Title

Genome reconstruction and characterization of microbial eukaryotes in complex microbial communities through genome-resolved metagenomics

Author

West, Patrick

Publication Date

2020

Peer reviewed|Thesis/dissertation

Genome reconstruction and characterization of microbial eukaryotes in complex microbial communities through genome-resolved metagenomics

By

Patrick T. West

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jillian F. Banfield, Chair
Professor N. Louise Glass
Professor Michael Freeling

Spring 2020

Abstract

Genome reconstruction and characterization of microbial eukaryotes in complex microbial communities through genome-resolved metagenomics

By

Patrick T. West

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian Banfield, Chair

Microbial eukaryotes are important pathogens, environmental quality indicators, integral components of natural microbial communities, and critical for understanding our own evolutionary history. Yet, microbial eukaryotes are an often neglected component of microbial ecology studies. Common metagenomic techniques, such as 16S rRNA gene sequencing, fully omit eukaryotes, and they are frequently ignored in shotgun-metagenomic sequencing projects. A methodology was developed for recovering eukaryotic genomes from metagenomes that relies upon a newly developed machine learning-based method, EukRep, to separate Eukaryotic scaffolds from prokaryotic scaffolds prior to binning. In this way, eukaryotic gene predictors can be applied to eukaryotic scaffolds, eliminating one of the largest challenges to properly binning eukaryotes in shotgun metagenomic samples. The effectiveness of EukRep was tested on both mock communities constructed from reference bacterial, archaeal, and eukaryotic genomes *in silico* as well as on natural microbial community samples and shown to enable the recovery of near-complete eukaryotic genomes including high-quality fungal, protist, and rotifer genomes from complex environmental samples. Thus, this approach enables consistent genome reconstruction and prediction of metabolic and behavioral potential for eukaryotes as well as their associated communities in a culture independent, natural microbial community context.

A EukRep-based approach was used to investigate the effect of addition of organic carbon to a geyser-associated microbial community. Crystal Geyser, a CO2-driven geyser in Utah (USA), provides large volumes of deeply sourced fluids, thus is well suited for studying microbial communities in high CO2 environments. Upon addition of organic carbon there was a substantial change of the community metabolism, with selection against almost all candidate phyla bacteria and archaea and for eukaryotes. Near complete genomes were reconstructed for three fungi placed within the Eurotiomycetes and an arthropod. While carbon fixation and sulfur oxidation were important functions in the geyser community prior to carbon addition, the organic carbon-impacted community showed enrichment for secreted proteases, secreted lipases, cellulose targeting CAZymes, and methanol oxidation. The results demonstrate the broader utility of EukRep for reconstruction and evaluation of relatively high-quality fungal, protist, and rotifer genomes from complex environmental samples. This approach opens the way for cultivation-independent analyses of whole microbial communities.

Fungi are common members of the human microbiome, but are often excluded from metagenomic studies due to the large size and complexity of Eukaryotic genomes. Here, targeted Eukaryotic genome recovery was performed on over a thousand metagenomes from premature infant fecal samples and twenty-eight metagenomes from the neonatal intensive care unit (NICU) housing the infants. Samples were screened for the presence of Eukaryotes using a machine learning classifier, and *de novo* genome assembly, curation, and annotation was performed on identified samples. Seventeen distinct Eukaryotic genomes were recovered (median completeness 91%; median size 15.6 Mbp), including genomes from four strains of *Candida albicans*, seven genera of fungi, and two organisms (Diptera (fly) and Rhabditid (nematode)) with no previously sequenced genomes of the same family. Seven percent of infants were colonized by a Eukaryote during the first months of life, and prevalence was significantly associated with administration of maternal antibiotics and particular bacterial taxa. All NICU samples had detectable fungal communities (median relative abundance 2%, full range 0.3-24.1%), and different locations in the NICU had distinct Eukaryotic microbiomes. Near-identical genomes of *Purpureocillium lilacinum* were recovered from both infant and NICU samples (99.999% average nucleotide identity), highlighting the potential for environmental NICU fungi to colonize premature infants. Zygosity and potential aneuploidy were determined for all assembled genomes, and regions with loss of heterozygosity (indicative of recent genome evolution) were detected in some *C. albicans* genomes. This study resolved Eukaryote dynamics in the NICU and premature infant gut samples, and reveals potential reservoirs of unexpected eukaryotic diversity within the hospital environment.

Candida parapsilosis is the third most common cause of invasive candidiasis. *C. parapsilosis* infections have been continually increasing in prevalence over the past two decades, and at significantly higher prevalence in neonates than other at risk populations, marking its importance as an emerging pathogen. Despite this, *C. parapsilosis* is understudied. The recovered C. *parapsilosis* genomes contain small genomic regions with highly elevated levels of Single Nucleotide Variants (SNVs), which we refer to as SNV hotspots. SNV hotspots are shared between strains, with some unique to *C. parapsilosis* strains from a single hospital. Four of the *C. parapsilosis* genomes have a high copy number (4-16) RTA3 gene, a lipid translocase previously implicated in antifungal resistance, potentially indicative of adaptation to antifungal treatment. Additionally, time course metatranscriptomics and metaproteomics were performed on a premature infant with a documented *C. parapsilosis* blood infection, offering a rare look at the *in vivo* expression and protein landscape of a Candida species. *C. parapsilosis in situ* expression is highly distinct from culture settings, but also highly variable, demonstrating the importance of studying Candida *in situ* in addition to culture settings.

Mono Lake, CA, is a high alkalinity, hypersaline lake with an unusually productive ecosystem largely supported by benthic and planktonic algae. A species of choanoflagellate from Mono Lake that forms a multicellular, hollow rosette filled with bacteria, but little was known about this choanoflagellate and its associated microbial community. This association is of interest given choanoflagellates are the closest living relatives to animals and the analogy between rosette-enclosed consortia and animal gut microbiomes. Metagenomic shotgun sequencing was performed in order to reconstruct genomes for the choanoflagellete and its associated community. EukRep was used for eukaryotic sequence identification and enabled genome recovery, genome completeness evaluation and prediction of metabolic potential of both the choanoflagellate nuclear and mitochondrial genomes. The nuclear draft genome measures 49

Mbp in length, contains 11052 predicted genes and appears to be near complete. Interestingly, its extracellular proteins have a higher isoelectric point compared to marine choanoflagellates, likely an adaptation to their saline, high pH environment. Characterization of bacterial communities leveraged samples taken from choanoflagellate rosette enriched and choanoflagellate rosette depleted samples in order to distinguish bacteria within and outside rosettes. Across all samples, 23 near-complete bacterial genomes were recovered, primarily belonging to Gammaproteobacteria, Bacteroidetes, and Spirochaeta. Of these, seven were found only in the choanoflagellete enriched samples, suggesting that these bacteria are partitioned into the rosette interior. Overall, the research provided insights into the composition and metabolic interactions between an ordered assemblage of single celled eukaryotes and its enclosed microbiome.

In this work, genome-resolved and culture-independent methods are employed to study microbial eukaryotes in a variety of natural community contexts, ranging from animal microbiomes, the hospital room, and environmental communities. The development of EukRep and subsequent incorporation into metagenomic pipelines represents an important methodological advance for the comprehensive study of the structure and ecology of natural microbial communities and provides new insights into community functioning.

Table of Contents

Introduction

Microbial eukaryotes are the evolutionary connection between us, and all the macroscopic life around us, to the microscopic world of unicellular organisms and our ancient, ancient ancestors. Two very open, and very large questions are how the domain of Eukaryotes emerged from the relatively simple, unicellular Prokaryotes and the identity of the Last Eukaryotic Common Ancestor (LECA). The recently discovered superphylum of Asgard Archaea, are the closest known living prokaryotes to eukaryotes, and suggest an archaeal host cell and an alphaproteobacterial (mitochondrial) endosymbiont as the origin of eukaryotic cellular complexity (Zaremba-Niedzwiedzka et al. 2017). Both phylogenetic analyses and the protein content of the Asgard archaea support this hypothesis. Interestingly, however, no clear phylogenetic root within the Eukaryotic domain has been resolved (Burki 2014; Williams 2014), leaving the identity of LECA a mystery.

Over the past two decades, culturing of microbial eukaryotes with no near phylogenetic relatives (e.g., the Picozoa; Seenivasan et al. 2013, and Ancyromonads; Janouškovec et al. 2017), as well as the recent discovery of the Hemimastigophora phylum (Lax et al. 2018), indicate that a large fraction of microbial eukaryotes remain to be described. In fact, not only is LECA currently undetermined, but entire eukaryotic phylums of diversity likely remain to be discovered (Keeling et al. 2019). Similar to bacteria and archaea, only a very small fraction of microbial eukaryotes can currently be isolated in culture (Caron et al. 2008; Pawlowski et al. 2012). Cultivation-independent methods have the potential to provide an expanded view of eukaryotic diversity, as they have done across the Domain Bacteria and Archaea (Hug et al. 2016).

Amplicon sequencing, such as 18S rRNA sequencing is a cultivation-independent method that has helped to reveal the breadth of eukaryotic diversity. However, this gene does not contain enough information to resolve deep phylogenies and does not provide information about metabolism or lifestyle of the identified organisms. Genome-resolved metagenomics is a promising, culture independent method of recovering whole genomes from environmental sequencing samples. In this method, DNA is extracted from whole communities and shotgun sequenced. Sequencing reads are then assembled *de novo* into scaffolds and subsequently binned into putative genomes. Recovering entire genomes is highly advantageous as it allows for detailed multi-protein phylogenies, as well as prediction of metabolic and lifestyle potential based on gene content.

Yet, only a handful of eukaryotic genomes have been recovered with genome-resolved metagenomics (Sharon et al. 2013; Kantor et al. 2015, 2017; Quandt et al. 2015; Mosier et al. 2016; Raveh-Sadka et al. 2016). This conspicuous absence of eukaryotic genomes can partially be attributed to the need for proper gene predictions for high quality binning, and the failure of standard metagenomic gene prediction algorithms to properly predict genes on eukaryotic scaffolds. In Chapter 1, I propose a modified genome-resolved metagenomic method incorporating EukRep, a machine learning-based algorithm for separating eukaryotic scaffolds from prokaryotic scaffolds, to aid in binning eukaryotes. I show it reliably enables proper gene

prediction on eukaryotic scaffolds and subsequently, the recovery of diverse eukaryotic genomes from a range of environments and *in silico* experiments.

In addition to their critical role in understanding our own evolutionary history, microbial eukaryotes are important opportunistic pathogens and members of the human microbiome. Although human diseases caused by fungi is a highly active area of research, relatively little is known about asymptomatic colonization of the gut early in life. Studies with varying methods have reported 0%, 26%, 50%, and 63% of premature infants being colonized by fungi (Baley et al. 1986; Stewart et al. 2012; Stewart et al. 2013; LaTuga et al. 2011). However, due to methodological limitations, all of these studies were only able to analyze the fungal components of the communities and thus, unable to answer basic questions about fungal abundance relative to gut bacterial community members. In addition, the extent to which the hospital environment serves as a reservoir for microbial eukaryotes and a source of colonization of infants is currently unknown. In chapters 2 and 3, I utilize genome-resolved metagenomics to characterize the presence of microbial eukaryotes in the infant gut and neonatal intensive care unit (NICU). We show microbial eukaryotes are diverse and frequently present in the gut microbiomes of infants, and that there may be transfer between infants and hospital rooms or vice versa.

Microbial eukaryotes exist in complex communities comprised of bacteria, archaea, and viruses; however, relatively little is known about their biology and behavior in a whole community context. Microbial community context is likely of high significance, given interactions between bacteria and eukaryotes influence the development, metabolism, and evolution of all types of eukaryotes, ranging from animals (Gilbert et al. 2018) to unicellular ciliated protozoa (Gong et al. 2016).

Indeed, microbial eukaryotes have been shown to be dramatically influenced by bacteria. The Choanoflagellates, the closest known living relatives of animals, typically graze on bacteria by trapping them in their apical collar (Hibberd et al. 1975). The Choanoflagellate *Salpingoeca rosetta* is primarily a unicellular organism, but formation of multicellular rosettes can be both induced or inhibited by lipids produced by the bacterium *Algoriphagus machipongonensis* (Cantley et al. 2016). Similarly, the bacterium *Vibrio fischeri* produces a chondroitinase, EroS, capable of inducing mating in *S. rosetta* (Woznica et al. 2017). As another example, the yeast *Candida albicans*, a human commensal and opportunistic pathogen, exhibits complex interactions with *Enterococcus faecalis*, a bacterial human gut commensal. *C. albicans* and *E. faecalis* negatively impact one another's virulence (Cruz et al. 2013), suggesting a mechanism that promotes commensal behavior in a gut microbial community context. The decrease in *C. albicans* virulence was attributed to inhibition of hyphal morphogenesis and by proteases secreted by *E. faecalis* (Cruz et al. 2013). However, these given examples of interactions were studied using constructed experiments rather than natural communities. In chapter 3, I utilize metatranscriptomics and metaproteomics combined with metagenomics of infant fecal samples to examine the metabolism and behavior of the yeast *Candida parapsilosis* in an infant gut context and show that it is significantly altered from what is typically observed in culture settings. In chapter 4, a complex co-culture derived from a Mono Lake water sample containing over 25 distinct species, including a microbial eukaryote belonging to the Choanoflagellates, is characterized.

Acknowledgments

The first person I would like to thank is my advisor, Jill Banfield., who has so generously poured countless hours of time and energy into my development as a person and a scientist. I am so grateful to have had an advisor that is as supportive of their students as Jill has been. She has taught me how to find common interests and collaborate with others, to write effectively and efficiently, exactly how to find what's interesting in a mountain of data, and many other skills I will take with me in life and my career.

I would like to thank all of the members of the Banfield lab for their bioinformatics support and mentorship. The lab has been a second home for me these past five years and my colleagues are who have made it that way. In particular, I would like to thank Chris Brown, Alex Probst, Karthik Anantharaman, Evan Starr, and Alex Crits-Christoph for their advice as well as their patience and willingness to share their expertise in microbiology with me. Christine He for her friendship, support, and the frequent drawing and painting sessions that have helped me stay sane. And finally, Matt Olm for being a great friend, roommate, and mentor.

To Kyle, who has been here for me from the beginning. Thank you for your emotional support and all that you've done for me. Your love has represented a steady foundation for me during these often stressful and turbulent years. I'm glad to have shared the hikes, camping trips, vacations, and my day to day life with you and have been incredibly lucky to have you here with me.

