohel E, Couteaudier Y, Papierok B, Cavelier N, Dedryver CA (1997) Ribosomal internal transcribed spacer size variation correlated with RAPD-PCR pattern polymorphisms in the entomopathogenic fungus Erynia neoaphidis and some closely related species. Mycological Research 101: 573-579.

47. Valente P, Gouveia FC, Lemos GA, Pimentel D, Mendonca-Hagler LC, et al. (1997) PCR-amplified ITS length variation within the yeast genus Metschnikowia. The Journal of general and applied microbiology 43: 179-181.

48. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A.

49. Chang CF, Liu YR, Chen SF, Naumov GI, Naumova ES, et al. (2012) Five novel species of the anamorphic genus Candida in the Cyberlindnera clade isolated from natural substrates in Taiwan. Antonie van Leeuwenhoek 102: 9-21.

50. Kurtzman CP, Robnett CJ, Basehoar-Powers E (2008) Phylogenetic relationships among species of Pichia, Issatchenkia and Williopsis determined from multigene sequence analysis, and the proposal of Barnettozyma gen. nov., Lindnera gen. nov. and Wickerhamomyces gen. nov. FEMS yeast research 8: 939-954.

51. Kurtzman CP (2011) Phylogeny of the ascomycetous yeasts and the renaming of Pichia anomala to Wickerhamomyces anomalus. Antonie van Leeuwenhoek 99: 13-23.

52. Lozupone CA, Knight R (2008) Species divergence and the measurement of microbial diversity. FEMS microbiology reviews 32: 557-578.

53. Dollive S, Peterfreund GL, Sherrill-Mix S, Bittinger K, Sinha R, et al. (2012) A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. Genome Biol 13: R60.

54. Dethlefsen L, Huse S, Sogin ML, Relman DA (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS biology 6: e280.

55. Hoffmann C, Hill DA, Minkah N, Kirn T, Troy A, et al. (2009) Community-wide response of the gut microbiota to enteropathogenic Citrobacter rodentium infection revealed by deep sequencing. Infect Immun 77: 4668-4678.

56. Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. Journal of bacteriology 186: 2629-2635.

57. Lee CM, Sieo CC, Abdullah N, Ho YW (2008) Estimation of 16S rRNA gene copy number in several probiotic Lactobacillus strains isolated from the gastrointestinal tract of chicken. FEMS microbiology letters 287: 136-141.

58. Herrera ML, Vallor AC, Gelfond JA, Patterson TF, Wickes BL (2009) Strain-dependent variation in 18S ribosomal DNA Copy numbers in Aspergillus fumigatus. Journal of clinical microbiology 47: 1325-1332.

59. Metwally L, Fairley DJ, Coyle PV, Hay RJ, Hedderwick S, et al. (2008) Improving molecular detection of Candida DNA in whole blood: comparison of seven fungal DNA extraction protocols using real-time PCR. Journal of medical microbiology 57: 296-303.

60. (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207-214.

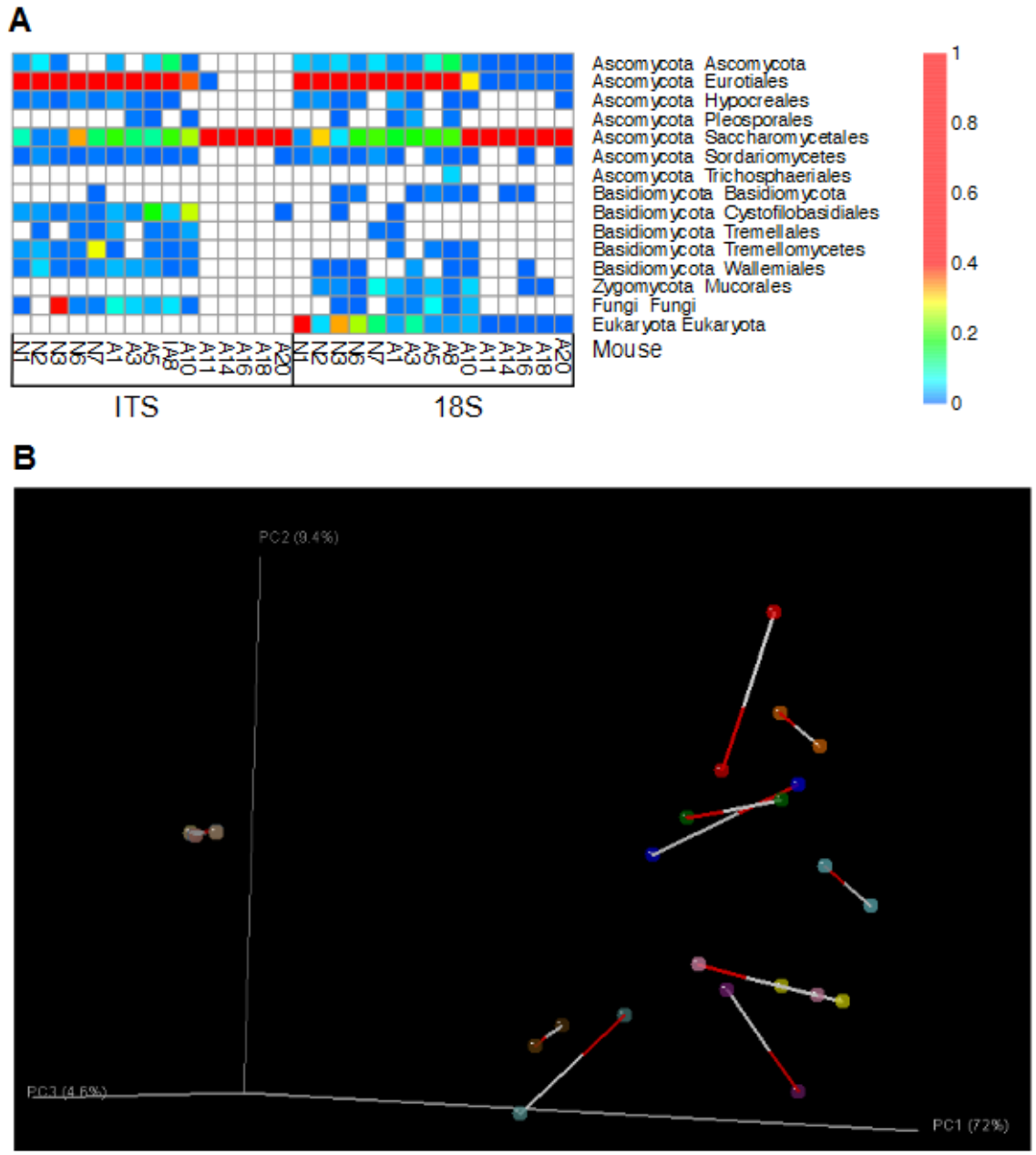
61. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73: 5261-5267.

62. Lozupone CA, Hamady M, Kelley ST, Knight R (2007) Quantitative and Qualitative {beta} Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. Applied and Environmental Microbiology 73: 1576-1585.

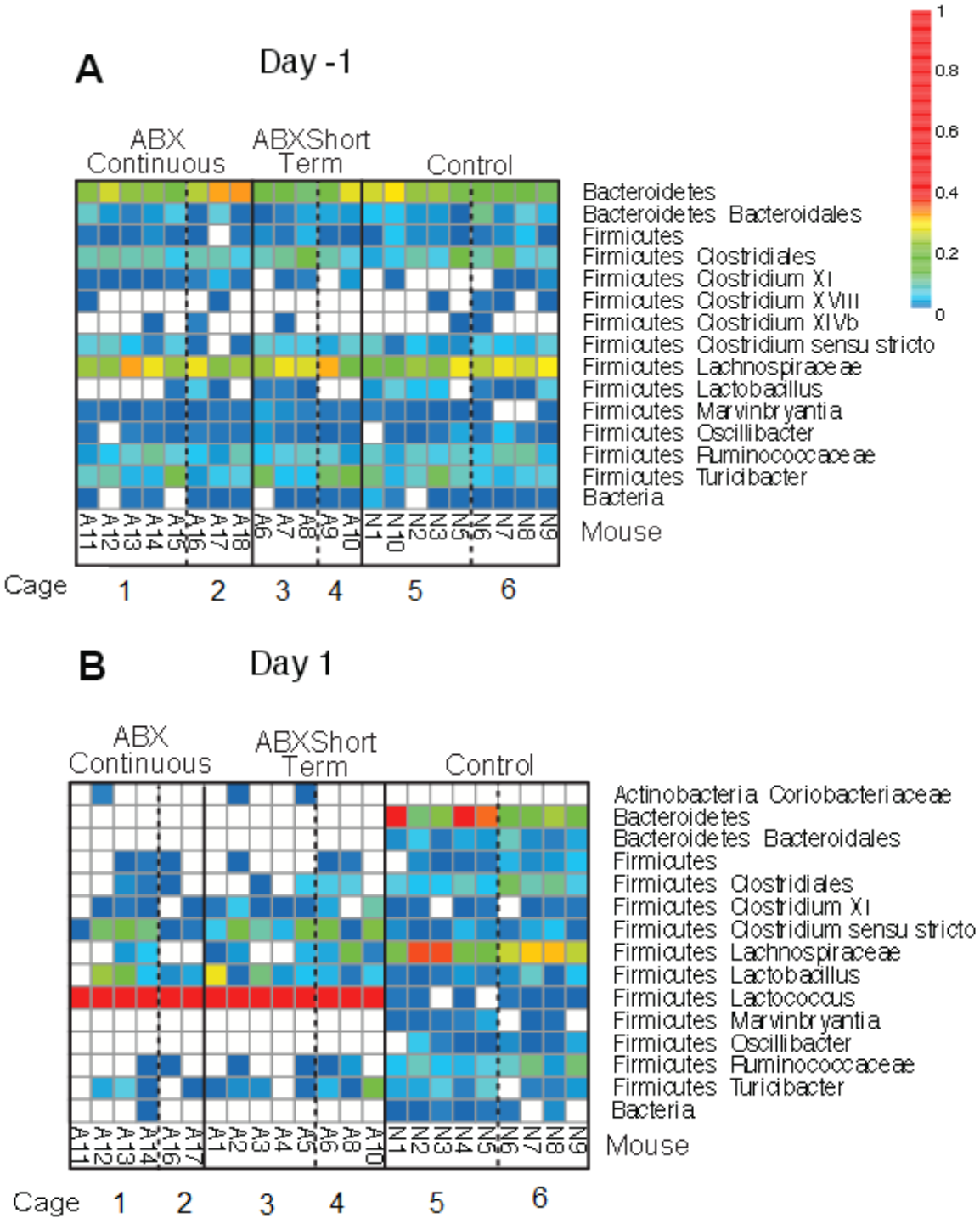
63. Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71: 8228-8235.