To my family, thank you for your love and support. You've pushed me through the years to be my best and I wouldn't be here without your guidance and support. Mom, the value you place on education and critical thinking, Dad the dedication and responsibility you display, and Mari the care and love you show for others have all helped shape me to be the person I am today.

I would like to thank Nathan Springer, who's lab was my introduction to academic research. Although I was an undergraduate student at the time, he treated me with respect as a colleague and gave me more opportunity than I could have hoped for. I would also like to thank Steve Eichten for introducing me to the field of bioinformatics and encouraging me to pursue it despite my initial hesitance. In addition, thank you to Mandy Waters, Peter Hermanson, and Qing li for their mentorship.

Much of the work presented here would not have been possible without the contributions of collaborators that have worked with Jill and I over the course of my PhD. Igor Grigoriev, Mike Morowitz, Sandy Johnson, Nicole King, and Bob Hettich have all generously shared their time, knowledge, and resources to make these projects a reality.

Thank you to many others for their friendship and emotional support, especially Andy for engaging with me in my many hobbies and always taking my ideas seriously, no matter how ridiculous; Antoine, for the late night conversations and hang outs; and Kris for ensuring I do not take Zellerbach Hall performances for granted.

1 Genome-reconstruction for eukaryotes from complex natural microbial communities

West, Patrick T., Probst, Alexander J., Grigoriev, Igor V. Thomas, Brian C. and Banfield, Jillian F.

Published in *Genome Research*, April 2018, doi: 10.1101/gr.228429.117

1.1 Abstract

Microbial eukaryotes are integral components of natural microbial communities, and their inclusion is critical for many ecosystem studies, yet the majority of published metagenome analyses ignore eukaryotes. In order to include eukaryotes in environmental studies, we propose a method to recover eukaryotic genomes from complex metagenomic samples. A key step for genome recovery is separation of eukaryotic and prokaryotic fragments. We developed a k-mer-based strategy, EukRep, for eukaryotic sequence identification and applied it to environmental samples to show that it enables genome recovery, genome completeness evaluation, and prediction of metabolic potential. We used this approach to test the effect of addition of organic carbon on a geyser-associated microbial community and detected a substantial change of the community metabolism, with selection against almost all candidate phyla bacteria and archaea and for eukaryotes. Near complete genomes were reconstructed for three fungi placed within the Eurotiomycetes and an arthropod. While carbon fixation and sulfur oxidation were important functions in the geyser community prior to carbon addition, the organic carbon-impacted community showed enrichment for secreted proteases, secreted lipases, cellulose targeting CAZymes, and methanol oxidation. We demonstrate the broader utility of EukRep by reconstructing and evaluating relatively high-quality fungal, protist, and rotifer genomes from complex environmental samples. This approach opens the way for cultivation-independent analyses of whole microbial communities.

1.2 Introduction

Microbial eukaryotes are important contributors to ecosystem function. Gene surveys or DNA "barcoding" are frequently used to identify eukaryotes in microbial communities and have demonstrated the breadth of eukaryotic diversity (Pawlowski et al. 2012). However, these approaches can only detect species and are unable to provide information about metabolism or lifestyle in the absence of sequenced genomes. The majority of fully sequenced eukaryotic genomes are from cultured organisms. Lack of access to cultures for a wide diversity of protists

and some fungi detected in gene surveys has resulted in major gaps in eukaryotic reference genome databases (Caron et al. 2008; Pawlowski et al. 2012). Single-cell genomics holds promise for sequencing uncultured eukaryotes and has generated partial genomes for some (Cuvelier et al. 2010; Yoon et al. 2011; Monier et al. 2012; Vaulot et al. 2012; Roy et al. 2014; Mangot et al. 2017). However, multiple displacement amplification limits the completeness of single-cell genomes (Woyke et al. 2010). Alternatively, metagenomic sequencing reads from environmental samples are mapped against reference genomes to detect organisms and constrain metabolisms, but this approach is restricted to study of organisms with sequenced relatives. Many current studies of natural ecosystems and animal or plant-associated microbiomes use an untargeted shotgun sequencing approach.

When the DNA sequences are assembled, tens of thousands of genome fragments may be generated, some of which are derived from eukaryotes. Exceedingly few metagenomic studies have systematically identified such fragments as eukaryotic, although some genomes for microbial eukaryotes have been reconstructed (Sharon et al. 2013; Kantor et al. 2015, 2017; Quandt et al. 2015; Mosier et al. 2016; Raveh-Sadka et al. 2016). In almost all cases, these genomes were recovered from relatively low-diversity communities where binning of genomes is typically less challenging than in complex environments. Here, we applied a new k-mer-based approach for identification of assembled eukaryotic sequences in datasets from diverse environmental samples. Identification of eukaryotic genome fragments enabled their assignment to draft genomes and improvement of the quality of gene predictions. Predicted genes on assembled metagenomic contigs provide critical inputs for further binning decisions that incorporate phylogenetic profiles as well as classification of the reconstructed genomes and assessment of their completeness. Our analyses focused on biologically diverse environmental samples, many of which came from groundwater. In addition, we investigated previously published metagenomes from infant fecal samples and a bioreactor community used to break down thiocyanate. Because the approach works regardless of a predetermined phylogenetic affiliation, it is now possible to reconstruct genomes for higher eukaryotes as well as fungi and protists from complex environmental samples.

1.3 Materials and methods

1.3.1 Crystal Geyser sample collection and DNA extraction

Details of filtration of groundwater for sample CG_bulk is given in Probst et al. (2016) (sample CG23_combo_of_CG06-09_8_20_14). Groundwater containing particulate wood was collected in a 50- mL Falcon tube. All samples were frozen on site on dry ice and stored at −80°C until further processing. The sample with the particulate wood was spun down, and DNA extraction was performed as described previously (Emerson et al. 2015).

1.3.2 Crystal Geyser DNA sequencing and assembly

Raw sequencing reads were processed with bbtools (http://jgi.doe. gov/data-and-tools/bbtools/) and quality-filtered with SICKLE Figure 6. Comparison of CG_WC and CG_bulk metabolic capacity. Log2 ratio of all annotated genes found within the CG_bulk sample against annotated genes found in the CG_WC sample. Annotated genes were grouped into categories based upon

scores with a custom set of metabolic pathway marker HMMs (Anantharaman et al. 2016), CAZyme HMMs (Cantarel et al. 2009), and protease and lipase HMMs from MEROPs and the Lipase Engineering Database, respectively. Putative proteases and lipases were also filtered to only those containing a secretion signal and less than three transmembrane domains (see Methods). Gene count (red) is the ratio of total number of genes in each category for each sample normalized by the total number of genes found in the sample. Relative abundance (blue) is the ratio of average read coverage depth of the contig containing a given annotated gene in each category normalized by the sample read count multiplied by read length. West et al. 576 Genome Research www.genome.org Downloaded from genome.cshlp.org on March 19, 2020 - Published by Cold Spring Harbor Laboratory Press with default parameters (version 1.21; https://github.com/ najoshi/sickle). IBDA_UD (Peng et al. 2012) was used to assemble and scaffold filtered reads. IDBA_UD was chosen as it is a widely used, publicly available program designed for metagenomic assemblies. Unlike almost all other such assemblers, it includes a scaffolding step. This is important because longer sequences can be more robustly binned. Scaffolding errors were corrected using MISS (I Sharon, BC Thomas, JF Banfield, unpubl.), a tool that searches and fixes gaps in the assembly based on mapped reads that exhibit inconsistencies between raw reads and assembly. The two Crystal Geyser samples used for binning and comparison in this study, CG_WC and CG_bulk, resulted in 874 and 529 Mbps of assembled scaffolds, respectively.

1.3.3 Prokaryotic genome binning and annotations

Protein-coding genes were predicted on entire metagenomic samples using MetaProdigal (Hyatt et al. 2012). Ribosomal RNA genes were predicted with Rfam (Nawrocki et al. 2015), and 16S rRNA genes were identified using SSU-ALIGN (Nawrocki 2009). Predicted proteins were functionally annotated by finding the best BLAST hit using USEARCH (UBLAST) (Edgar 2010) against UniProt (The UniProt Consortium 2017), UniRef90 (Suzek et al. 2007), and KEGG (Kanehisa et al. 2016). Prokaryotic draft genomes were binned through the use of emergent self-organizing map (ESOM)-based analyses of tetranucleotide frequencies. Bins were then refined through the use of ggKbase (ggkbase.berkeley.edu) to manually check the GC, coverage, and phylogenetic profiles of each bin.

1.3.4 EukRep training and testing

EukRep, along with trained linear SVM classifiers, are available at https://github.com/patrickwest/EukRep. A diverse reference set of 194 bacterial genomes, 218 archaeal genomes, 27 opisthokonta, and 43 protist genomes was obtained from NCBI and JGI (Supplemental Table S1). Hug et al. (2016), JGI Mycocosm database (jgi.doe.gov/fungi), and the NCBI taxonomy browser were used as references for selecting genomes from a broad taxonomic range. The contigs comprising these genomes were split into 5-kb chunks for which 5-mer frequencies were calculated (Anvar et al. 2014). Contigs shorter than 3 kb were excluded. The 5-mer frequencies were used to train a linear-SVM (scikit-learn, v. 0.18, default parameters with C = 100) to classify sequences as either of opisthokonta, protist, bacterial, or archaeal origin. The hyperparameter C was optimized using a grid-search with cross-validation and accuracy on a subset of test genomes used for scoring. To classify an unknown or test sequence, the sequence was split into 5-kb chunks, and 5-mer frequencies were determined for each chunk. Contigs

shorter than 3 kb were excluded. The trained classifier was then used to predict whether the sequence is of opisthokonta, protist, bacterial, or archaeal origin. Once classified, the 5-kb chunks were stitched back together into their parent scaffold, and the parent scaffold's taxonomy was determined based upon majority rule of its 5-kb chunks. Accuracy for a given genome was considered to be the percent of total base pairs correctly identified as either eukaryotic or prokaryotic. To compare the effect of k-mer length on prediction accuracy, k-mer frequencies ranging in length from 4 to 6 bp from the same training set were used to train separate linear-SVM models. To determine the minimum sequence length cutoff, test genomes were fragmented into pieces of n length, and sequences shorter than n length were filtered out. To test EukRep, a separate set of 97 eukaryotic and 393 prokaryotic genomes was obtained from NCBI and JGI (Supplemental Table S2). Genomes assembled into less than 10 contigs were fragmented into 100-kb pieces in order to better represent metagenomic data sets. EukRep was then run on each genome individually. Accuracy for a given genome was measured by dividing the total number of base pairs correctly classified by the total number of base pairs tested

1.3.5 EukRep training and testing

Scaffolds predicted to be eukaryotic scaffolds by EukRep were binned into putative genomes using CONCOCT (Alneberg et al. 2014). Eukaryotic genome bins smaller than 5 Mbp were not included in further analyses. Gene predictions were performed individually on each bin with the MAKER2 pipeline (v. 2.31.9) (Holt and Yandell 2011) with default parameters and using GeneMark-ES (v. 4.32) (Ter-Hovhannisyan et al. 2008), AUGUSTUS (v. 2.5.5) (Stanke et al. 2006) trained with BUSCO (v. 2.0) (Simão et al. 2015), and the proteomes of *Chylamydomonas reinhardtii* (Merchant et al. 2007), *Neurospora crassa* (Galagan et al. 2003), and *Reticulomyxa filosa* (Glöckner et al. 2014) for homology evidence. These gene prediction strategies were employed due to their ability to be automatically trained for individual genomes. Completeness of the combined MAKER2 predicted gene set as well as the individual gene predictor gene sets were compared, and the most complete based upon BUSCO analysis was used in future analyses. Phylogenetic classification of the predicted genes along with presence or absence of single-copy orthologous genes was then used to refine each binned genome. CAZYmes were detected in both eukaryotic and prokaryotic bins through the use of HMMER3 (v. 3.1b2) (Eddy 1998) and a set of HMMs obtained from dbCAN (Yin et al. 2012). The presence or absence of various metabolic pathways was determined by using a custom set of metabolic pathway marker gene HMMs (Anantharaman et al. 2016) and HMMER3. Protease and lipases were predicted by using lipase HMMs from the Lipase Engineering Database (Fischer and Pleiss 2003) and BLASTing against a protease database obtained from MEROPS (Rawlings et al. 2016). Putative excreted proteases and lipases were identified by searching for predicted proteases and lipases with secretion signals identified with SignalP (Petersen et al. 2011) and no more than one transmembrane domain with TMHMM (Krogh et al. 2001). To find potentially contaminating prokaryotic scaffolds, predicted genes were BLASTed against UniProt. Scaffolds in which the majority of best hits belonged to prokaryotic genes were removed. Read data sets for previously published metagenomes are available under Sequence Read Archive (SRA) accession numbers SRA052203 and SRP056932 at (SRA; http://www.ncbi.nlm.nih. gov/sra) and BioProjects PRJNA294605 and PRJNA279279.

1.3.6 Eukaryotic gene set comparisons

Nine gene sets were obtained from JGI's mycocosm database (Grigoriev et al. 2011) and NCBI. For each genome, genes were predicted without transcriptomic evidence by running assembled sequences through the MAKER2 pipeline with AUGUSTUS trained with BUSCO and GeneMark-ES in self-training mode. Gene sets predicted with transcriptomic evidence were obtained from the JGI portal and NCBI. For comparison against eukaroytic MetaProdigal predicted gene sets, MetaProdigal was run with the '-meta' flag.

1.3.7 Eukaryote genome completeness estimates

Genome completeness of predicted eukaryotic genomes was estimated based on the presence of conserved, low-copy-number genes. BUSCO (v. 2.0) (Simão et al. 2015) was run with default parameters using the "eukaryota_odb9" lineage set composed of 303 core eukaryotic genes. Completeness was considered to be Metagenomic reconstruction of eukaryotic genomes Genome Research 577 www.genome.org Downloaded from genome.cshlp.org on March 19, 2020 - Published by Cold Spring Harbor Laboratory Press the percent of the total 303 core genes that were present in either single or duplicated copies. Additionally, the number of genes identified as duplicated was used as a way to estimate how much of a given binned genome appeared to be from a single organism.