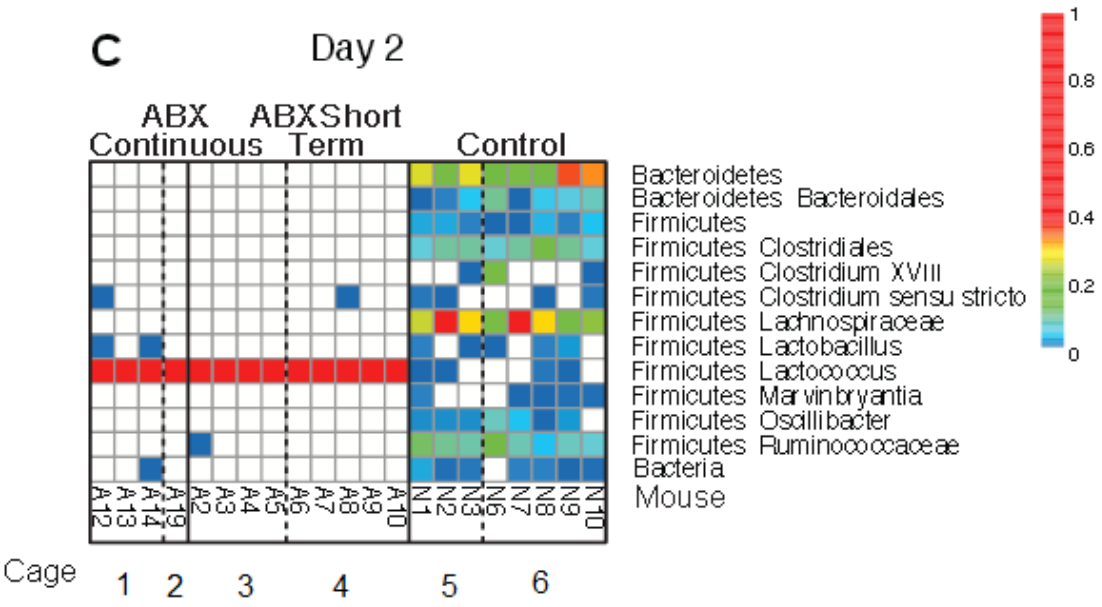
64. Ivanov, II, Atarashi K, Manel N, Brodie EL, Shima T, et al. (2009) Induction of intestinal Th17 cells by segmented filamentous bacteria. Cell 139: 485-498.