1.3.8 Eukaryote genome completeness estimates

Bulk soil was collected from the Eel River Critical Zone Observatory (CZO) in Northern California. DNA extraction was performed as described previously (Emerson et al. 2015). Raw sequencing reads were processed with bbtools (http://jgi.doe.gov/data-and-tools/ bbtools/) and quality-filtered with SICKLE with default parameters (version 1.21; https://github.com/najoshi/sickle). IBDA_UD (Peng et al. 2012) was used to assemble and scaffold filtered reads. The genome of Choanephora cucurbitarum was obtained from the NCBI genome database and spiked into the assembled soil metagenome. MetaProdigal was used to obtain gene predictions for the entire sample. EukRep was then used to classify scaffolds as eukaryotic. CONCOCT was used to bin predicted eukaryotic sequences, and gene predictions were reperformed on the Choanephora bin with the MAKER2 pipeline using GeneMark-ES and AUGUSTUS for gene prediction.

1.3.9 taxator-tk comparison

The microbial-full_20150430 database was obtained from the taxator-tk (Dröge et al. 2015) website and was used for mapping. Mapping of test genomes against the reference database was performed using BLASTN with default alignment parameters and output format described in the taxator-tk manual. In a second round of testing, scaffolds belonging to test genomes were removed from the test set to simulate genomes from novel organisms. Taxonomic assignment and binning were performed as described in the taxator-tk manual without filtering alignments.

1.3.10 Phylogenetic analyses

To determine ANI between genomes, dRep was used (Olm et al. 2017). To estimate taxonomic composition of Crystal Geyser samples, rpS3 proteins were searched against KEGG (Kanehisa et al. 2016) with USEARCH (UBLAST) (Edgar 2010), and the taxonomy of the top hit was used to assign identified rpS3s to taxonomic groups. Abundance of identified rpS3s was determined by calculating the average coverage depth of the scaffolds containing annotated ribosomal protein

S3 (rpS3) genes. Average coverage depth was calculated by dividing the number of reads mapped to the scaffold by the scaffold length. Abundances were normalized for comparison across samples by multiplying the average coverage depth by the sample read count times read length. Four hundred sixty-one protein sets were obtained from binned eukaryotic genomes, publicly available genomes from the Joint Genome Institute's IMG-M database (img.jgi.doe.gov; Chen et al. 2016), NCBI, the Candida Genome Database (http ://www.candidagenome.org/), and a previously developed data set (Hug et al. 2016). For each protein set, 16 ribosomal proteins (L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24, S3, S8, S10, S17, and S19) were identified by BLASTing a reference set of 16 ribosomal proteins obtained from a variety of protistan organisms against the protein sets. BLAST hits were filtered to a minimum e-value of $1.0 \times 10{-}5$ and minimum target coverage of 25%. The 16 ribosomal protein data sets were aligned with MUSCLE (v. 3.8.31) (Edgar 2004) and trimmed by removing columns containing 90% or greater gaps. The alignments were then concatenated. A maximum likelihood tree was constructed using RAxML (v. 8.2.10) (Stamatakis 2014), on the CIPRES web server (Miller et al. 2010), with the LG plus gamma model of evolution (PROTGAMMALG) and with the number of bootstraps automatically determined with the MRE-based bootstopping criterion.

1.4 Results

1.4.1 Crystal Geyser community structure

The deep subsurface microbial community at Crystal Geyser, Utah has been well characterized as being dominated by chemolithoautotrophic bacteria and archaea, including many organisms from candidate phyla (CP) (Probst et al. 2014, 2016; Emerson et al. 2015). It is our current understanding that a wide diversity of novel bacteria and archaea are brought to the surface by geyser eruptions (Probst et al. 2018). Such deep sedimentary environments are unlikely to have high organic carbon compound availability. Thus, we hypothesized that organic carbon addition to this system would profoundly shift the community composition by selecting against the novel geyser microorganisms and enriching for better known heterotrophs. To test this prediction, we analyzed a sample of wood that was added to the shallow geyser and had decayed in the groundwater conduit (hereafter referred to as CG_WC). This sample and a wood-free sample (CG_bulk) that was collected the day before CG_WC were subjected to metagenomic analysis. We identified 124 and 316 distinct strains in the CG_WC and CG_bulk samples, respectively. The CG_WC sample contained abundant eukaryotic sequences (Fig. 1.1A) that were not present in the surrounding geyser water (Fig. 1.1B). Twelve strains were present in both samples (Fig. 1.1C), including the archaeon Candidatus "*Altiarchaeum hamiconexum*" (Probst et al. 2014), which dominated the CG_bulk sample. A phylum-level comparison of the microbial communities is presented in Figure 1.1D. The presence of decaying wood strongly enriched for Actinobacteria and Proteobacteria, as well as eukaryotes such as Ascomycota, Basidiomycota, and an organism classified as part of the Arthropoda. A low abundance alga from the class Bacillariophyta was detected in both samples.

As predicted, the CG_WC sample contains very few CP bacteria and archaea, with the notable exception of three members of Saccharibacteria (TM7). Two Saccharibacteria genomes were >90% complete, and one 1.01 Mbp genome was circularized and curated to completion. To evaluate for the accuracy of the complete genome, we ruled out the presence of repeat sequences

that could have confounded the assembly and carefully checked the consistency of paired reads mapped across the entire genome (Supplemental Data 1.1). The cumulative GC skew was used to identify the origin and terminus of replication (Brown et al. 2016). Although the skew has generally the expected form (consistent with genome accuracy), the origin defined based on GC skew was offset from the *dnaA* gene by ~46 kbp (Supplemental Fig. S1.1A). Short repeat sequences often associated with the origin were absent both from the predicted origin and the region encoding dnaA, although they were identified close to the origin for another candidate phyla radiation bacterium (Anantharaman et al. 2016). We identified the origin region for a previously reported complete Saccharibacteria RAAC3_TM7 genome using cumulative GC skew and showed that repeats were not present in this genome either and that the predicted origin is 7.6 kb from the *dnaA* gene (Kantor et al. 2013).

1.4.2 EukRep tested on reference data sets

Typically, only prokaryotic gene prediction is performed on metagenomic samples, as these are the only algorithms specifically designed for this application (e.g., MetaProdigal) (Hyatt et al. 2012). For samples containing both prokaryotic and eukaryotic DNA, such as CG_WC, obtaining high-quality gene predictions for eukaryotes is complicated by the fact that distinct gene prediction tools are used for prokaryotic vs. eukaryotic sequences due to differences in gene structure. Specifically, eukaryote genomes have more complex promoter regions, regulatory signals, and genes spliced into introns and exons, variable between species. For this reason, it is not surprising that we found that prokaryotic gene predictors underperform when used on eukaryotic sequences. This can impact binning by affecting taxonomic profiling of scaffolds and bin quality metrics such as the presence or absence of single-copy genes (Supplemental Fig. S1.2). To address this issue and obtain high-quality eukaryotic gene predictions from metagenomes, we present EukRep, a classifier that utilizes k-mer composition of assembled sequences to identify eukaryotic genome fragments prior to gene prediction (Fig. 1.2A). When previously used to taxonomically classify metagenomic sequences, machine learning algorithms have shown promise, but their success was limited when samples contained many different species (Vervier et al. 2016). We hypothesized that a supervised classification method could be applied to accurately classify sequences at the domain level for gene prediction purposes, avoiding complications from having a large number of taxonomic categories.

The EukRep model was trained using a diverse reference set of bacterial, archaeal, opisthokonta, and protist genomes (3.40 Gbps of sequence) (Supplemental Table S1.1). The k-mer frequencies were calculated for each 5-kb interval, resulting in 581,376 individual instances that were used to train a linear-SVM (scikit-learn) (Pedregosa et al. 2011). We found that 5-mer frequencies represented the best compromise between speed and accuracy for classifying eukaryotic scaffolds and that sequences can be classified with high accuracy at lengths of 3 kb or greater (Fig. 1.2B; Supplemental Fig. S1.3). A validation set of 486 independent genomes (Supplemental Table S1.2) was assembled to test the prediction power of EukRep. An important goal of EukRep is to be able to classify novel as well as known eukaryotic sequences and to avoid overfitting for existing eukaryotic sequences. Thus, the training and validation sets were chosen so as to taxonomically overlap at a maximum of genus level. Using the described validation set to test EukRep, we found that the classifier was able to accurately predict the domain of 97.5% of total tested eukaryotic sequence length and 98.0% of prokaryotic sequence length.

An important note is that EukRep is designed so as to miss as little eukaryotic sequence as possible. To ensure this, the program classifies every sequence in a sample, even sequences whose composition signals will be weak because the sequences are relatively short. Further, given the continuum between confident and less confident classification of eukaryote sequences, we chose settings that maximized classification outcomes (recall). The 2% of incorrect classifications of prokaryote as eukaryote sequences represent false positives that can be removed using standard binning methods (especially those that include phylogenetic signal).

We examined classifier accuracy on a per-genome basis to test whether the classifier performance varied for organisms of widely different types (Fig. 1.2C). This metric differs from that reported above because it refers to the accuracy of classifying individual artificially fragmented genomes rather than overall accuracy on all scaffolds tested from every genome. Ninety-four percent of tested eukaryotic genomes were classified with >90% accuracy, whereas 88% of tested prokaryotic genomes were classified with >90% accuracy. In a small number of prokaryotic genomes, more than half of the contigs were misclassified as eukaryotic. Notably, all of these were small genomes of organisms inferred to be parasites or symbionts. However, almost all of the sequences composing the eukaryotic genomes tested were correctly classified, indicating this method can successfully identify scaffolds whose analysis would benefit from a eukaryotic gene prediction algorithm.

In a complex metagenomic sample, obtaining sequences from novel lineages is a relatively common occurrence, and EukRep's ability to classify novel eukaryotic sequences is critical. We tested the ability of EukRep to do this by having it classify both eukaryotes (n = 18) and prokaryotes (n = 46) from phyla not represented in EukRep's training set (Supplemental Fig. S1.3). Although the genomes were fragmented into 3-kb pieces, EukRep maintained an overall accuracy of 90%. When tested on sequences fragmented to 20 kb, accuracy improved to 98%. Thus, we conclude that EukRep can be relied upon to correctly classify the majority of genomes from potentially entirely new phyla, even when fragmented.

Other taxonomic binning algorithms such as taxator-tk (Dröge et al. 2015) rely upon alignment to reference databases to make taxonomic classifications. Although these algorithms are typically designed for classifying reads at the lowest taxonomic level possible (e.g., species), they can potentially classify scaffolds at the domain level and perform the same function as EukRep. In order to test whether EukRep represents a significant improvement in this specific area, we compared EukRep to taxator-tk by classifying genomes from phyla unrepresented in EukRep's training set at the domain level. taxator-tk was selected for comparison because it includes eukaryotes in its prebuilt reference data set. taxator-tk was run twice. In the first test, many of the fragments to be classified were present as genomes in the reference data set (11/18), and it classified 47% of the total eukaryotic sequence tested as eukaryotic at 3 kb and 76% at 20 kb (Supplemental Fig. S1.3). In the second test, where the test genomes were removed from the reference set at the genus level so that the fragments represented genomes from novel genus level organisms at a minimum, the tool classified 24% at 3 kb and 44% at 20 kb of total eukaryotic sequence as eukaryotic (Supplemental Fig. S1.3). Due to the fact that EukRep does not rely upon alignment-based methods, it also does not require a reference database and can process metagenomes quickly, at a rate of up to two Gbp an hour on a single core. Thus, we

conclude that EukRep represents an improvement over this approach for the purpose of identifying scaffolds for eukaryotic gene prediction.

1.4.3 Testing eukaryotic gene predictions on reference genomes

Eukaryotic gene prediction algorithms rely on a combination of transcriptomic evidence or protein similarity (AUGUSTUS [Stanke et al. 2006]; SNAP [Korf 2004]) and sequence signatures (GeneMark-ES [Ter-Hovhannisyan et al. 2008]) to make predictions. Given the frequent lack of sequenced close relatives to organisms identified in metagenomes and the lack of transcript data in many metagenomic studies, we tested how well eukaryotic gene predictors function in a diversity of eukaryotic genomes without transcriptomic evidence or homology evidence from close relatives. We applied the MAKER2 pipeline (Holt and Yandell 2011) with GeneMark-ES in self-training mode along with AUGUSTUS trained using BUSCO (Simão et al. 2015) to nine diverse eukaryotic genomes obtained from JGI's portal (Grigoriev et al. 2011) and NCBI's genome database (Fig. 1.3A; NCBI Resource Coordinates 2017). The proteomes of *Chlamydomonas reinhardtii* (Merchant et al. 2007), *Neurospora crassa* (Galagan et al. 2003), and *Reticulomyxa filosa* (Glöckner et al. 2014) were also used as homology evidence. In each case, MAKER2-derived gene predictions were compared to reference gene predictions that incorporate transcriptomic evidence. The majority of the gene predictions identified without transcriptomic evidence were supported by reference gene predictions (78%–98%), and the majority of reference gene predictions overlapped a MAKER2-derived gene prediction (75%–98%). Estimated completeness of the predicted gene sets was measured by using BUSCO (Simão et al. 2015) to search for 303 eukaryotic single-copy orthologous genes within the predicted gene sets. The number of single-copy, duplicated, fragmented, and missing genes showed minimal differences with and without transcriptomic evidence (Fig. 1.3A). These results show the pipeline we assembled for eukaryotic gene prediction, even without transcriptomic evidence, is capable of detecting near complete gene sets similar to those from reference genomes, with the exception of untranslated regions and alternative splicing patterns.

To ensure that our proposed methodology can result in improved eukaryotic gene predictions in the context of a complex metagenomic sample, we spiked the genome of *Choanephora cucurbitarum* (Min et al. 2017) into a complex, 15-Gbp, soil shotgun metagenomic sample (Fig. 1.3B). The genome of *C. cucurbitarum* was used because it is a fragmented draft (N50 = 24,238 bp) with scaffold lengths similar to what is often encountered in a metagenome and because it has gene models with many introns that would particularly benefit from eukaryotic gene prediction. EukRep was run on this mock data set and recovered 40.6 Mbp of sequence classified as eukaryotic. Of this, 26.6 Mbp were the Choanephora genome (91.6% of the entire genome, 99.6% of the genome longer than the 3-kb minimum sequence length cutoff). Next, 93.2% of the identified genome was placed into a single bin. Training and running eukaryotic gene predictors on this bin substantially improved gene predictions, increasing estimated completeness via single-copy genes from 36% to 97% (Fig. 1.3C). The gene models were substantially more similar to reference gene models in terms of total gene count and gene length than those predicted using MetaProdigal (Supplemental Fig. S1.4).