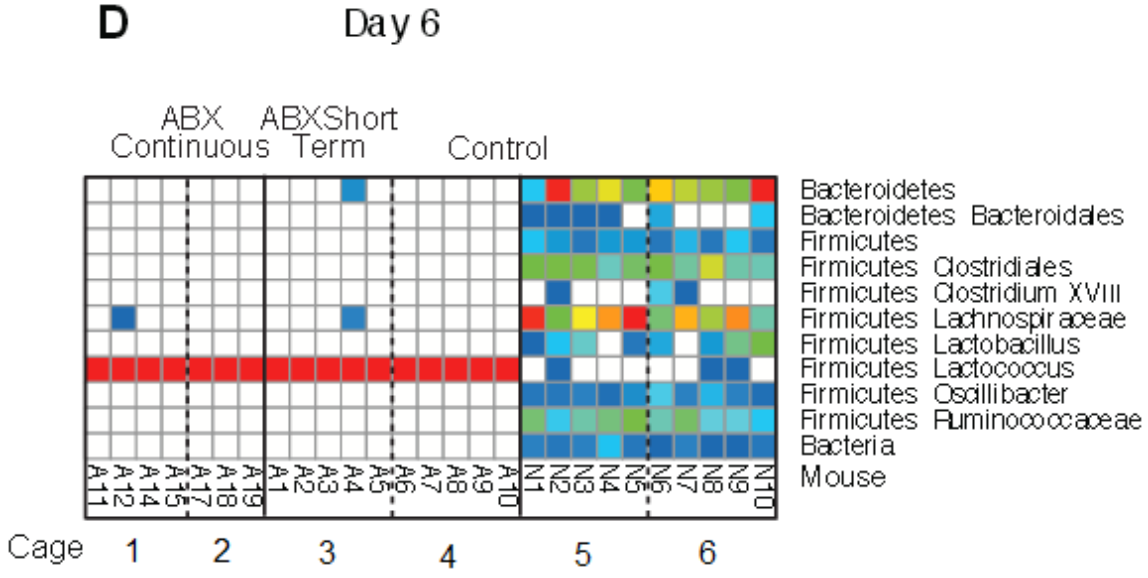
Supplementary Information

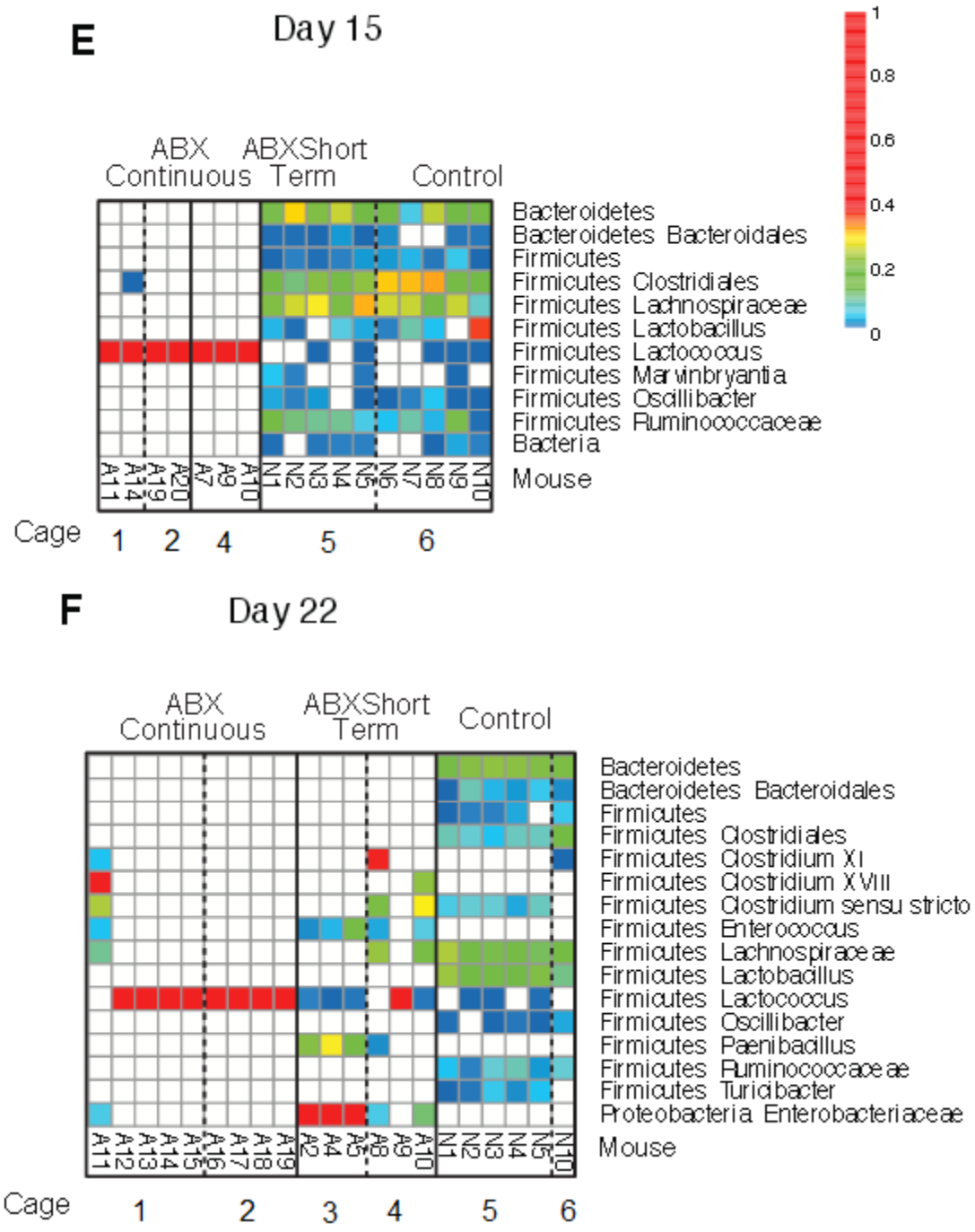


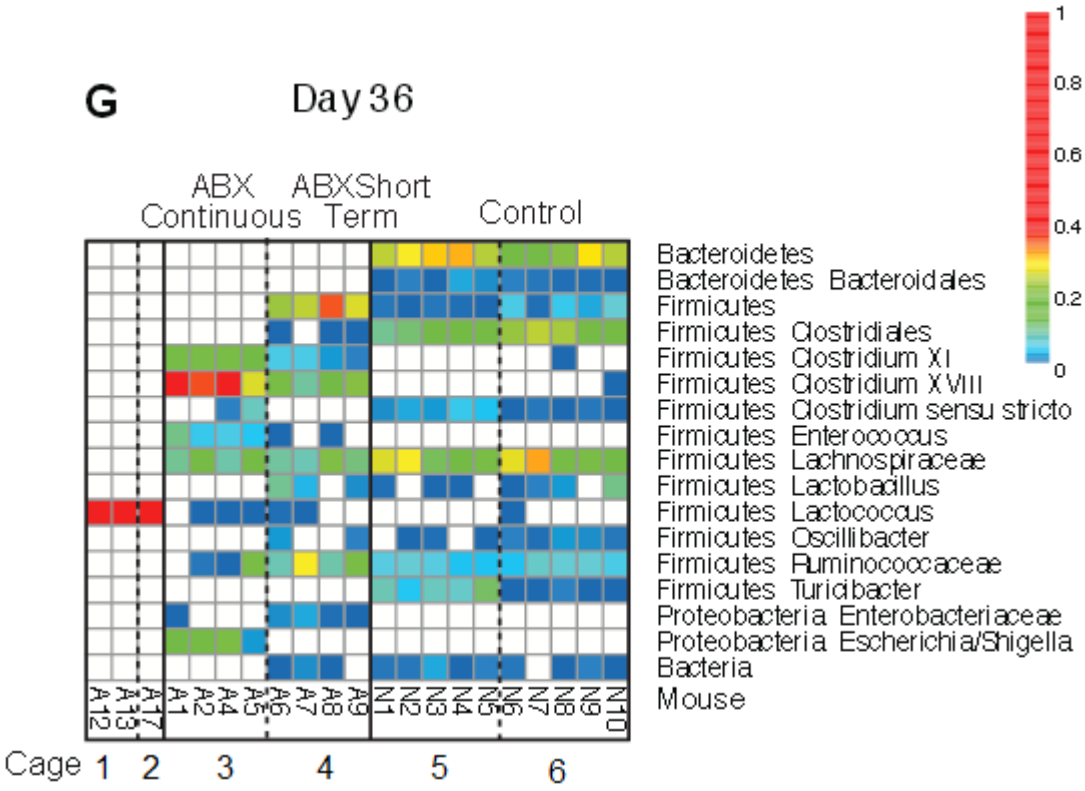
Supplementary Figure 4-1. Figure S1. Comparison of microeukaryote lineages specified by the ITS and 18S amplicons. A) Heat maps comparing selected samples analyzed using both the 18S and ITS amplicons. Each column shows the average for mice in the group and at the time point indicated rarefied to 200 reads per individual. The color code to the right indicates the scale. B) Procrustes analysis comparing results for the 18S and ITS analysis. Data from the 18S and ITS amplicons for each mouse are shown by balls connected by a line.