1.4.4 Analysis of newly reconstructed eukaryotic genomes

After benchmarking EukRep on reference data sets, the algorithm was applied to the CG_WC sample, and 214.8 Mbps of scaffold sequence was classified as eukaryotic. Because eukaryotic gene predictors are designed to be trained and run on a single genome at a time, CONCOCT (Alneberg et al. 2014), an automated binning algorithm, was applied to the identified eukaryotic scaffolds to generate two preliminary eukaryote genomes. In this way, GeneMark-ES and AUGUSTUS gene prediction could be performed, as described above, on each bin individually as if running on a single genome.

The availability of relatively confident gene predictions for eukaryotic contigs enabled re-evaluation of genome completeness based on the presence or absence of 303 eukaryotic single-copy genes as identified by BUSCO (Table 1.1; Fig. 1.4). An obvious finding was that one of the CONCOCT bins was a megabin. Using information about single-copy gene inventories, along with tetranucleotide frequencies, coverage, and GC content, we assigned the eukaryotic scaffolds into four genome bins. BLASTing gene predictions against UniProt identified three of the bins as likely fungi and a fourth as a likely metazoan. Gene prediction was redone on the new fungal bins with GeneMark-ES in self-training mode and AUGUSTUS trained with BUSCO. The bins ranged in size from 24.5 Mbps to 99.0 Mbps and encoded between 8947 and 18,440 genes. BUSCO single-copy orthologous gene analysis showed all four bins were relatively complete individual genomes based on gene content, with the lowest containing 243/303 (80%) and the highest containing 288/303 (95%) single-copy orthologous genes (Table 1.1; Fig. 1.4). Some genes expected to be in single copy were duplicated, as is often found with BUSCO analysis of complete genomes. The assembly quality of one bin, WC_Fungi_A, appeared to be quite high, with 50% of its sequences contained in scaffolds longer than 599 kb. We reduced potential contamination of eukaryotic bins with prokaryotic sequence by BLASTing predicted proteins against UniProt and removing scaffolds with the majority of best hits belonging to prokaryotic genes.

A phylogenetic tree constructed from a set of 16 predicted, aligned, and concatenated ribosomal proteins (Hug et al. 2016) placed three of the bins within the fungal class Eurotiomycetes (Fig. 1.5). Each of these three bins ranged in size from 24.6 to 39.2 Mbps and in gene count from 8963 genes to 15,756 genes, within the range observed in previously sequenced Ascomycete fungi. The closest sequenced relative to all three bins was *Phaeomoniella chlamydospora*, a fungal plant pathogen known for causing Esca disease complex in grapevines (Morales-Cruz et al. 2015). The fourth bin, 99.7 Mbps in length and estimated to be 92% complete, was placed within the Arthropoda (Fig. 1.5). Its closest, although distant, sequenced relative is *Orchesella cincta* (FaddeevaVakhrusheva et al. 2016). *Orchesella cincta* is a member of the hexapod subclass Collembola (springtails), a diverse group basal to insects known primarily to be detritivorous inhabitants of soil. Although ribosomal protein S3 (rpS3) sequences belonging to Dictyosteliida, Heterolobosea, and Basidiomycota were detected, there were no genomes reconstructed for these organisms, likely due to low abundance or genome fragmentation.

1.4.5 Analysis of newly reconstructed eukaryotic genomes

To test whether the presence of organic carbon within the CG_WC sample would enrich for heterotrophic metabolic pathways (and against members of chemolithoautotrophic communities typically associated with the Crystal Geyser community), we searched the CG_WC and CG_bulk samples using HMMs for CAZymes grouped by substrate (Cantarel et al. 2009), lipase HMMs from the Lipase Engineering Database (Fischer and Pleiss 2003), and a protease BLAST database from MEROPS (Rawlings et al. 2016). Predicted proteases and lipases were filtered to specifically identify putative excreted proteases and lipases by searching for proteins with secretion signals identified with SignalP (Petersen et al. 2011) and one or less transmembrane domains with TMHMM (Krogh et al. 2001).

Pathways previously described as dominant within the Crystal Geyser such as the Wood Ljungdahl carbon fixation pathway and Ni-Fe hydrogenases were depleted in CG_WC as compared to CG_bulk. Instead, genes encoding CAZymes targeting cellulose, hemicellulose, pectin, starch, and other polysaccharides were enriched in CG_WC, indicating an increased capacity for degradation of complex carbohydrates (Fig. 1.6). A strong enrichment for excreted lipases and proteases was also detected, further indicative of an increase in the amount of heterotrophic metabolisms (Fig. 1.6). CG_WC also had a strong enrichment for methanol oxidation.

The four binned eukaryotic genomes contributed substantially to the putative heterotrophic categories (Supplemental Table S1.3). Fungi are known to exhibit different CAZyme profiles based upon their lifestyle (Ohm et al. 2012; Kim et al. 2016). An analysis of the CAZyme profiles of the three fungal bins focused on plant cell wall-targeting CAZymes supports the role of these fungi as possible plant pathogens or saprotrophs (Supplemental Table S4; Floudas et al. 2012; Ohm et al. 2012; Kim et al. 2016). A profile of CAZymes found within the Arthropoda bin revealed a large number of chitin-targeting CAZymes (Supplemental Table S1.3).

1.4.6 Testing EukRep in recovery of eukaryote genomes from other ecosystems

To test the broader application of EukRep, we applied the method to infant fecal samples and thiocyanate reactor samples in which eukaryotes had previously been identified (Sharon et al. 2013; Kantor et al. 2015, 2017; Raveh-Sadka et al. 2015, 2016). By using EukRep, we were able to quickly and systematically scan 226 samples for the presence of eukaryotic sequences. Six relatively complete fungal genomes were recovered from fecal samples from three infants (Fig. 1.4). Three are *Candida albicans* and were reconstructed from two different infants. The two genomes from the same infant are indistinguishable and very closely related to that from the third infant. All three are closely related to but distinguishable from the *C. albicans* reference strain WO-1 (Fig. 1.4; Supplemental Fig. S1.5A). The other three fungal genomes are strains of *Candida parapsilosis* that all occurred in a single infant. These are essentially indistinguishable from each other and from the *C. parapsilosis* strain CDC317 reference genome, with which they share >99.7% average nucleotide identity (ANI) (Fig. 1.4; Supplemental Fig. S1.5A,B; Sharon et al. 2013; Raveh-Sadka et al. 2015, 2016). *C. albicans* and *C. parapsilosis* are both clinically relevant human pathogens (Trofa et al. 2008; Kim and Sudbery 2011).

Within thiocyanate reactor samples, genomes of a rotifer, Rhizaria, and a relative of the slime mold *Fonticula alba* had previously been identified (Kantor et al. 2015, 2017). With EukRep, we

were able to rapidly identify these eukaryotic genomes and evaluate their completeness. Genome completeness analysis benefited from improved gene predictions for single-copy orthologous genes and showed that the identified genomes ranged in completeness Figure 1.4. Overview of binned eukaryotic genomes. Genomes that share greater than 99% average nucleotide identity (ANI) are indicated by black bars. ANI comparisons are shown in more detail in Supplemental Figure S1.3. Genic regions refer to sequence located within predicted gene models whereas intergenic refers to all other sequence. Genes containing a PFAM domain were identified with PfamScan (Mistry et al. 2007). Genome completeness is measured as the percent of 303 eukaryotic single-copy orthologous genes found within a genome in a particular form with BUSCO. West et al. 574 Genome Research www.genome.org Downloaded from genome.cshlp.org on March 19, 2020 - Published by Cold Spring Harbor Laboratory Press from 69%–91%. (Fig. 1.4). As previously reported (Kantor et al. 2017), the rotifer was present in seven different samples (Rotifer_A-G) (Fig. 1.4), consistent with its persistence in the thiocyanate reactor community. All seven bins shared greater than 99% ANI (Supplemental Fig. S1.5B) indicating they are likely the same species.

1.5 Discussion

Using a newly acquired and two previously reported whole-community metagenomic data sets, we demonstrated that it is possible to rapidly recover high-quality eukaryotic genomes from metagenomes for phylogenetic and metabolic analyses. The key step implemented in this study was the presorting of eukaryotic genome fragments prior to gene prediction. By training and using eukaryotic gene predictors, we achieved much higher quality eukaryotic gene predictions than those obtained using a prokaryotic gene prediction algorithm on the entire data set (i.e., without separation based on phylogeny). This was critical for draft genome recovery and evaluation of genome completeness.

Classification of assembled genome fragments at the domain level was surprisingly accurate, with 98.0% (Fig. 1.2C) of eukaryotic sequences being correctly identified as eukaryotic, despite no close relative in the training set in many cases (Supplemental Table S1.2). The high accuracy of separation suggests some underlying pattern of k-mer frequencies that is different in eukaryotes compared to prokaryotes. In part, the signature may arise from different codon use patterns associated with the different genetic codes for bacteria and eukaryotes.

We anticipate that reexamination of environmental metagenomic data sets using the same approach as implemented here will yield high quality genomes for previously unknown eukaryotes. An important benefit from this and future sequencing efforts will be an expanded knowledge of the diversity, distribution, and functions of microbial eukaryotes, which are widely acknowledged as understudied (Pawlowski et al. 2012). Increasing the diversity of sequenced eukaryotic genomes would benefit evolutionary studies. Current eukaryotic multigene trees form a solid backbone of the eukaryotic tree of life (Parfrey et al. 2010) but suffer from sparse eukaryotic taxon sampling. Single-gene trees, which are possible to construct from gene surveys, lack the resolution of multigene trees (Rokas and Carroll 2005). Comprehensive sequencing of full genomes would help diminish the sparse taxon sampling problem in multigene trees and improve eukaryotic evolutionary reconstructions, with implications for understanding of

eukaryotic protein function. For example, Ovchinnikov et al. (2017) demonstrated that it is possible to accurately predict protein structure by utilizing residue-residue contacts inferred from evolutionary data, but such analyses require large numbers of aligned sequences. More diverse eukaryotic sequences could expand the utility of this method for eukaryotic protein family analyses. Furthermore, a broader diversity of eukaryotic genomes would provide new insights regarding gene transfer patterns and whole-genome evolution.

EukRep, applied in the context of metagenomics, may prove useful for genome sequencing projects where isolation of the organism of interest may be difficult or not technically feasible. For example, it could be applied to study populations of bacteria Figure 1.5. Phylogenetic placement of binned eukaryotic genomes with maximum likelihood analysis of 16 concatenated ribosomal protein alignments. Genomes from Crystal Geyser, infant-derived fecal samples, and thiocyanate reactor samples are identified with blue, red, and purple circles, respectively. Branches with greater than 50% bootstrap support are labeled with their bootstrap support. Reference ribosomal proteins were obtained from Hug et al. (2016), JGI (Grigoriev et al. 2011), and NCBI (NCBI Resource Coordinators 2017). Metagenomic reconstruction of eukaryotic genomes Genome Research 575 www.genome.org Downloaded from genome.cshlp.org on March 19, 2020 - Published by Cold Spring Harbor Laboratory Press within the hyphae of arbuscular mycorrhizal fungi (Hoffman and Arnold 2010).

Eukaryotic cells frequently contain multiple sets of chromosomes (diploid or polyploid). These are often very similar but not identical and can result in the genome assembly alternating between collapsing and splitting contigs representing homologous genomic regions (Margarido et al. 2015). If reads are only allowed to map to one location when determining genome coverage, this could lead to variation of coverage values across different portions of a genome. As differential coverage of contigs is a parameter commonly used to help bin genomes, ploidy can complicate genome recovery. Another potential problem could relate to contamination of eukaryotic genome bins with some bacterial fragments. This will occur to some extent, given that some bacterial and archaeal contigs were wrongly classified as eukaryotic. Phylogenetic profiling of the predicted genes can be used to screen out most prokaryotic sequences.

During development, we noted that the frequency of correct identification of bacterial genomes was improved by increasing the number and diversity of eukaryote sequences used in classifier training. Further improvements are anticipated as the variety of reference sequences increases. However, there may be biological reasons underpinning incorrect profiles. The small number of cases where EukRep profiled bacteria as eukaryotes or vice versa may be interesting targets for further analysis. Notably, almost all are inferred or known symbionts or parasites, raising the question of whether their sequence composition has evolved to mirror that of their hosts.

We demonstrated the value of EukRep-enabled analyses through study of an ecosystem that had been perturbed by addition of a carbon source. The results clearly show a large shift in the community composition and selection for fungi. Of the binned genomes, the fungi have by far the most cellulose-, hemicellulose-, and pectin-degrading enzymes, consistent with their enrichment in response to high organic carbon availability from degrading wood. We also genomically characterized what appears to be a macroscopic hexapod that is related to springtails (Collembola), organisms known to feed on fungi (Chen et al. 1996). Given that the hexapod

genome has a large number of chitin-degrading enzymes (Supplemental Table S1.3), we speculate that it may be part of the community supported by the fungi in the decaying wood. However, it is also possible that it was associated with the wood prior to its addition to the geyser conduit. Interestingly, the eukaryote-based community contains very few members of the candidate phyla radiation (CPR) and an archaeal radiation known as DPANN and other CP bacteria. These novel organisms are mostly predicted to be anaerobes and are highly abundant in groundwater samples that were likely sourced from deep aquifers under the Colorado Plateau (Probst et al. 2018). The results of the current study indicate that CPR and DPANN in the Crystal Geyser system are adapted to an environment relatively low in carbon availability, a finding that may guide future laboratory enrichment studies that target these organisms.

1.6 Conclusions

Overall, the results reported here demonstrate that comprehensive, cultivation-independent genomic studies of ecosystems containing a wide variety of organism types, including eukaryotes, are now possible. Examples of future applications include analysis of the distribution and metabolic capacities and potential pathogenicity of fungi in the human microbiome, tracking of eukaryotes (including multicellular eukaryotes) in reactors used in biotechnologies, profiling of the built environment, and natural ecosystem research.

1.7 Figures

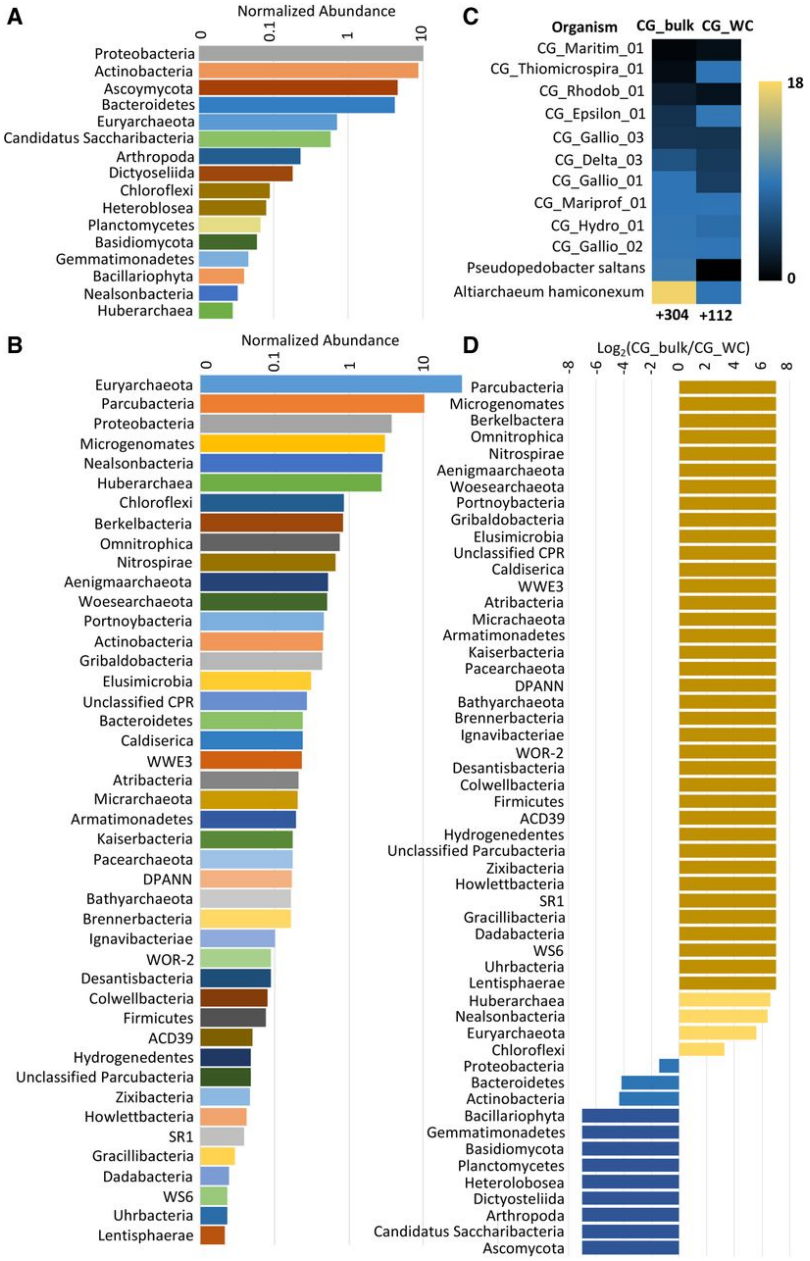

Figure 1.1 Comparison of CG_WC and CG_bulk community composition. The relative abundances of taxonomic groups in CG_WC (A) and CG_bulk (B) are depicted. Abundance was determined as the average coverage depth of the scaffolds containing annotated ribosomal protein S3 (rpS3) genes. Abundances were normalized for comparison across samples by multiplying the average coverage depth by the sample read count and read length. (C) Normalized coverage of rpS3 containing scaffolds of strains common to both samples. The number of additional strains detected in each sample is listed below the respective sample heat map. (D) Log2 ratio of normalized coverage of taxonomic groups from A and B. Taxonomic groups identified in only one sample are indicated by the darker yellow and blue bars.

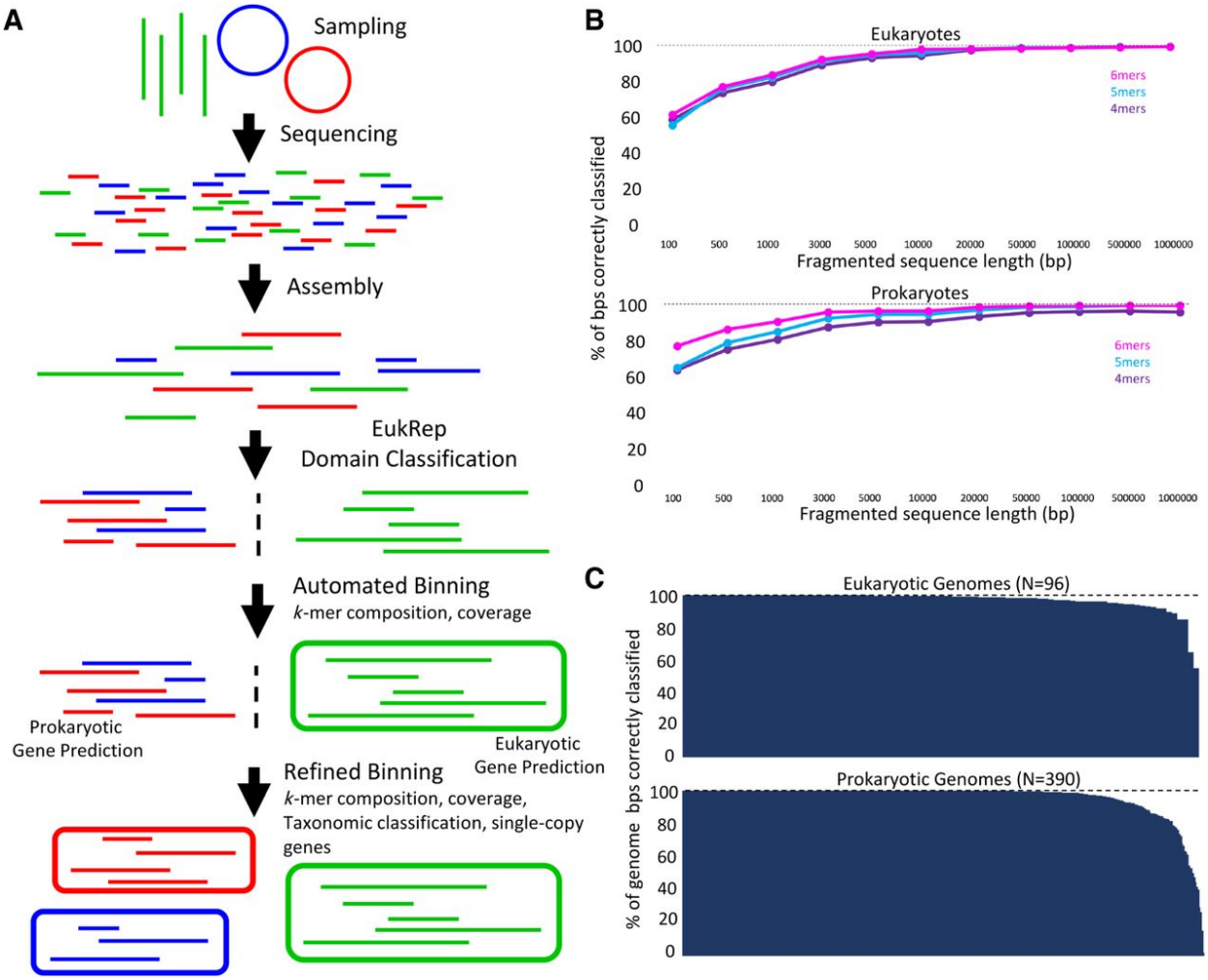

Figure 1.2 Identification of scaffolds for eukaryotic gene prediction with EukRep. (A) Schematic of the analysis pipeline used to identify and bin both eukaryotic and prokaryotic genomes within this paper. (B) A subset of genomes from Supplemental Table S2 was used to compare prediction accuracy of linear-SVM models trained on k-mer frequencies of k-mers ranging in length from 4 to 6 bp. For each sequence size category, sequences longer than the specified length were fragmented to the specified length and sequences shorter were excluded. (C) Accuracy of EukRep domain prediction on a per-genome level for both eukaryotes and prokaryotes. Percent of the genome correctly classified is defined as the percent of base pairs within a given genome predicted to belong to the genome's known domain. Each bar represents the percent of a single genome that was classified correctly. Genomes used for training and testing of EukRep along with their prediction results are listed in Supplemental Tables S1 and S2.

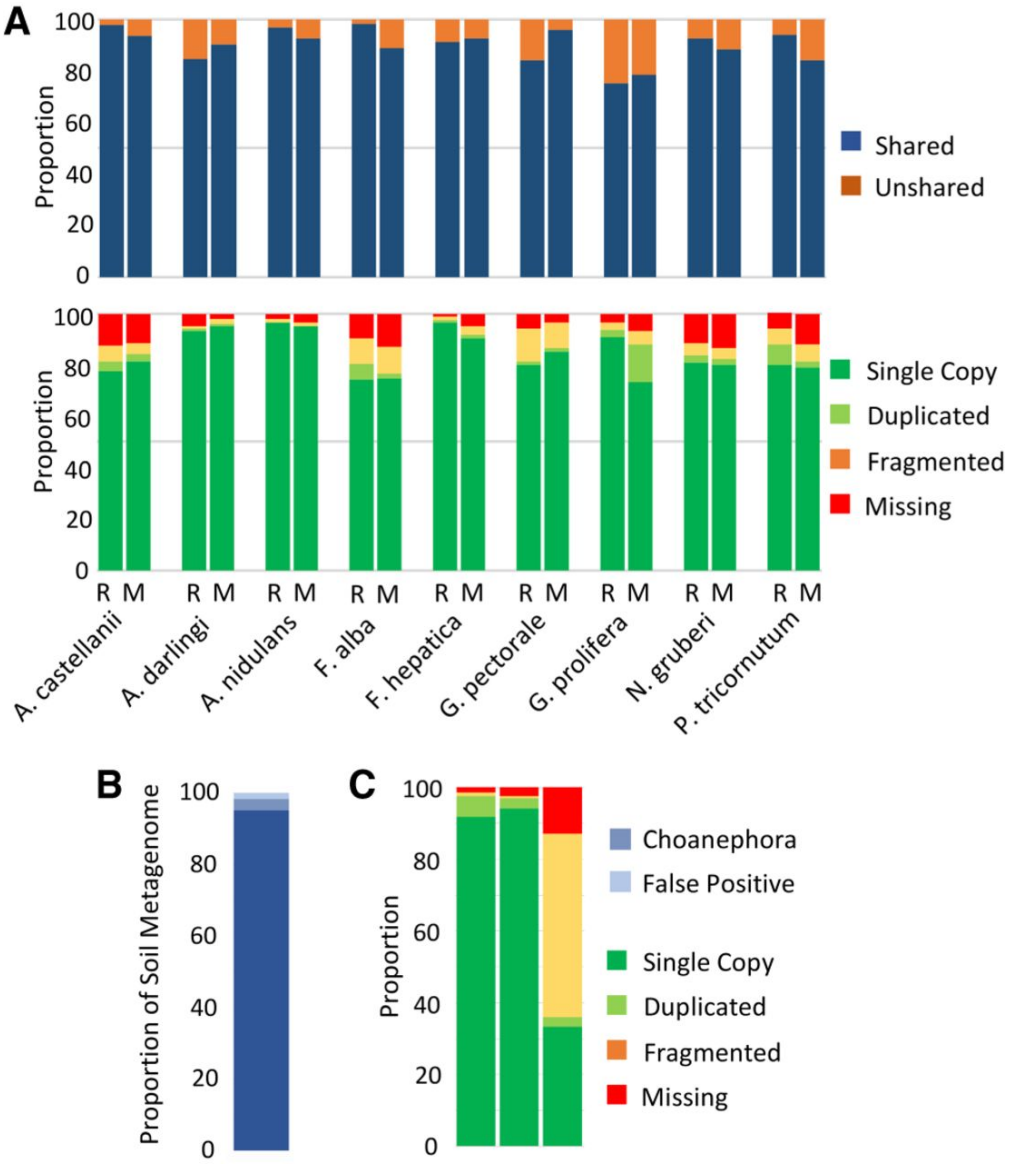

Figure 1.3. Eukaryotic gene prediction on metagenomic scaffolds. (A) Gene predictions for nine diverse eukaryotic organisms including fungi, a Metazoa, a Stramenopile, an Archaeplastida, and a Rhizaria. Columns labeled "R" refer to reference gene sets, whereas M columns refer to gene sets predicted without transcript or close homology evidence. The top panel displays the proportion of total genes either overlapping (shared) or not overlapping (unshared) a gene model from the other respective gene set for a given genome. The bottom panel is an analysis of presence or absence of single-copy genes in each gene set as determined by BUSCO using the eukaryota_odb9 lineage set. (B) Proportion of a soil metagenome spiked with the genome of *Choanephora cucurbitarum* predicted to be either noneukaryotic, eukaryotic and belonging to the Choanephora, or predicted to be eukaryotic but has homology to prokaryotic sequences. (C) BUSCO analysis of the binned *Choanephora cucurbitarum* genome with protein sets from (left to right) the reference protein set, trained MAKER2 output, and whole metagenome MetaProdigal output.