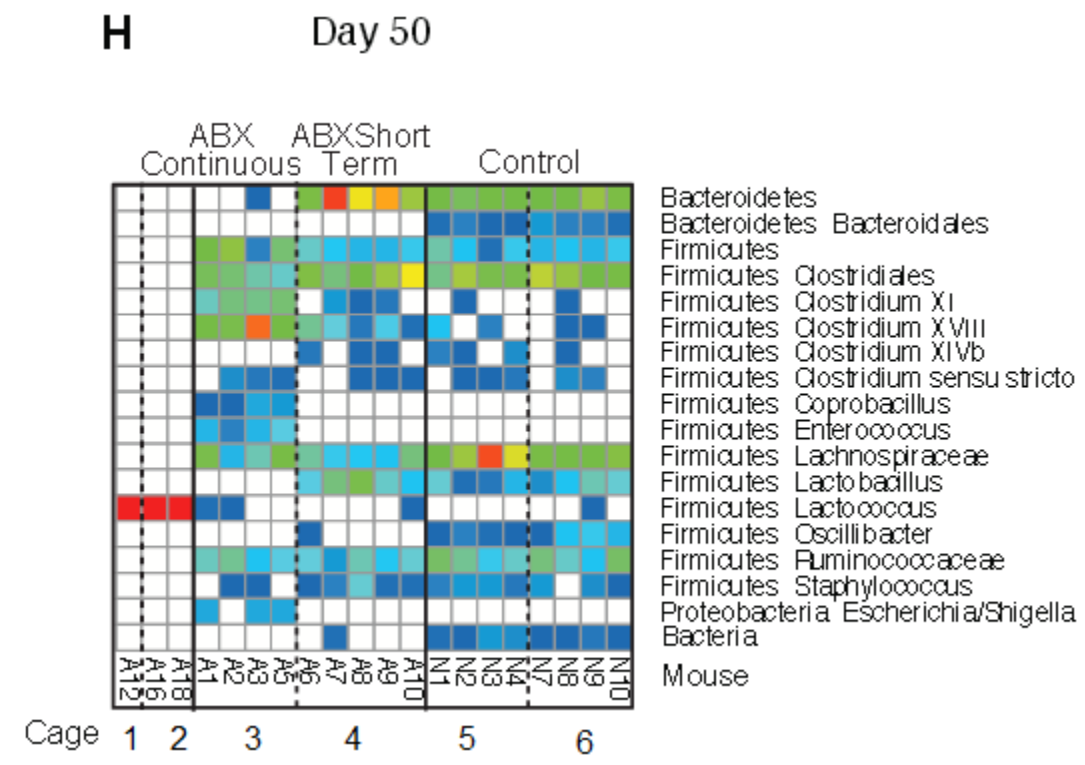


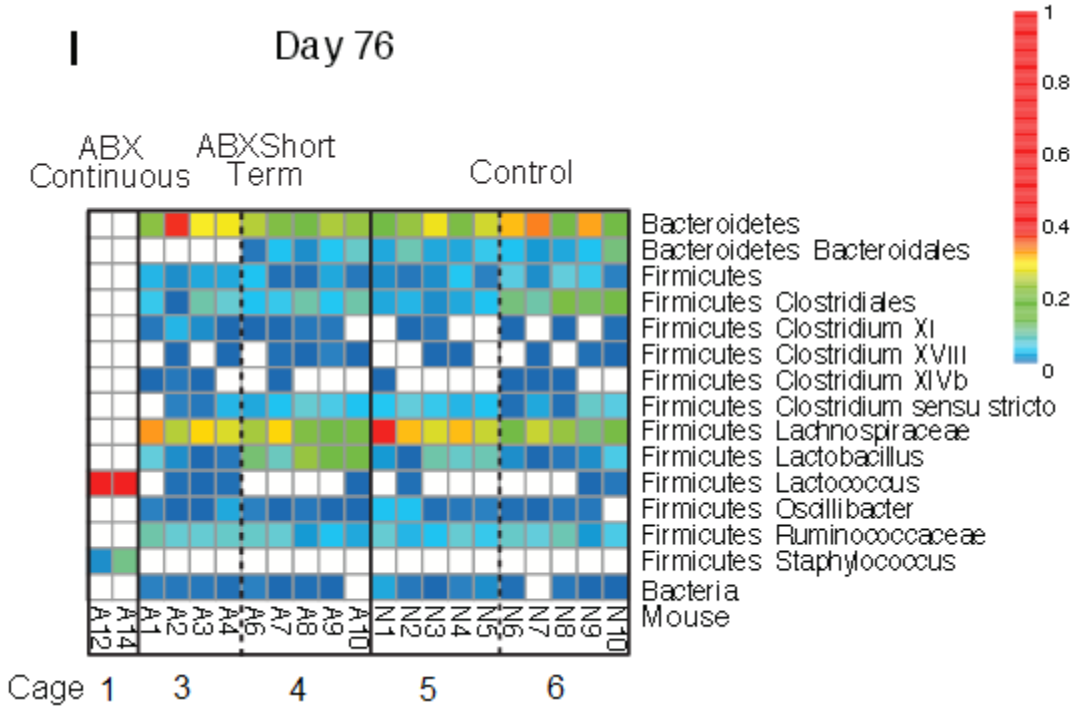




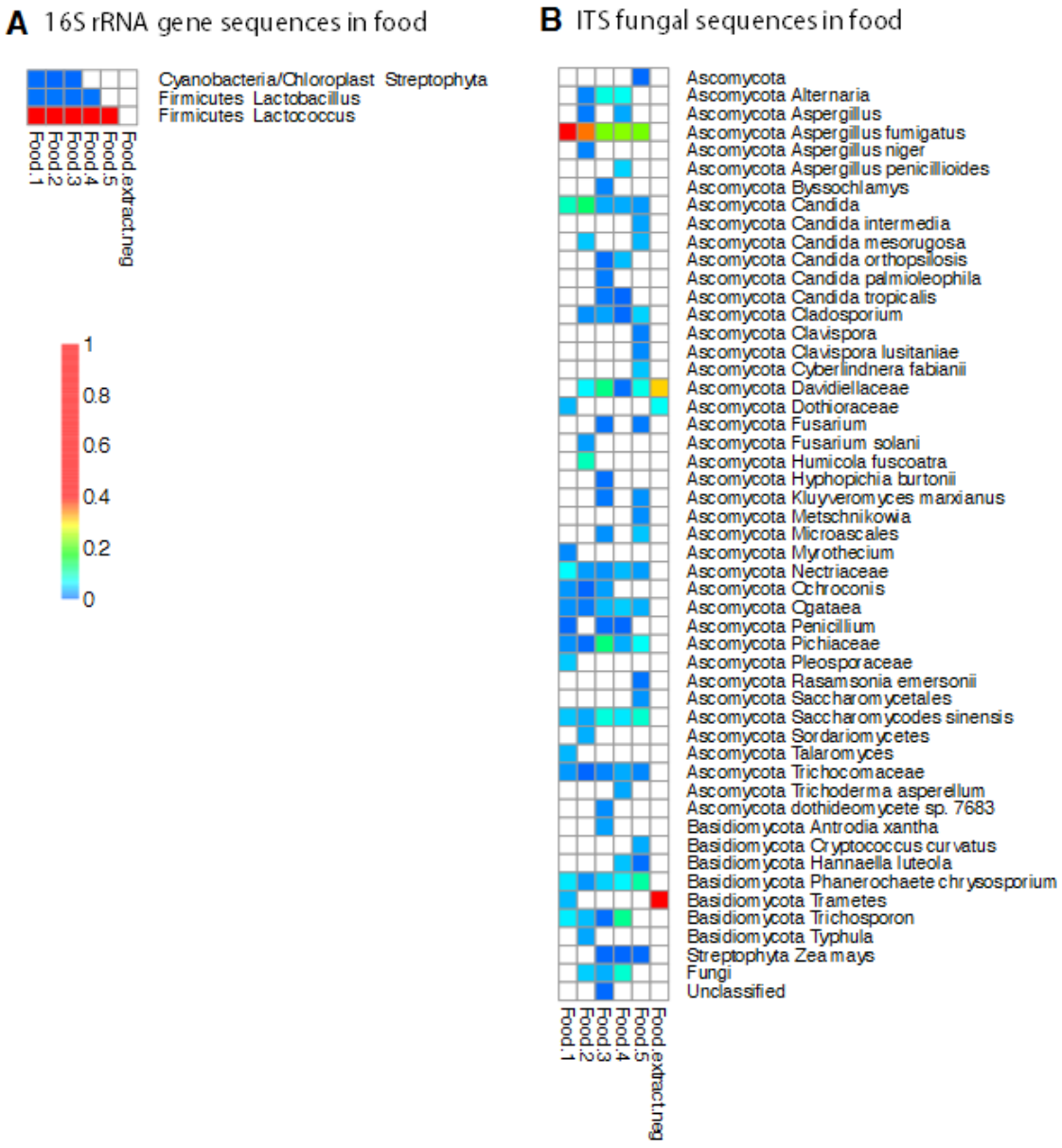




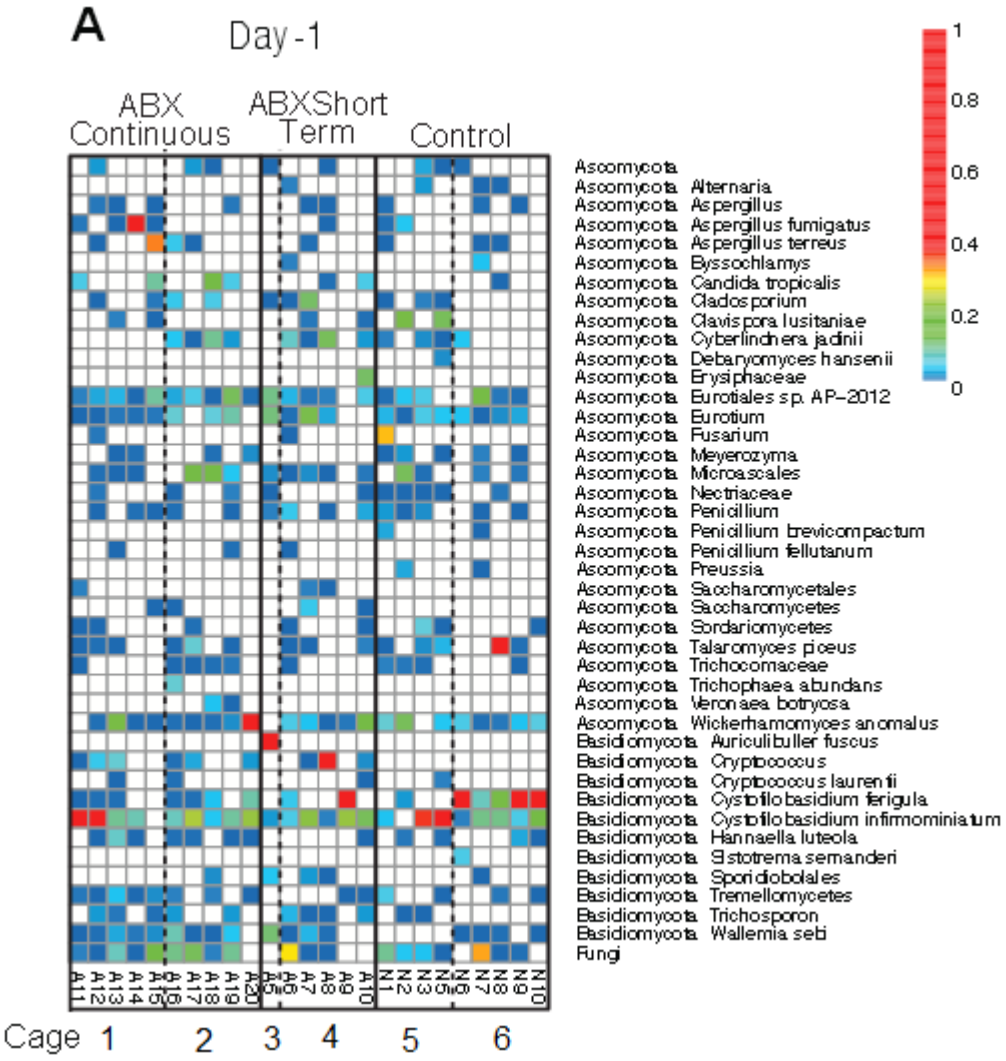


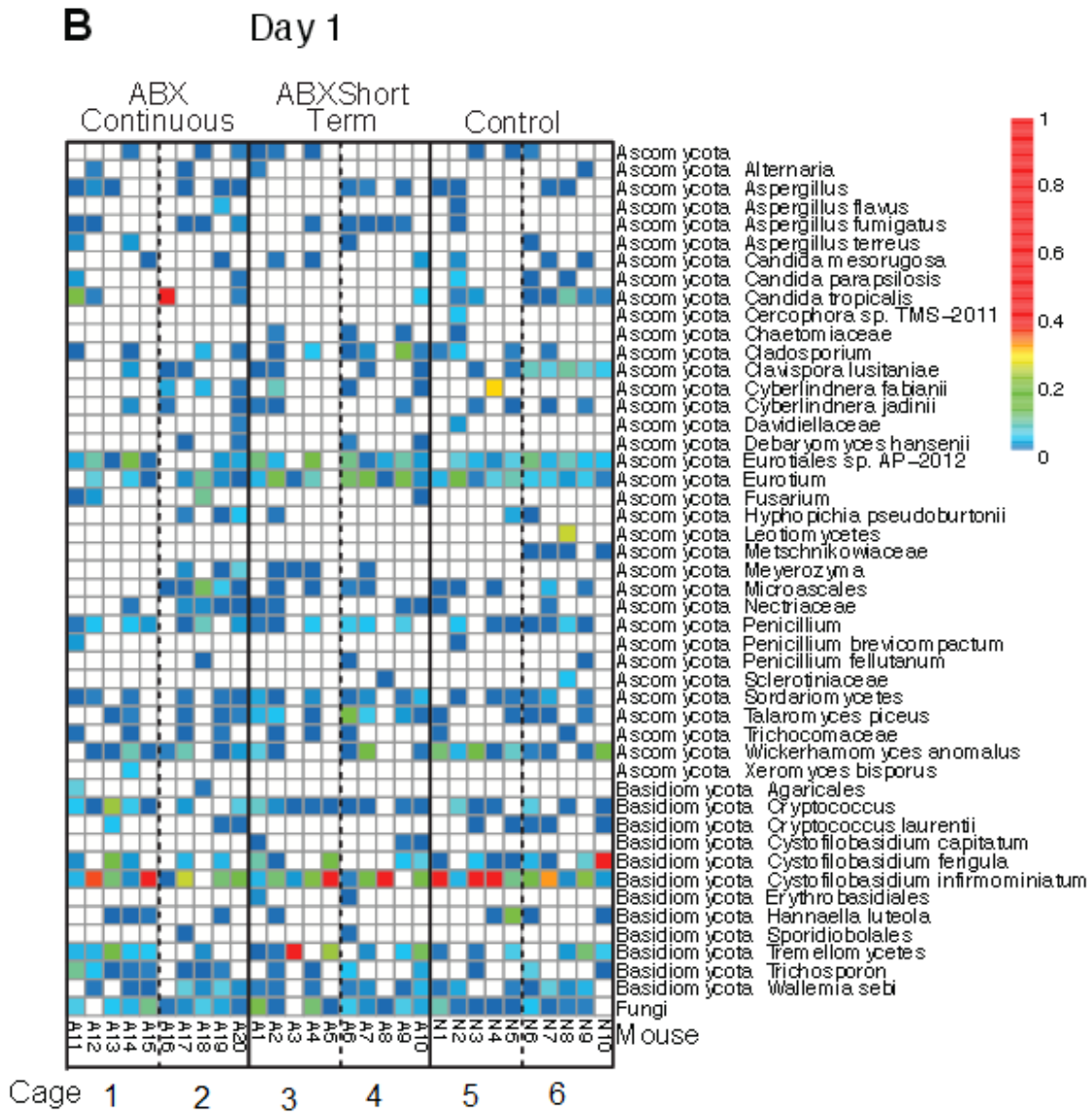


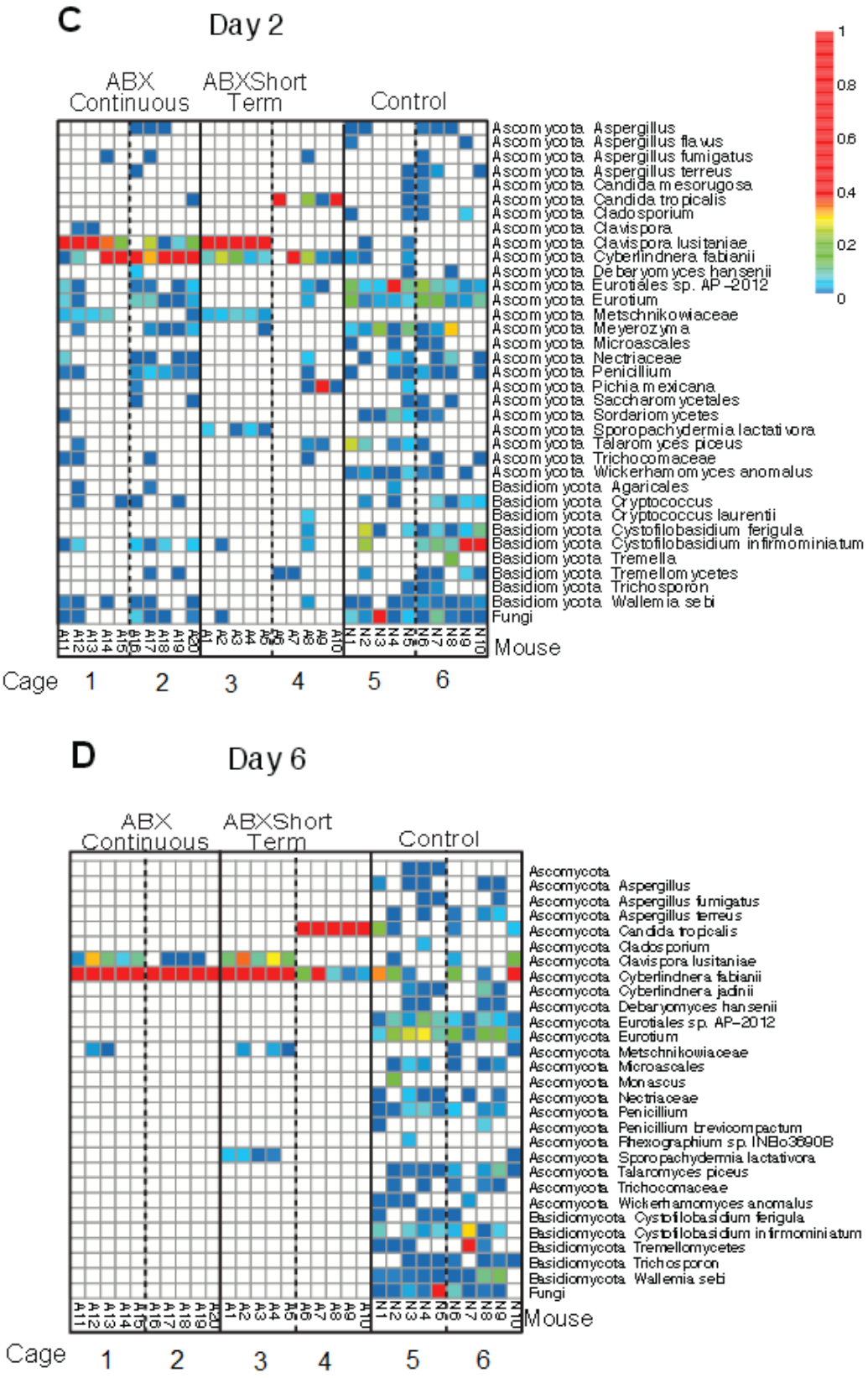
Supplementary Figure 4-2. Heat maps showing the composition of bacterial communities inferred from 16S sequence data for each time point (A-I), with