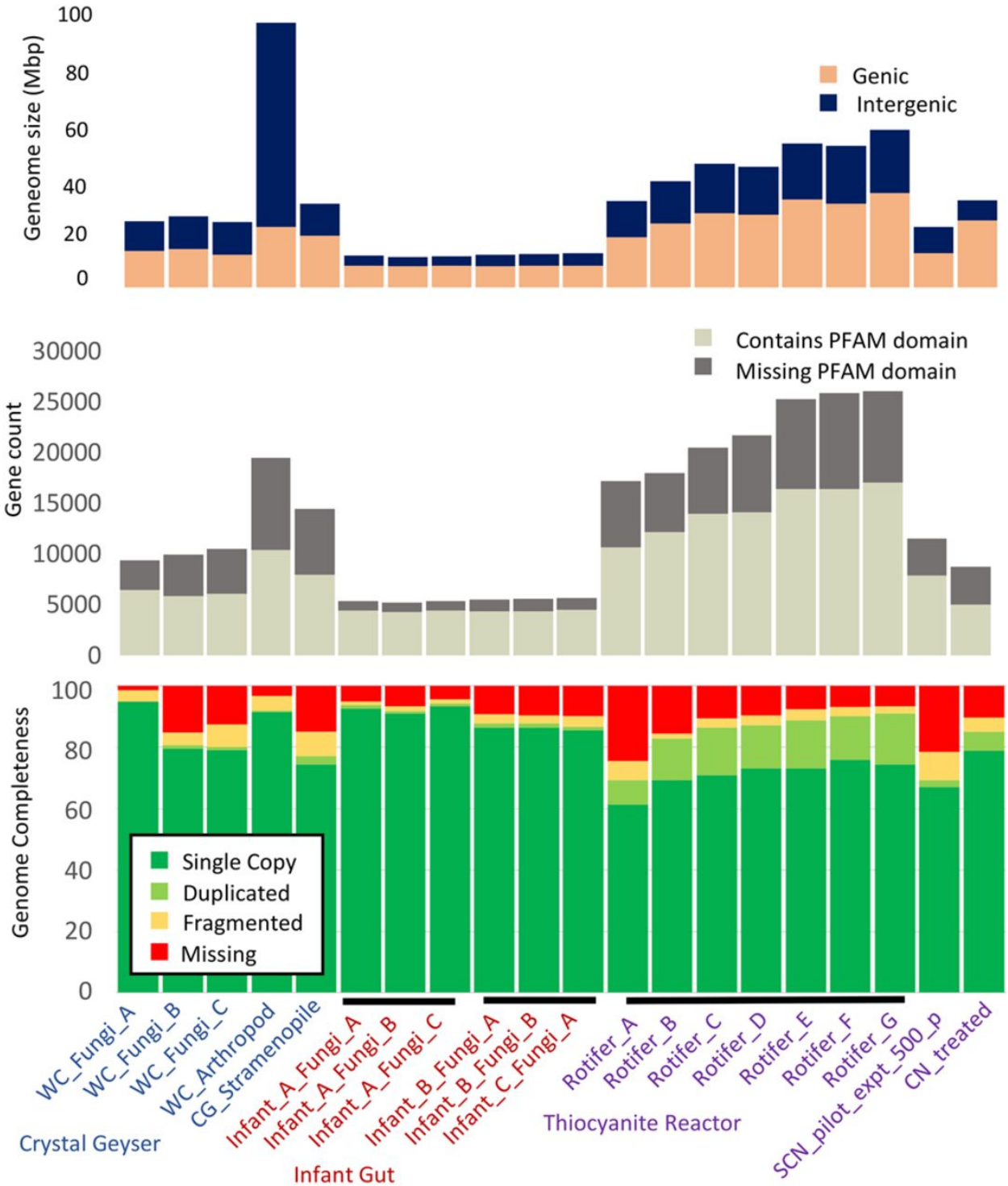

Figure 1.4. Overview of binned eukaryotic genomes. Genomes that share greater than 99% average nucleotide identity (ANI) are indicated by black bars. ANI comparisons are shown in more detail in Supplemental Figure S3. Genic regions refer to sequence located within predicted gene models whereas intergenic refers to all other sequence. Genes containing a PFAM domain were identified with PfamScan (Mistry et al. 2007). Genome completeness is measured as the percent of 303 eukaryotic single-copy orthologous genes found within a genome in a particular form with BUSCO.

Figure 1.5. Phylogenetic placement of binned eukaryotic genomes with maximum likelihood analysis of 16 concatenated ribosomal protein alignments. Genomes from Crystal Geyser, infant-derived fecal samples, and thiocyanate reactor samples are identified with blue, red, and purple circles, respectively. Branches with greater than 50% bootstrap support are labeled with their bootstrap support. Reference ribosomal proteins were obtained from Hug et al. (2016), JGI (Grigoriev et al. 2011), and NCBI (NCBI Resource Coordinators 2017).

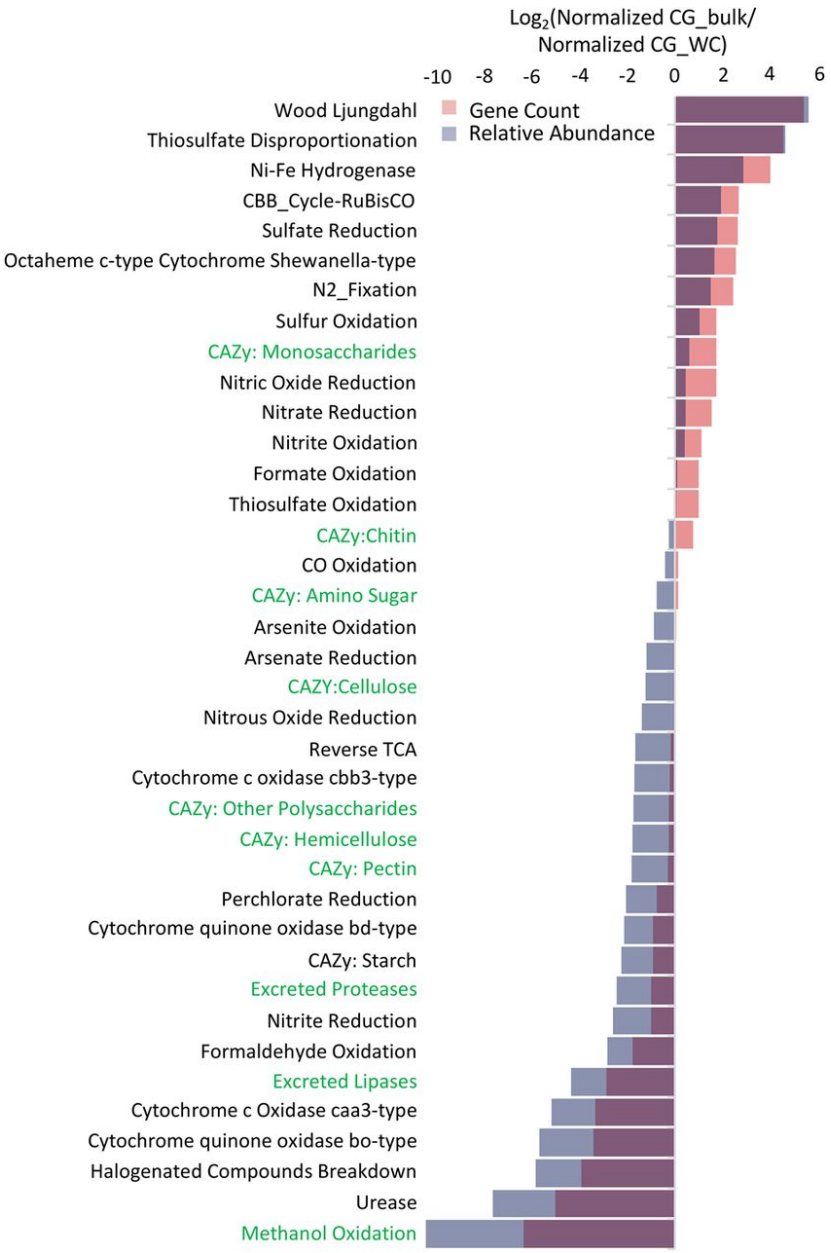

Figure 1.6. Comparison of CG_WC and CG_bulk metabolic capacity. Log2 ratio of all annotated genes found within the CG_bulk sample against annotated genes found in the CG_WC sample. Annotated genes were grouped into categories based upon scores with a custom set of metabolic pathway marker HMMs (Anantharaman et al. 2016), CAZyme HMMs (Cantarel et al. 2009), and protease and lipase HMMs from MEROPs and the Lipase Engineering Database, respectively. Putative proteases and lipases were also filtered to only those containing a secretion signal and less than three transmembrane domains (see Methods). Gene count (red) is the ratio of total number of genes in each category for each sample normalized by the total number of genes found in the sample. Relative abundance (blue) is the ratio of average read coverage depth of the contig containing a given annotated gene in each category normalized by the sample read count multiplied by read length.

For supplemental figures, tables, and information for Chapter 1, see
https://doi.org/10.1101/gr.228429.117

2 Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms

Olm, Matthew R.*, West, Patrick T.*, Brooks, Brandon, Firek, Brian A., Baker, Robyn, Morowitz, Michael J., Banfield, Jillian F.

Published in *Microbiome* February 2019, doi: https://doi.org/10.1186/s40168-019-0638-1

2.1 Abstract

2.1.1 Background

Fungal infections are a significant cause of mortality and morbidity in hospitalized preterm infants, yet little is known about eukaryotic colonization of infants and of the neonatal intensive care unit as a possible source of colonizing strains. This is partly because microbiome studies often utilize bacterial 16S rRNA marker gene sequencing, a technique that is blind to eukaryotic organisms. Knowledge gaps exist regarding the phylogeny and microdiversity of eukaryotes that colonize hospitalized infants, as well as potential reservoirs of eukaryotes in the hospital room built environment.

2.1.2 Results

Genome-resolved analysis of 1174 time-series fecal metagenomes from 161 premature infants revealed fungal colonization of 10 infants. Relative abundance levels reached as high as 97% and were significantly higher in the first weeks of life ($p = 0.004$). When fungal colonization occurred, multiple species were present more often than expected by random chance ($p = 0.008$). Twenty-four metagenomic samples were analyzed from hospital rooms of six different infants. Compared to floor and surface samples, hospital sinks hosted diverse and highly variable communities containing genomically novel species, including from Diptera (fly) and Rhabditida (worm) for which genomes were assembled. With the exception of Diptera and two other organisms, zygosity of the newly assembled diploid eukaryote genomes was low. Interestingly, Malassezia and Candida species were present in both room and infant gut samples.

2.1.3 Conclusions

Increased levels of fungal co-colonization may reflect synergistic interactions or differences in infant susceptibility to fungal colonization. Discovery of eukaryotic organisms that have not been sequenced previously highlights the benefit of genome-resolved analyses, and low zygosity of assembled genomes could reflect inbreeding or strong selection imposed by room conditions.

2.2 Introduction

Eukaryotes are common members of the human microbiome (Schulz et al. 2009; Baley et al. 1986; Tamburini et al. 2016). The colonization density and diversity of eukaryotes are lower than their bacterial counterparts (Schulz et al. 2009; Ott et al. 2008; Parfrey et al. 2011), but they can have substantial health consequences. The yeast *Saccharomyces boulardii* can significantly reduce rates of antibiotic-associated diarrhea (Surawicz et al. 1989), protozoa limit Salmonella populations through predation (Wildschutte et al. 2004), and high abundances of Candida and Rhodotorula are associated with asthma development in neonates (Fujimura et al. 2016). Fungal disease is most prevalent in immunocompromised patients, including premature infants (Fridkin et al. 1996; Manzoni et al. 2015), although their incidence has declined in recent decades (Aliaga et al. 2014).

While infant fungal disease is an active area of study, little is known about asymptomatic colonization of premature infants by fungi or other eukaryotes. Studies have reported 0%, 26%, 50%, and 63% of premature infants being colonized by fungi (Baley et al. 1986; Stewart et al. 2012; Stewart et al. 2013; LaTuga et al. 2011), with variation in methodological sensitivity probably at the heart of these differences. Methods used to analyze the mycobiome, including culturing, DGGE, and ITS sequencing, identify the fungal fraction of the microbial community separate from the community at large. This has left basic knowledge gaps about the relative abundance of fungi in early life, an important point as fungi-infant interactions in early life are known to affect allergy development (Fujimura et al. 2016; Bush et al. 2001; Fujimura et al. 2010). In fact, recent review articles have referred to eukaryotes as a "Missing Link in Gut Microbiome Studies" (Laforest-Lapointe et al. 2018), and stated that "Studies addressing how the infant mycobiome develops and shapes the host immune system will be required for a more comprehensive understanding of the early-life microbiome." (Tamburini et al. 2016). Particular highlighted knowledge gaps relate to the ecological roles, growth dynamics, and source of eukaryotes in the human and hospital microbiomes (Laforest-Lapointe et al. 2018; Huffnagle et al. 2013).

The hospital is a known source for bacterial infant colonists (Brooks et al. 2017). The built environment has been implicated in fungal outbreaks (Sanchez et al. 1993; Vazquez et al. 1993; Pfaller et al. 1996; Mesquita-Rocha et al. 2013), yet the eukaryotic built environment microbiome remains understudied. This is because the vast majority of high-throughput studies of the hospital microbiome and the human gut microbiome use bacteria-specific 16S rRNA marker gene sequencing, and thus are blind to eukaryotes. Of five recent studies of the hospital microbiome, only one included primers to target the internal transcribed spacer (ITS) sequences to detect eukaryotes (Oberauner et al. 2013; Lax et al. 2017; Shin et al. 2015; Hewitt et al. 2013; Bokulich et al. 2013). It remains to be seen if eukaryotes in the room have the genetic potential to colonize infants, and if so, where in the room these eukaryotes are located.

An alternative approach to microbiome characterization involves shotgun metagenomics. In this method, all DNA from a sample is sequenced regardless of its organismal source or genetic context. In some studies, mapping of the sequencing reads to reference genomes has enabled identification of pathogens (Wilson et al. 2018). However, the reads can be assembled, and new methods aid in reconstructing eukaryotic genomes from these datasets (West et al. 2018),

enabling understanding of these organisms in the context of their entire communities, which also include bacteria, archaea, bacteriophage, viruses, and plasmids. Relative to amplicon sequencing, genome assembly has several distinct advantages for understanding communities that contain eukaryotes. First, genomes provide information about *in situ* ploidy (number of distinct chromosome sets per cell), heterozygosity (here used to refer to the fraction of alleles in a diploid genome that have two versus one abundant sequence types), and extent of population microdiversity (here used to refer to additional sequence types that constitute low-abundance alleles). Second, strain-tracking can be performed using high-resolution genomic comparisons. Last, newly assembled eukaryotic sequences expand the diversity of genomically defined eukaryotes in public databases, enabling comparative and evolutionary studies.

Here, we used genome-resolved metagenomics to study eukaryote-containing microbiomes of premature infants and their NICU environment. We evaluated the incidence of eukaryotes in room and infant samples and investigated the time period during which infant microbiomes contained eukaryotes. Genomes were assembled for 14 eukaryotic populations, and their ploidy, zygosity, and population microdiversity defined. The same species of eukaryotes were found in infant microbiome and the NICU environment, and a subset of other microbial eukaryotes in NICU rooms was classified as types that can cause nosocomial infections.

2.3 Results

2.3.1 Recovery of novel eukaryotic genomes from metagenomes

In this study, we analyzed 1174 fecal metagenomes and 24 metagenomes from the NICU environment, totaling 5.31 Tb of DNA sequence (Additional file 1: Table S2.1). Fecal samples were collected from 161 premature infants primarily during the first 30 days of life (DOL) (full range of DOL 5–121; median 18), with an average of 7 samples per infant. NICU samples were taken from six patient rooms within the hospital housing the infants (Magee-Womens Hospital of UPMC, Pittsburgh, PA, USA). Three NICU locations were sampled in each room: swabs from frequently touched surfaces, wipes from other surfaces, and swabs from sinks (Brooks et al. 2017). Eukaryotic genomes were assembled from all samples using a EukRep-based pipeline (West et al. 2018; see the "Methods" section for details). The bacterial component of some of the datasets was analyzed previously (see the "Methods" section).

Fourteen novel eukaryotic genomes were recovered in total, with a median estimated completeness of 91% (Table 2.1). Detailed genome assembly information is available in Additional file 2: Table S2. Genomes were assembled from organisms of a wide phylogenetic breadth, and four are the first genome sequences for their species (Fig. 2.1). Twelve of the genomes are classified as fungal and are described in more detail below. The two other genomes (both recovered from hospital sink samples) represent the first genomes of their phylogenetic families. Diptera S2_005_002R2 is within the phylogenetic clade of Diptera (true flies) and is equally related to *Drosophila melanogaster* (fruit fly) and *Lucila cuprina* (Australian sheep blowfly). Rhabditida S2_005_001R2 is within the family Rhabditida (nematode) and is related to both pathogenic and non-pathogenic roundworms. In both cases, BLAST searches of the rpS3 protein sequence against NCBI revealed no significant hits, and furthermore, comparing the mitochondrial cytochrome c oxidase subunit I gene and protein against the Barcode Of Life Database (BOLD) (Ratnasingham

et al. 2007) and NCBI revealed no hits with high identity. Thus, we are unable to tie our genomes to any morphologically described species.

2.3.2 Fungal contaminants in extraction controls

Four negative extraction controls were subjected to metagenomic sequencing to detect sequences resulting from reagent contamination. One of the four extraction controls harbored *Purpureocillium lilacinum* DNA, with > 50% of sample reads mapping to the genome and with a breadth of coverage (percentage of the genome covered by at least one read) of 87% (Additional file 3: Figure S2.1A). The average nucleotide identity (ANI) was calculated between *P. lilacinum* reads in the extraction control, *P. lilacinum* genomes assembled in the study, and all previously sequenced *P. lilacinum* genomes in NCBI (Additional file 3: Figure S2.1B). *P. lilacinum* reads from the extraction control were extremely similar to genomes assembled from the NICU and infant gut, and divergent from previously sequenced genomes (Additional file 3: Figure S2.1B). Thus, *P. lilacinum* genomes assembled from room and gut samples are probably due to reagent contamination and not actually present in the environment.

Reads from three of the four extraction controls mapped to *Malassezia restricta* S2_018_000R1, all at low abundance (< 3% of reads with a genome breadth of coverage of 1.3–14.2% using reads from the four samples) (Additional file 3: Figure S2.1C). It was not possible to calculate the ANI between the genomes in samples and controls due to the low sequencing coverage of *M. restricta* S2_018_000R1 in the extraction controls. Malassezia is a near-ubiquitous skin-associated fungus (Gaitanis et al. 2012). Based on the depth of coverage (2.37×), the genome had a very low breadth of coverage (88% expected vs. 13% actual) (Additional file 4: Figure S2.9), indicating that the genome sampled from the hospital surface is different to that of the Malassezia that contaminated the reagents. For this reason, the Malassezia in infant and room samples were not excluded from further analysis.

2.3.3 Fungal microbiome of the premature infant gut

Excluding *P. lilacinum*, fungi were detected in 10 of the 161 premature infants profiled in this study (6%) (Fig. 2.2a; Additional file 5: Table S2.3). The limit of detection for eukaryotic organisms was calculated as 0.05% of the total community (Additional file 6: Figure S2.2) (see the "Methods" section for details). Eukaryotes were detected significantly more often early in life, and significantly more often when antibiotics were recently administered (Fig. 2.2b). Antibiotics were given significantly more often early in life ($p = 5.3E{-}8$; Wilcoxon rank-sum test), making it difficult to determine which of these two variables is driving the association.

Fungal colonization was not significantly associated with gestational age, twin status, birth weight, mode of delivery, or other clinical metadata. (Additional file 7: Tables S2.4, Additional file 8: Table S2.5). Further, fungal colonization was not associated with bacterial community composition. *P. lilacinum*, presumed to be a metagenomic contaminant (Additional file 3: Figure S2.1), decreases in abundance as infants age (Additional file 9: Figure S2.8), probably because increased bacterial biomass in later collected samples overwhelms the contaminant DNA, as shown previously (Salter et al. 2014). Given this, we infer that the decrease in relative abundance of fungi present in the microbiomes of later-collected samples is due to bacterial growth.

All seven species detected colonizing the premature infants have been previously implicated as agents of nosocomial infection (Table 2.2), yet no infants colonized by eukaryotes in this study received antifungals or showed any symptoms consistent with acute fungal infection. However, asymptomatic colonization has been shown to be a risk factor for future fungemia (Huang et al. 1998). Seven different eukaryotic species were detected in at least one infant, with only *Candida albicans* and *Candida parapsilosis* colonizing more than one infant (Fig. 2.2a). Infant N2_070 was colonized by two fungi, and infant N5_275 was colonized by three. A permutation test was performed to determine if fungi were unevenly distributed among the infants of this study (i.e., if having one fungi predisposes colonization by another). The probability of observing 13 fungi colonize ≤ 10 unique individuals from a total population of 161 individuals was determined (Fig. 2.2c), with a resulting p value of 0.008. Thus, in this study, multiple fungi colonized the same infant more often than expected random chance.

2.3.4 Fungal microbiome of the neonatal intensive care unit

Eukaryotic organisms were detected in 18 of the 24 metagenomes of the NICU room environment (Fig. 2.3). Eukaryotic DNA made up an average of 1.23%, 1.22%, and 0.03% of the communities in highly-touched surfaces, sinks, and counters and floors, respectively. In order to compare the influence of room occupants and sampling location on the room mycobiome, we performed a multidimensional scaling (MDS) analysis (Fig. 2.3a). Communities were differentiated based on sampling location rather than infant room.

The mycobiome of the NICU surfaces is dominated by species of Malassezia (Fig. 2.3b). The eukaryotic organisms found in NICU sinks are distinct from, and more diverse than, those found on surfaces. Sink communities contained *Necteria haematococca*, *Candida parapsilosis*, Exophiala, and Verruconis, all of which were detected in multiple rooms and samples. Additionally, sinks in three separate NICU rooms contain DNA from Rhabditidia S2_005_000R1 (a novel nematode; see the previous section for details). Diptera S2_005_002R2 (fly) also makes up about 2% of the entire community for a single time-point in the sink in infant S2_005's room (Fig. 2.3b). No macroscopic organisms were noted during the sample collection process. It remains to be seen whether these organisms contribute to the dispersal of organisms throughout the NICU or affect the communities themselves.

Candida parapsilosis was detected in both the NICU and in a premature infant, as were organisms of the genus Malassezia. To contextualize the similarity between *C. parapsilosis* strains in both communities, genomes assembled from both the infant and room environments were compared to all available reference genomes and each other using dRep (Olm et al. 2017). *C. parapsilosis* genomes from the NICU sink of infant S2_005 and gut of infant N3_182 were more similar to reference genomes than each other (Additional file 10: Figure S2.3), and thus do not represent direct strain transfer events.

2.3.5 Sequence analysis of new genomes

De novo assembly of eukaryotic genomes from metagenomes allows not only for the detailed genomic comparison and detection of novel organisms, but also for the determination of ploidy, aneuploidy (abnormal number of chromosomes in a cell), heterozygosity, and population microdiversity of organisms in vivo. Changes in ploidy and aneuploidy have been observed in many eukaryotes, especially yeasts (Butler et al. 2009; Kathovade et al. 2010), and are thought to

be a strategy for relatively quick adaptation to shifts in environmental conditions. To determine the ploidy of genomes reconstructed in this study (Table 2.1), we examined the read count for each allele at a given variant site. For a diploid genome, alleles are expected to have a read count of 50%; for a triploid genome, alleles are expected to have a read count of either 33% or 67%. At low coverage, determining allele frequency with read mapping has more stochasticity relative to high coverage. Simulated reads for haploid, diploid, and triploid genomes at 10× and 100× coverage suggest it is possible to determine ploidy in even our low coverage genomes (Additional file 11: Figure S2.4). Based upon this analysis, all but one of our reconstructed genomes are diploid (Additional file 12: Figure S2.5). *C. lusitaniae* is likely haploid. Similarly, aneuploidy can be detected by searching for regions where allele frequencies and/or read coverage differs from the rest of the genome. For diploid genomes reconstructed from metagenomes, the sequences for each chromosome are a composite of sequences from the two alleles. Population microdiversity can be detected based on read counts that exceed the expected ratio of 50%. Measuring population microdiversity in this way can be confounded by sequencing error and stochastic read coverage variation (Additional file 11: Figure S2.4). Genomic datasets for isolates are not expected to have population microdiversity but will display sequencing error and stochastic read coverage variation. Consequently, we could separate sequencing noise from true population microdiversity by comparing the patterns we observed in our population genomic data to microdiversity found in isolate genomic datasets (Dawson et al. 2007). For *C. parapsilosis* N3_182_000G1, the peak of allele frequencies is wider than that of the sequenced *Candida parapsilosis* isolate (Fig. 2.4a), suggesting considerable population microdiversity. The *P. lilacinum* contaminant also displayed substantial microdiversity (Additional file 15: Figure S2.10). To avoid the stochasticity introduced by low sequencing coverage (Additional file 11: Figure S2.4), only genomes with over 50× sequencing coverage were analyzed for population microdiversity in this way.

Another method of measuring population microdiversity involves determining the number of multiallelic sites (sites with more than two sequence variants). Tests with simulated reads were performed to confirm that non-specific mapping of reads from unrelated species does not bias results (see the "Methods" section). All of our genomes have more multiallelic sites than isolate-sequenced genomes (Fig. 2.4b), suggesting that all of our genomes have appreciable population microdiversity. Further, genomes from the room had higher microdiversity than those from the gut, although this comparison is not statistically significant ($p = 0.09$).

Finally, overall heterozygosity for each genome was measured by calculating the number of heterozygous SNPs per kilo-base pair (Fig. 2.4c). A wide range of heterozygosity was observed within genomes. For most organisms, there was low heterozygosity, and for *C. albicans* and *C. parapsilosis*, comparable to that of reference isolates. *Malassezia restricta* S2_018_000R1 has both a particularly high rate of SNPs per kilo-base pair and high population microdiversity.

2.4 Discussion

2.4.1 Eukaryotic genome recovery from metagenomes augments information from isolate studies

In contrast with prior studies that have investigated microbial eukaryote genomes via sequencing of isolates, we employed a whole-community sequencing approach and could detect population microdiversity in both NICU and infant-derived samples. Malassezia on NICU surfaces has

particularly high population microdiversity. Given that Malassezia are skin-associated fungi (Gaitanis et al. 2012), their high population microdiversity may be the consequence of the accumulation of numerous strains throughout the hospital via shedding of skin from different individuals. This could also reflect naturally large population variation present within the skin of a single individual, as has been reported for skin-associated bacteria (Oh et al. 2014; Tsai et al. 2016).

In the current analysis, most of the samples contained one dominant eukaryotic genotype, presumably one well adapted to the habitat, but other allele variants indicate the presence of lower-abundance genotypes (Fig. 2.4b). Given this dominance, it was possible to directly estimate genome heterozygosity. Prior studies have reported that *C. albicans* grows clonally *in vivo* (Hirakawa et al. 2015), yet Candida, when expressing a certain phenotype, undergoes mating (Bennett et al. 2003), most likely via a parasexual cycle (Bennett et al. 2003). For *C. albicans*, the measured heterozygosity was comparable to that of previously sequenced isolate genomes (Hirakawa et al. 2015; Jones et al. 2004). Despite high heterozygosity of *C. albicans*, we see low strain heterogeneity. It has been hypothesized that *C. albicans* mating may occur primarily on the skin (Lachke et al. 2003). We speculate there may be more strain heterogeneity on the skin or other areas of the human microbiome besides in the gut, as it is probable that heterozygosity in Candida populations in the human and room microbiomes arises due to mating with distinct coexisting strains.

The heterozygosity measurements of all other fungi except Malassezia were low, possibly indicating diversity reduction due to inbreeding and/or strong selection for specific alleles. We speculate that this reflects a long history of colonization of a habitat type that strongly selects for a specific genotype, so genome structure reflects the relatively low probability of recombination with strains with divergent alleles (in other words, the presence of gut-adapted and sink-adapted strains). However, without the availability of similar genomes to compare to from other habitats, we cannot rule out genetic bottlenecks that took place prior to introduction to the hospital.

An important aspect of the current study is the sequencing of reagent controls, which allowed us to identify *P. lilacinum* as a likely contaminant. It is interesting to note that peak allele frequency analysis indicated high population microdiversity for the contaminant. Genomic microdiversity of the reagent-associated population may indicate its long-term persistence in the reagents, analogous to that shown for Delftia metagenome contamination that was present in Pippin size selection cassettes for many years (Olm et al. 2017). Given the increasing use of metagenomic sequencing for pathogen detection and prior reports of *P. lilacinum* as both a contaminant and disease agent (Shivaprasad et al. 2013; Luangsa-ard et al. 2011), it will be important to rule out a reagent source of *P. lilacinum* in future diagnostic studies.

2.4.2 Premature infants are colonized by eukaryotes early in life

Six percent of infants in this study were colonized by fungi, lower than most previous studies of infants (Baley et al. 1986; Stewart et al. 2012; Stewart et al. 2013; LaTuga et al. 2011). Compared to shotgun sequencing, DGGE and ITS methods should be more sensitive due to the use of PCR, and thus may be more suitable for broad ecological surveys. However, the ability to amplify very rare sequences from organisms present at exceedingly low abundance levels

complicates interpretation of the measured colonization frequencies. Our shotgun sequencing-based methods provide a more balanced view of community composition than methods that rely on PCR, and detection of populations that comprise more than ~ 0.05% of the community DNA is possible with read-mapping (Additional file 1: Table S2.1; Additional file 6: Figure S2.2). Further, whole-community sequencing measures the relative abundance of eukaryotes in the context of the whole community, something that cannot be done using ITS, DGGE, or culturing-based methods. Fungi are generally considered low-abundance members of the gut microbiome (Schulze et al. 2009), yet in this study, they reached levels as high as 55%, 78%, and 96% of the entire community (Fig. 2.2). Differences in fungal communities during early life are known to have effects on infant health later in life (Fujimura et al. 2016), and it remains to be seen if extreme abundance levels like this have long-lasting effects.