each mouse shown individually. The scale of relative proportions is shown on the right.

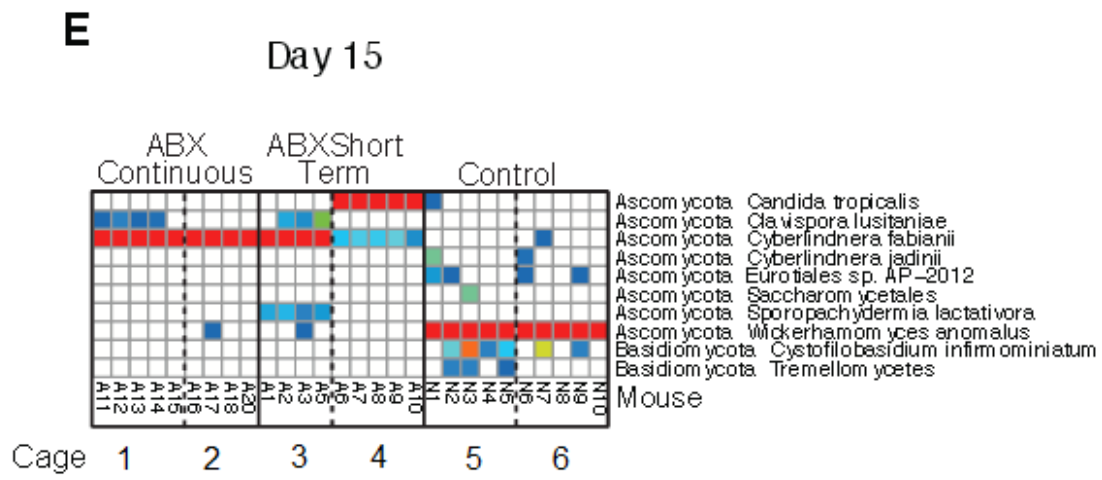


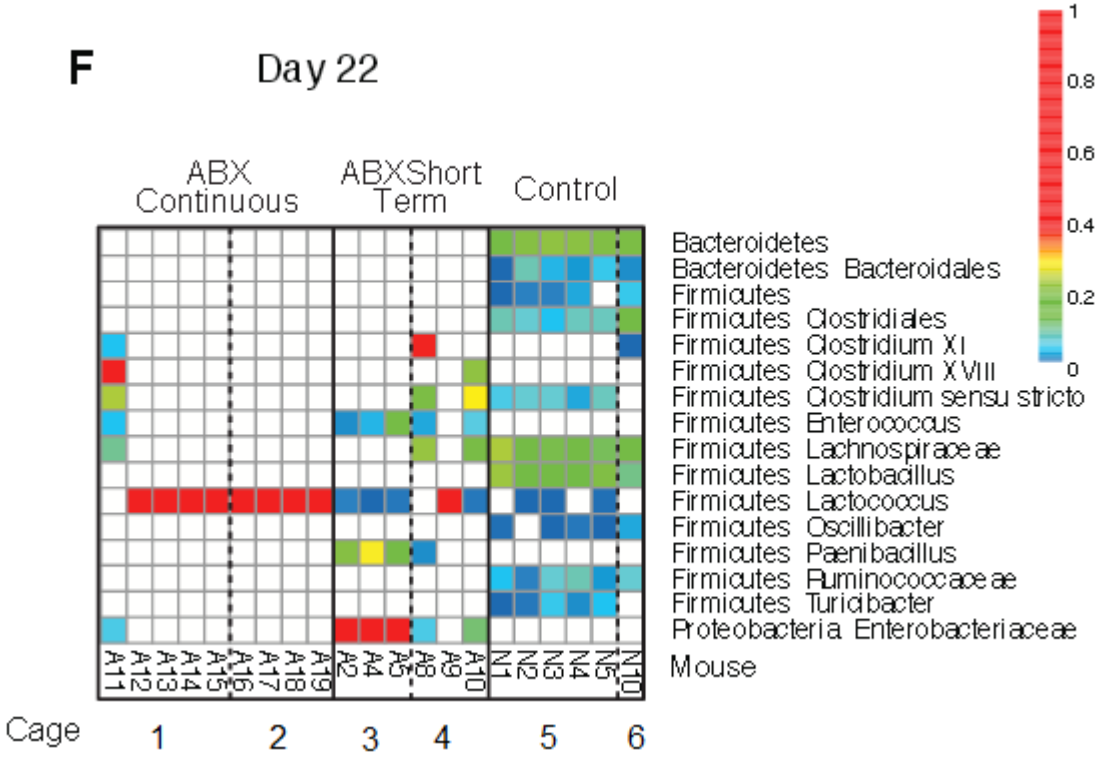
Supplementary Figure 4-3. 16S and ITS sequences recovered from five samples of mouse chow. A) Sequences from the 16S analysis. B) Sequences from the ITS analysis. The scale of relative proportions is shown on the far right.

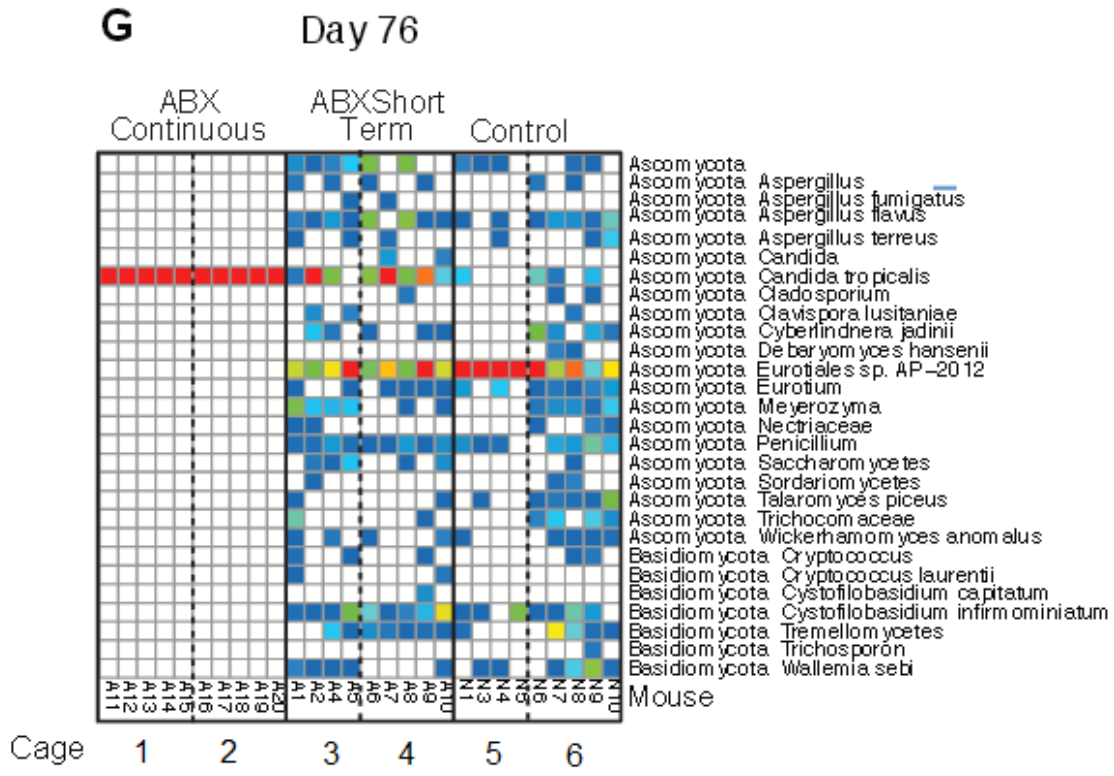




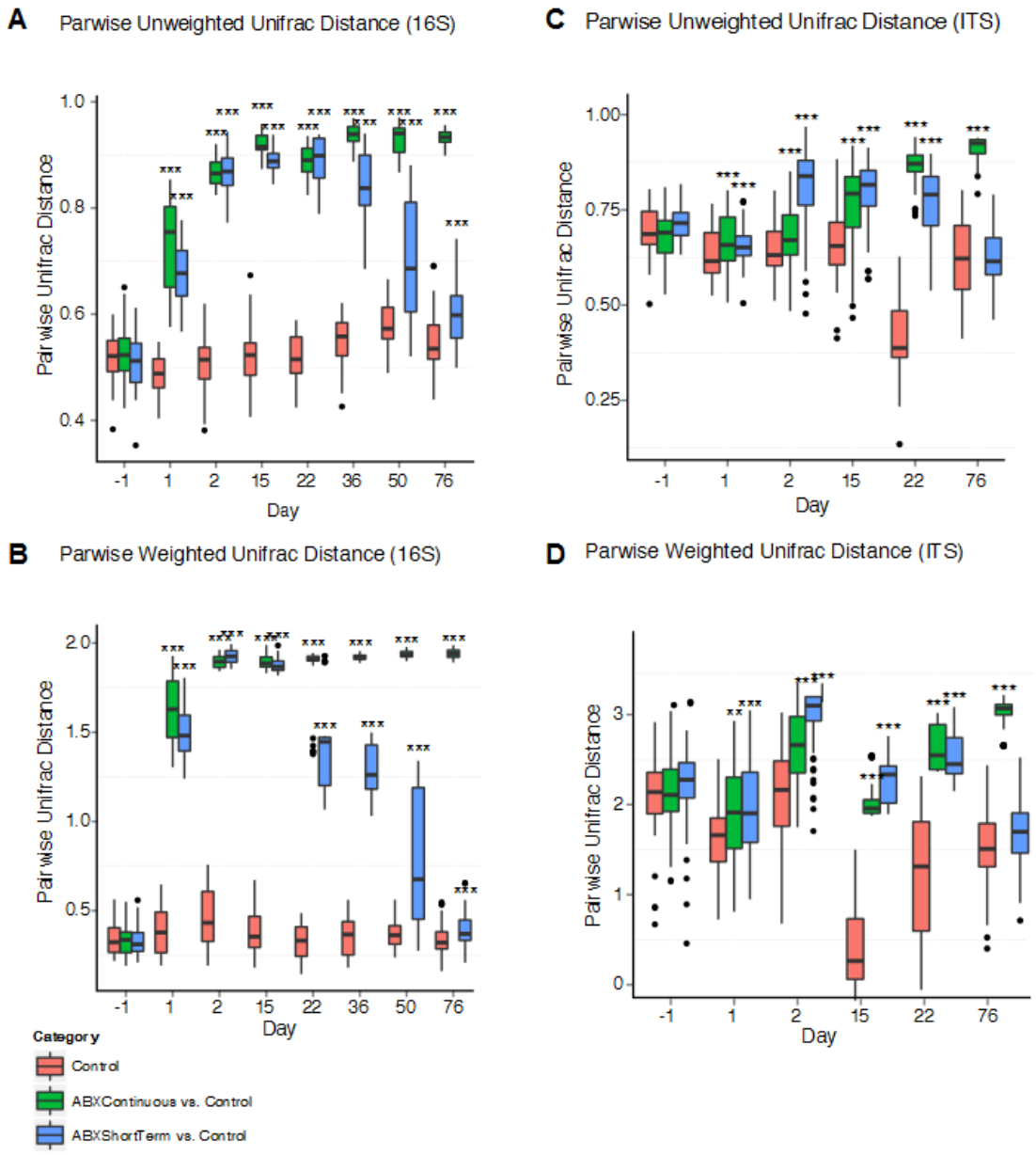






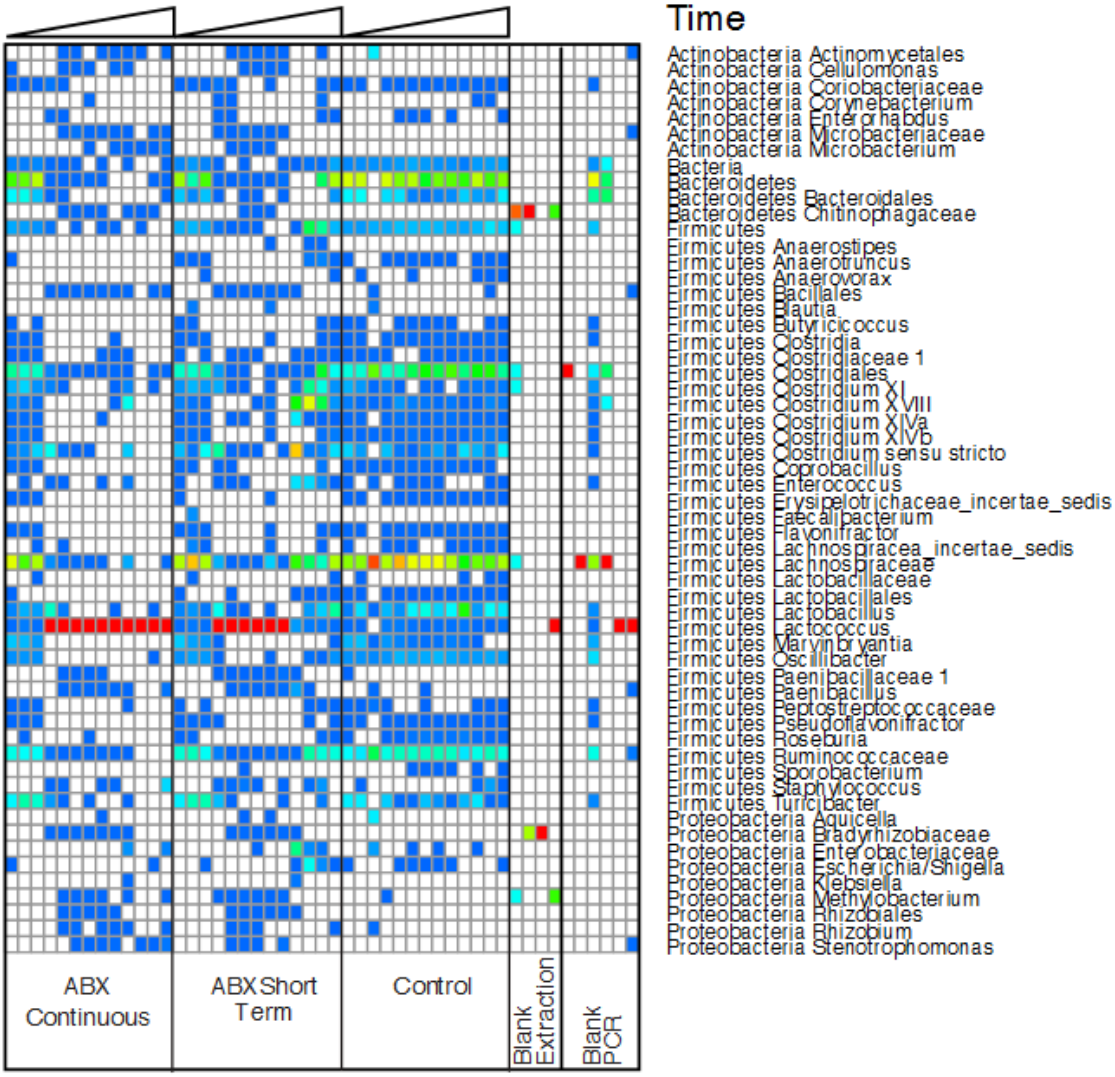


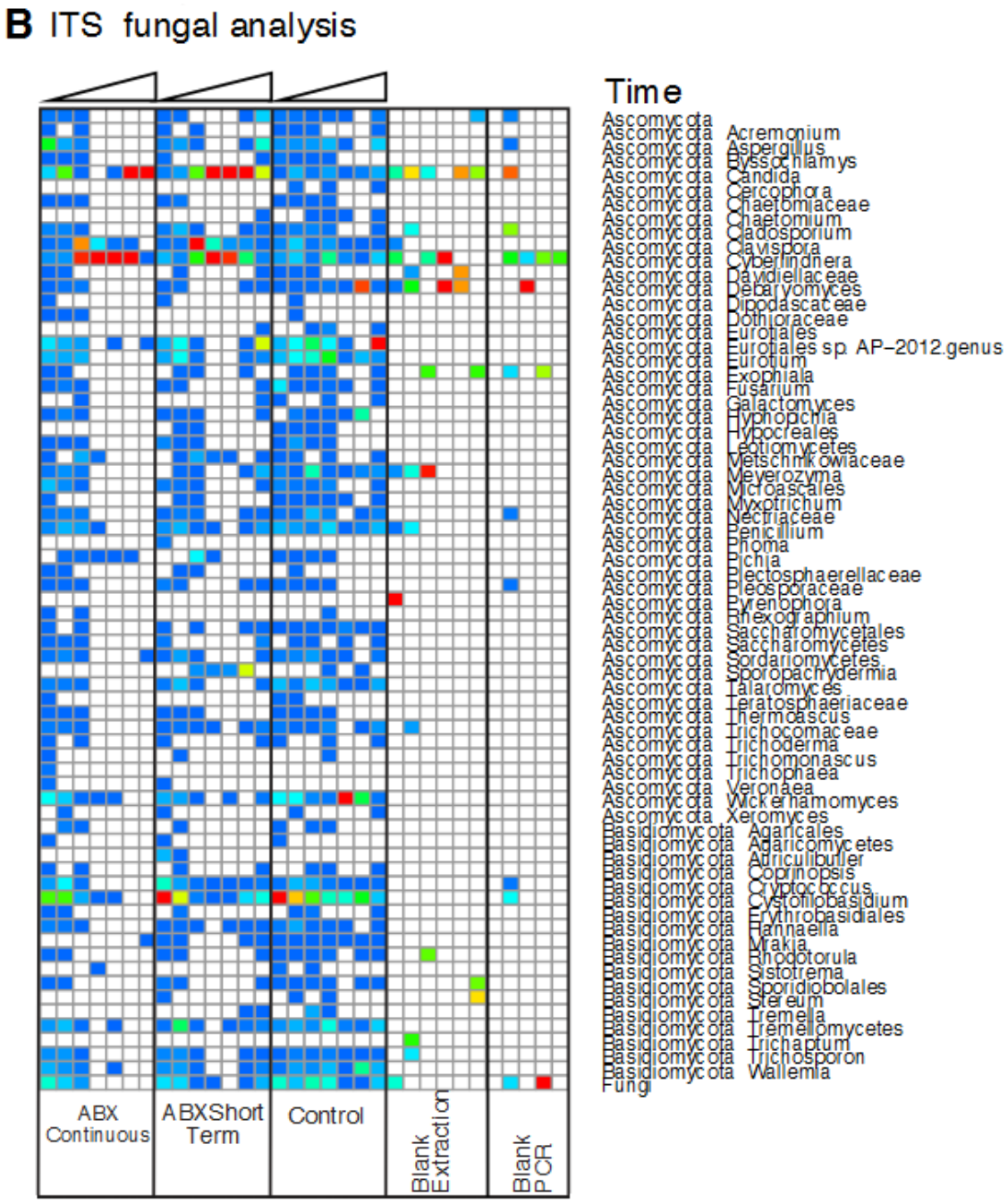
Supplementary Figure 4-4. Heat maps showing the composition of fungal communities inferred from ITS sequence data for each time point (A-G), with each mouse shown individually. The scale of relative proportions is shown on the right.



Supplementary Figure 4-5. PCoA analysis distances measures for bacterial (16S sequence data) and fungal (ITS sequence data) communities. Distances matrices were calculated using weighted or unweighted UniFrac, then the pairwise distances between a treatment group and the control group compared to the distances within the control group on that day. Asterisks above each box and whisker plot indicate whether the comparison was significantly different.

A 16S rRNA gene analysis





Supplementary Figure 4-6. Comparison of contamination controls to experimental samples for the 16S A) and ITS B) amplicons. "Extraction control" indicates sequences derived from blank purifications using DNA-free water. Each column showing mouse data is an average over all reads at that time point.

Primer Name	Sequence	Amplicon	Clade Targeted
Euk_NSR399	TCTCAGGCTCCYTCTCCGG	18S qPCR, 454	Microeukaryotes
18S-67a-deg	AAGCCATGCATGYCTAAGTATMA	18S qPCR, 454	Microeukaryotes
ITS1Fungal	CTTGGTCATTTAGAGGAAGTAA	ITS1 454	Fungi
ITS2	GCTGCGTTCTTCATCGATGC	ITS1 454	Fungi
BSR357	CTGCTGCCTYCCGTA	16S 454	Bacteria
BSF8	AGAGTTTGATCCTGGCTCAG	16S qPCR, 454	Bacteria
BSR65	FAM-TA+ACA+CATG+CA+AGTC+GA-BHQ1	16S qPCR probe	Bacteria

+ precedes the position of LNA base

Supplementary Table 4-1. Oligonucleotides used in this chapter.

Sample	DNA yield per pellet; average (ng)	DNA yield per pellet (ng); (SD)	Number of Bacterial genomes/ pellet; average	Number of Bacterial genomes/ pellet; (SD)	Number of microeukaryote genomes/ pellet; Average	Number of microeukaryote genomes/pellet; SD
Control (Baseline)	1.08E+03	4.47E+02	3.95E+07	2.14E+07	1.22E+04	5.14E+03
Control (Day 15)	6.27E+02	4.02E+02	1.10E+06	9.99E+05	8.14E+03	5.34E+03
Control (Day 76)	4.50E+02	2.72E+02	1.19E+06	9.05E+05	4.62E+03	3.89E+03
AbxShortTerm (Baseline)	4.51E+02	1.56E+02	1.06E+07	6.26E+06	2.93E+03	1.82E+03
AbxShortTerm (Day 15)	1.13E+02	7.42E+01	9.89E+01	8.97E+01	1.47E+05	5.69E+04
AbxShortTerm (Day 76)	3.94E+02	1.88E+02	7.01E+05	5.69E+05	2.20E+03	1.71E+03
ABXContinuous (Baseline)	4.39E+02	1.79E+02	9.48E+06	7.28E+06	1.90E+03	8.07E+02
ABXContinuous (Day 15)	1.48E+02	7.32E+01	1.72E+02	1.19E+02	3.24E+05	1.58E+05
ABXContinuous (Day 76)	1.08E+02	1.32E+01	3.82E+02	3.38E+02	1.00E+05	3.27E+04

Supplementary Table 4-2. DNA yields and numbers of genomes inferred from ribosomal gene copies data.

	Amount	ratio (no Abx/Abx)
Average DNA yield no Abx (ng)	6.10E+02	4.68E+00
Average DNA yield day 15 Abx (ng)	1.30E+02	
Average bacterial genomes no Abx	1.24E+07	9.13E+04
Average bacterial genomes day 15 Abx	1.35E+02	
Average fungal genomes no Abx per pellet	5.95E+03	2.53E-02
Average fungal gneomes day 15 Abx per pellet	2.35E+05	
Average DNA yield no Abx (ng)	6.10E+02	5.65E+00
Average DNA yield day 76 Abx (ng)	1.08E+02	
Average bacterial genomes no Abx per pellet	1.24E+07	3.24E+04
Average bacterial genomes day 76 Abx per pellet	3.82E+02	
Average fungal genomes no Abx	5.95E+03	5.95E-02
Average fungal gneomes day 76 Abx	1.00E+05	

Supplementary Table 4-3. Ratios for DNA yields, average bacterial genomes, and average fungal genomes for selected comparisons between groups.

CHAPTER 5: Conclusions and Future Directions

In this dissertation I present novel methods for characterizing understudied clades of the gut microbiome and an application that highlights the utility and necessity for such methods. Species in the gut interact across domains, forming a single community [1]. Bacteria, archaea, and eukarya all contain different metabolic capabilities, which can be strung together to form metabolic pathways with different steps taking place in different species. Nevertheless, most studies have focused on studying the bacterial community in isolation, which provides a useful yet incomplete picture of microbiome dynamics. Subsequently, we developed methods to characterize the archaeal and eukaryotic components and explore the selective effects of antibiotic treatment across bacterial and eukaryotic gut communities.

In Chapter 2, I present a sample-to-analysis pipeline for characterizing microeukaryotic communities within the gut microbiome. Development of a targeted assay to study microeukaryotes is important, because there is evidence that microeukaryotes, and particularly fungi, are underrepresented in metagenomic sequencing assemblies [2]. This method was particularly challenging to develop because of the homology between the host genome and the genomes of the commensal microeukaryotes. We tested several primers, looking at breadth and specificity. Ultimately, 18S rDNA and ITS primers proved to capture their targeted communities accurately. The 18S amplicon detected a wide range of fungi and protozoa. The ITS amplicon only detected fungi but provided superior differentiation between closely related species. We also tested several DNA extraction methods, and determined that harsher lysis methods improve fungal DNA yield, as has been corroborated by others [3]. In addition to wet side methods, I also developed the BROCC program to automate taxonomic attribution of microeukaryotic sequences through voting on high quality, classified BLAST hits. Ultimately, this work provides a framework for future studies of the eukaryotic components of the gut microbiome.

While this study made great strides in understanding the fungal and protozoan gut community, community characterization was hampered by poorly curated databases. Specifically, many OTUs matched database sequences with high identity, but these database sequences were classified only to the phylum or kingdom level (i.e. 'uncultured fungi'). Many of these sequences likely come from understudied microeukaryotic lineages. Recently, several projects have produced useful databases characterizing the genomes of eukaryotic species of interest to humanity [4,5,6]. The creation of a large amplicon specific database with good coverage of all known lineages, particularly of the fungi, for the ITS gene would greatly aid future studies.

In Chapter 3, I present a wet lab workup protocol to characterize archaeal communities in the gut microbiome. This method was developed through testing different primer pairs, thermocycling conditions, and extraction techniques. Through limited Sanger sequencing I found *Methanobrevibactor* in human and macaque samples and *Methanobrevibactor*, *Methanococcus*, and *Thermogymnomonas* in murine samples. Previously, archaeal communities were characterized indirectly through qPCR [7] or metagenomic shotgun sequencing [8]. These techniques are not ideal for community characterization because qPCR does not easily differentiate between different species, and metagenomic sequencing is vulnerable to coverage related assembly problems [9], which can be exacerbated in rare community members such as archaea. The method I developed enables direct determination of the identity of different members of an archaeal microbiome community.

In this study, we detected the acidophilic genus *Thermogymnomonas* in stool. As was observed in both Chapters 2 and 4, genomic DNA is capable of surviving a trip through the digestive tract and being detected in stool. This is a key observation, because many members of the *Thermogymnomonas* genus and its class *Thermoplasmata* live between pH 0 and 4 [10,11,12], and the type species for the genus *Thermogymnomonas acidicola* grow optimally at pH 3[13].

Yet, species classified as *Thermogymnomonas* were detected in stool samples which are generally thought to be a good representation of the lower GI tract, even though the colon has a pH close to 7 [14,15]. Subsequently, it is important to determine whether the *Thermogymnomonas* detected was actually inhabiting the colon through some evolutionary adjustments that allow it to survive at higher pH levels, or was inhabiting the stomach, with an approximate pH between 1 and 5 depending on the location and presence of food [16], which would be much more habitable for an acidophile. Currently, the only microbe known to commonly inhabit the stomach is *Helicobacter pylori* [17], which can cause ulcers and other gastric disease or live asymptomatically [18]. Discovery and characterization of additional stomach microbiome inhabitants could provide more information about the causes of gastric disease and aid its treatment.

In Chapter 4, I analyze the longitudinal effects of heavy antibiotic use during and after treatment in a mouse model using a cocktail of ampicillin, neomycin, vancomycin, and metronidazole. We found that this mixture of antibiotics effectively suppresses the commensal bacterial community to the point where the only detectable species is *Lactococcus lactis*, which is likely derived from food manufacturing. Eight weeks after antibiotic treatment was stopped, many bacterial groups that were common before treatment returned, and cell counts returned to their previous levels. However, species composition and diversity were significantly different from untreated controls, indicating potential long term perturbation of the bacterial community. Under antibiotic treatment, we also observed three species of *Saccharomycetales* yeasts grow out in succession, increasing the fungal genome count by one to two orders of magnitude. After cessation of antibiotic treatment, fungal counts returned to normal levels within one week and to normal diversity and community composition 8 weeks later compared to untreated controls. Unexpectedly we also observed that fungal communities in the controls rapidly change

composition without intervention, indicating the fungal gut community is less stable than bacterial gut community.

The effects of heavy antibiotic use are not well characterized in the human gut. I would expect different effects from our findings in Chapter 4 because unlike laboratory mice, most humans have had prior antibiotic exposure, and studies have indicated that the human gut is a reservoir for antibiotic resistance genes [19]. Heavy antibiotic use may have a less potent effect on humans than naïve mice. Also, as we saw in Chapter 2, many of the fungal sequences in human studies are suspected to be derived food sources, and it is unclear if these cells are alive and viable. Suppression of the bacterial community should cause a fungal outgrowth, but it is unknown if *Candida albicans*, which has been documented to grow out with antibiotic treatment, would grow out alone or with other fungi because other yeasts have been documented to grow out as well [20]. It would be informative to see if any species that are thought to be food derived grow out as well. Species like *Saccharomyces cerevisiae* have been documented to cause infection in rare cases [21], but it is unknown if it is a dynamic member of the gut community.

Furthermore, the longitudinal effects of narrower spectrum and single antibiotic treatments on the gut microbiota still need to be studied. Many antibiotic regimens prescribed today consist of a single antibiotic used for a short time period to resolve acute bacterial infection [22]. However, long term side effects have been associated with antibiotic use, such as obesity [23]. Understanding the potential long term effects of more typical treatments on the gut microbiome may aid clinicians when prescribing antibiotics.

This dissertation presents a framework for characterizing understudied clades of the microbiome and direct experimentation through use of antibiotics. The methods developed here enable researchers to characterize archaea, fungi, and protozoans in microbiome samples. The

antibiotic study not only describes the effects of heavy antibiotic treatment on bacterial and fungal communities, but it provides a framework for analyzing future antibiotic microbiome studies, and highlights the quirks of using such a model, such as detection of bacterial DNA in food and observing a succession of fungal species dominate the gut, thereby enabling researchers to avoid potential pitfalls. This research can aid a wide variety of future studies of microbial and microbiome communities within the fields of microbiology, immunology, and even ecology.

References

1. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA (2008) Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. Nature Reviews Microbiology 6: 121-131.

2. Morgan JL, Darling AE, Eisen JA (2010) Metagenomic sequencing of an in vitro-simulated microbial community. PloS one 5: e10209.

3. Latgé JP (2007) The cell wall: a carbohydrate armour for the fungal cell. Molecular microbiology 66: 279-290.

4. Spatafora JW (2005) Assembling the fungal tree of life (AFTOL). Mycological Research 109: 755-756.

5. Stajich JE, Harris T, Brunk BP, Brestelli J, Fischer S, et al. (2012) FungiDB: an integrated functional genomics database for fungi. Nucleic acids research 40: D675-D681.

6. Aurrecoechea C, Brestelli J, Brunk BP, Fischer S, Gajria B, et al. (2010) EuPathDB: a portal to eukaryotic pathogen databases. Nucleic acids research 38: D415-D419.

7. Dridi B, Henry M, El Khechine A, Raoult D, Drancourt M (2009) High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. PLoS ONE 4: e7063.

8. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science (New York, NY) 312: 1355-1359.

9. Chuang Howe A, Pell J, Canino-Koning R, Mackelprang R, Tringe S, et al. (2012) Illumina Sequencing Artifacts Revealed by Connectivity Analysis of Metagenomic Datasets.

10. Dopson M, Baker-Austin C, Hind A, Bowman JP, Bond PL (2004) Characterization of Ferroplasma isolates and Ferroplasma acidarmanus sp. nov., extreme acidophiles from acid mine

drainage and industrial bioleaching environments. Applied and environmental microbiology 70: 2079-2088.

11. Golyshina OV, Yakimov MM, Lünsdorf H, Ferrer M, Nimtz M, et al. (2009) Acidiplasma aeolicum gen. nov., sp. nov., a euryarchaeon of the family Ferroplasmaceae isolated from a hydrothermal pool, and transfer of Ferroplasma cupricumulans to Acidiplasma cupricumulans comb. nov. International journal of systematic and evolutionary microbiology 59: 2815-2823.

12. Schleper C, Puehler G, Holz I, Gambacorta A, Janekovic D, et al. (1995) Picrophilus gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0. Journal of bacteriology 177: 7050-7059.

13. Itoh T, Yoshikawa N, Takashina T (2007) Thermogymnomonas acidicola gen. nov., sp. nov., a novel thermoacidophilic, cell wall-less archaeon in the order Thermoplasmatales, isolated from a solfataric soil in Hakone, Japan. International journal of systematic and evolutionary microbiology 57: 2557-2561.

14. Cummings J, Pomare E, Branch W, Naylor C, Macfarlane G (1987) Short chain fatty acids in human large intestine, portal, hepatic and venous blood. Gut 28: 1221-1227.

15. Pye G, Evans D, Ledingham S, Hardcastle J (1990) Gastrointestinal intraluminal pH in normal subjects and those with colorectal adenoma or carcinoma. Gut 31: 1355-1357.

16. Simonian HP, PHRMD LV, Doma S, Fisher RS, Parkman HP (2005) Regional postprandial differences in pH within the stomach and gastroesophageal junction. Digestive diseases and sciences 50: 2276-2285.

17. Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. Nat Rev Genet 13: 260-270.

18. Dooley CP, Cohen H, Fitzgibbons PL, Bauer M, Appleman MD, et al. (1989) Prevalence of Helicobacter pylori infection and histologic gastritis in asymptomatic persons. The New England journal of medicine 321: 1562.

19. Sommer MO, Dantas G, Church GM (2009) Functional characterization of the antibiotic resistance reservoir in the human microflora. Science (New York, NY) 325: 1128-1131.

20. Samonis G, Gikas A, Anaissie EJ, Vrenzos G, Maraki S, et al. (1993) Prospective evaluation of effects of broad-spectrum antibiotics on gastrointestinal yeast colonization of humans. Antimicrobial agents and chemotherapy 37: 51-53.

21. Muñoz P, Bouza E, Cuenca-Estrella M, Eiros JM, Pérez MJ, et al. (2005) Saccharomyces cerevisiae fungemia: an emerging infectious disease. Clinical infectious diseases 40: 1625-1634.

22. Gonzales R, Malone DC, Maselli JH, Sande MA (2001) Excessive antibiotic use for acute respiratory infections in the United States. Clinical infectious diseases 33: 757-762.

23. Raoult D (2008) Obesity pandemics and the modification of digestive bacterial flora. European Journal of Clinical Microbiology & Infectious Diseases 27: 631-634.

APPENDIX

Appendix 1- Plasmids Developed During Thesis Project

Acession Number	Gene	Species	Plasmid Backbone
1143	18S-Near Full	Candida	Topo4
1142	18S-Near Full	Candida	Topo4
1145	ITS1, 5.S, and ITS2	Candida	Topo4
1146	16S -Near Full	Methanococcus maripaludis	Topo4
1147	16S -Near Full	Sulfolobus solfataricus	Topo4
1144	16S-Partial	Methanobrevibactor	Topo4