All infants profiled in this study received 2–7 days of prophylactic antibiotics upon birth, meaning antibiotic use is highly correlated with earlier days of life (Additional file 7: Table S2.4). While both antibiotic administration and DOL were significantly correlated with eukaryote abundance, consistent with previous studies of fungal colonization of low birth weight infants (Baley et al. 1986; Huang et al. 2000), infants who received antibiotics later in life were not colonized by eukaryotes. This suggests that day of life is the more important factor. However, eukaryotes may have not been detected in later collected microbiome samples from those infants due to increased relative abundance of bacteria. In other words, the sensitivity of the shotgun sequencing method may be insufficient to detect fungi that persist at low abundance.

Interestingly, permutation testing revealed that fungi colonized the same infants more often than expected by random chance. There may be several explanations for this phenomenon. For example, some infants may be more genetically predisposed to fungal colonization. Alternatively, fungi may interact synergistically, with the first colonizing species establishing a niche in the gut that makes it more suitable for other fungi. Should this effect prove to be important, it may help to explain how fungal colonization contributes to development of asthma or allergies (Fujiumura et al. 2016).

2.4.3 Differences in colonization patterns of NICU sinks and surfaces

Yeasts of the genus Malassezia, a common member of the skin microbiome (Parfrey et al. 2011; West et al. 2018), NICU surfaces (Parfrey et al. 2011; Gaitanis et al. 2012). This result is analogous to findings of previous studies, which showed that typically skin-associated bacteria dominate consortia associated with hospital surfaces and parts of other built environments (Brooks et al. 2017; Sanchez et al. 1992; Hewitt et al. 2013; Jones et al. 2004; Oh et al. 2014).

The same eukaryotes were never detected in sinks and surfaces, and the sinks hosted a comparatively diverse and variable eukaryotic community (Fig. 3). Sinks are inherently heterogeneous environments with different moisture levels and chemical conditions. Punctuated cleaning events may also give rise to temporal variation. Diptera S2_005_002R2 (fly), which was present in only one sink sample, may be explained by sequencing of sink-associated eggs, as no macroscopic organisms were detected during the collection process. Recent studies have suggested that insects play significant roles in the dispersal of fungi, and this may occasionally occur in the NICU (Tsai et al. 2016).

The other metazoan detected, the worm Rhabditida S2_005_001R2, was found in sinks from multiple rooms and samples collected months apart. These organisms may also be a source of fungi, and like the fly, could impact the overall NICU microbiome. Intriguingly, the partial genome appears to derive from an organism that is equally related to a bovine lungworm and *Caenorhabditis elegans* and is potentially novel at the class level (Fig. 2.1). Although we cannot evaluate its medical importance, the organism may have been macroscopically described but lack of a reference genome prevents identification.

2.5 Conclusions

We applied genome-resolved metagenomics to study eukaryotes in the gut microbiomes of infants and their NICU rooms and detected eukaryotes associated with pathogenesis of immunocompromised humans, commensals of human skin, and fungi typical of environments such as soil and drain pipes. Genomic analysis of diploid organisms found low rates of heterozygosity that may be explained by persistence of hospital-associated lineages in environments that impose strong selective pressure. The application of this approach in other contexts should greatly expand what is known about eukaryotic genomic diversity and population variation.

2.6 Methods

2.6.1 Subject recruitment, sample collection, and metagenomic sequencing

This study made use of many different previously analyzed infant datasets. These datasets have previously published descriptions of the study design, patient selection, and sample collection, and are referred to as NIH1 [51, 52], NIH2 [19], NIH3 [53], NIH4 [54], Sloan2 [19], and SP_CRL [55]. Infants were chosen for inclusion in this study irrespective of fungal disease state. Negative extraction controls were performed and sequenced during the sequencing of the Sloan2 cohort. The last well of the extraction block (H12) was left empty, and this well was treated the same as all other samples throughout the extraction protocol. It is therefore a control for the kit reagents, the sterility of the kit tubes/plates, and the aseptic technique of the technician who performed the extraction. S2_CON_001E1, S2_CON_002E1, and S2_CON_003E1 were all on different extraction blocks, and S2_CON_002E2 was a second well on the same block as S2_CON_002E1.

This study also involved the collection and processing of an additional 269 samples from 53 infants. Newly collected infant fecal samples followed the same sample collection and DNA extraction protocol as described previously [53, 56]. Metagenomic sequencing of newly collected infant fecal samples was performed in collaboration with the Functional Genomics and Vincent J. Coates Genomics Sequencing Laboratories at the University of California, Berkeley. Library preparation on all samples was performed using the following basic protocol: (1) gDNA shearing to target a 500 bp average fragment size was performed with the Diagenode Bioruptor Pico, (2) end repair, A-tailing, and adapter ligation with an Illumina universal stub with Kapa Biosystems Hyper Plus Illumina library preparation reagents, and (3) a double AMpure XP bead cleanup, followed by indexing PCR with dual-matched 8 bp Illumina compatible primers. Final sequence ready libraries were visualized and quantified on the Advanced Analytical Fragment

Analyzer, pooled into 11 subpools based on mass, and checked for pooling accuracy by sequencing on Illumina MiSeq Nano sequencing runs. Libraries were then further purified using 1.5% Pippin Prep gel size selection assays collecting library pools from 500 to 700 bp. Pippin pools were visualized on fragment analyzer and quanted with Kapa Illumina library quant qPCR reagents and loaded at 3 nM. The 11 pools were then sequenced on individual Illumina HiSeq4000 150 paired-end sequencing lanes with 2% PhiX v3 spike-in controls. Post-sequencing bcl files were converted to demultiplexed fastq files per the original sample count with Illumina's bcl2fastq v2.19 software. New metagenomic data was processed in the same manner as in the prior studies, and as described previously [54].

Environmental metagenomes were described and published previously as part of the Sloan2 cohort study [19]. All samples were collected over a roughly one-year period from the same NICU at the University of Pittsburgh Magee-Womens Hospital. In order to generate enough DNA for metagenomic sequencing, DNA was collected from multiple sites in the NICU and combined into three separate pools for sequencing. Highly-touched surfaces included samples originating from the isolette handrail, isolette knobs, nurses hands, in-room phone, chair armrest, computer mouse, computer monitor, and computer keyboard. Sink samples included samples from the bottom of the sink basin and drain. Counters and floors consisted of the room floor and surface of the isolette. See previous publications for details [19, 57].

2.6.2 Eukaryotic genome binning and gene prediction

Reads from each sample were assembled independently using IDBA-UD [58] under default settings. A co-assembly was also performed for each infant, consisting of reads from all samples taken from that infant concatenated together. Binning assembled sequence scaffold into eukaryotic genomes was performed using a EukRep-based pipeline, described in detail in West et al. [30]. In cases where time-series data were available, samples were pre-binned using time-series information and eukaryotic bins were then subsequently identified with EukRep. In cases where multiple genomes of the same organism were recovered from multiple samples from the same infant, the most complete genome was selected for further analysis. In addition to the gene prediction methodology outlined previously [30], a second homology-based gene prediction step was performed. Ribosomal S3 (rpS3) proteins were identified in genomes using a custom ribosomal protein S3 (rpS3) profile HMM, and identified sequences were searched against the NCBI database [59] and UniProt [60] using BLAST [61]. For each de novo-assembled genome, gene sets for the top 1–3 most similar organisms were used as homology evidence for a second-pass gene prediction step with AUGUSTUS [62], as implemented in MAKER [63]. For Rhabditida S2_005_001R2, first-pass gene predictions were used, as homology evidence decreased overall estimated genome completeness. Genome completeness was estimated using BUSCO [64] and is based on the number of detected single-copy orthologs. N50 was calculated using the program checkM [65].

To verify bins, the taxonomy of each scaffold was determined by searching gene sequences against the UniProt database [53]. All bins were found to have a consistent phylogenetic signal, except the bin created from sample S2_009_000R2. Scaffolds had similar GC content and sequencing coverage, but were either dominated by genes with homology to the class Sordariomycetes or Eurotiomycetes. Scaffolds from the original "megabin" were split into two separate bins based on this phylogenetic signal, resulting in the genomes *Nectria haematococca*

S2_009_000R2 and Exophiala sp. S2_009_000R2. Gene prediction was run again for both of these genomes, as described above.

2.6.3 Phylogenetic analyses

In order to construct a phylogenetic tree, rpS3 proteins from each de novo genome were detected as described above and searched against the NCBI database using BLAST. Protein sets of the 3–5 most similar organisms on NCBI were downloaded for inclusion. Other phylogenetically important genomes, such as *A. thaliana*, were included as well. For each protein set, 16 ribosomal proteins (bacterial ribosomal protein names L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24, S3, S8, S10, S17, and S19) were identified using custom-built hidden Markov models (HMMs) with HMMER [66], using the noise cutoff (NC). The 16 ribosomal protein datasets were then aligned with MUSCLE [67] and trimmed by removing columns containing 90% or greater gaps. The alignments were then concatenated. A maximum likelihood tree was constructed using RAxML v.8.2.10 [68] on the CIPRES web server [69] with the LG plus gamma model of evolution (PROTGAMMALG) and with the number of bootstraps automatically determined with the MRE-based bootstrapping criterion. The constructed tree was visualized with Interactive Tree of Life (ITOL) [70].

Average nucleotide identity (ANI) between binned genomes and reference genomes was determined with dRep [35]. Resulting whole genome ANI values were used in combination with a 16 ribosomal protein phylogenetic tree to determine the taxonomy of de novo genomes. For genomes without a species-level taxonomy, genomes were searched against the entire NCBI nucleotide database using BLAST. This resulted in a species-level call for *Malassezia restricta* S2_018_000R1. For genomes without a genus-level taxonomy (Rhabditida S2_005_001R2 and Diptera S2_005_002R2), an additional step was taken. Mitochondrial cytochrome c oxidase subunit I (COI) genes were identified by searching *D. melanogaster* and *C. elegans* COI genes against our PRODIGAL [71] predicted genes sets with UBLAST [72]. Significant hits from our protein sets were then searched against the Barcode Of Life Database (BOLD) [31] and NCBI in order to identify sequences with high identity to our novel genomes. No significant hits were identified.

2.6.4 Mapping-based genome detection

To detect eukaryotes in an assembly-free manner, reads were mapped to a curated genome collection. This genome collection consists of all fungal genomes in RefSeq (accessed 9/14/17) [73], as well as genomes assembled in this study with no close representatives in RefSeq (average nucleotide identity of 90% or higher according to Mash [74]). The six genomes with no close representatives in RefSeq were Malassezia restricta S2_018_000R1, Diptera S2_005_002R2, Exophiala sp. S2_009_000R2, Verruconis sp. S2_005_001R2, and Rhabditida S2_005_001R2. Candida parapsilosis CDC317 was also included, as there were no genomes of C. parapsilosis in RefSeq.

Reads from all samples were mapped to this reference genome list using Bowtie 2 [75]. To determine which organisms were present in each sample, we primarily relied on breadth of coverage as reported by strainProfiler (https://github.com/MrOlm/strainProfiler). In NICU samples, all genomes with 50% breadth of coverage or above were considered present. For infant samples, reads resulting from concatenating all samples belonging to the same infant were first

used to determine which fungi are reliably detected. Genomes with 50% breadth of coverage or above were considered present with two exceptions, *Malassezia pachydermatis* and *Malassezia sympodialis*, at ~ 0.2 and 0.4 breadth, respectively. Considering the extensive and distributed breadth of coverage for these genomes (Additional file 3: Figure S1C), they were considered present in the infant despite having low breadth of coverage overall. Reads from each individual sample from each infant were then mapped to all fungi considered to be present in that infant to determine changes over time. Relative abundance of genomes was determined using the formula: (number of reads mapping to genome/total number of reads in sample).

The lowest coverage genome with this breadth threshold was 1.1× coverage. To determine the limit of detection, we first determined the relative abundance needed to achieve 1.1× coverage using the median infant co-assembly depth (27.5 Gb) and the median eukaryotic genome length in our database of organisms that were detected at least once (13.7 Mbp). We then calculated the limit of detection using the formula ((min coverage × median length)/median co-assembly depth). This led to an estimated limit of detection of 0.05% relative abundance for infant fungi detection, through this number has significant variability depending on how deep each individual infant was sequenced.

2.6.5 Negative extraction control analysis

Sequences resulting from negative extraction controls were computationally processed in an identical manner to other samples. Reads from all control samples were mapped to the curated genome collection described above, and the relative abundance of all genomes with at least 10% breadth was plotted in Additional file 3: Figure S1. The program strainProfiler (https://github.com/MrOlm/strainProfiler) was used to compare reads in sample S2_CON_000E3 to *P. lilacium* genomes assembled in this study and all publicly available *P. lilacinum* genomes. Version 0.2 of the program was run with default settings, resulting in an average nucleotide identity measure between sample S2_CON_000E3 and all *P. lilacinum* genomes. Next, dRep v1.4.3 [35] was used to compare the *P. lilacinum* genomes with each other using the command "dRep cluster --SkipMash". The resulting distance matrix was merged with the values generated from strainProfiler to generate the dendrogram in Additional file 3: Figure S1B. Full code for implementation is available at https://github.com/MrOlm/InfantEukaryotes.

All publically available Malassezia genomes were acquired by searching for the term "Malassezia" in the assembly section of NCBI and downloading them manually. Genomes were compared to each other, and representative genomes were chosen using dRep v1.4.3 and the commands "dRep compare --SkipMash" and "dRep choose --noQualityFiltering -sizeW 0.5". A concatenation of all negative extraction control sequences was then mapped to the resulting genomes using Bowtie 2. Custom scripts were used to determine the breadth of coverage of each 10,000 bp window of each fungal genome in each sample, and each window with at least 50% breadth was marked with a tick using Circos [76] to visualize. Open source code detailing this analysis is available at https://github.com/MrOlm/InfantEukaryotes.

To determine the expected breadth of coverage (percentage of genome base pairs with at least one read) for a given depth of coverage (average number of reads at any given genome base pair), a simulation was performed. Metagenomic reads were first simulated for *Escherichia coli* and *Candida albicans* reference genomes using pIRS (https://github.com/galaxy001/pirs).

Simulated reads were mapped back to the original reference genome, and the resulting .bam file was subset 20 times to simulate various depths of coverage. The breadth and depth of coverage was plotted and an exponential line of best fit was calculated using SciPy [77]. The line had an R2 value over 0.99 and was defined using the equation:
breadth = (− 1 ×e^(− 0.883 × coverage)) + 1. This equation was used to determine the expected breadth of coverage for a given depth of coverage.

2.6.6 Statistical analyses and generation of MDS plot

To compare the eukaryotic communities present in NICU room samples, multidimensional scaling (MDS) based on Bray-Curtis distance was performed. The Bray-Curtis distance was calculated based on the relative abundance of each eukaryote present in a sample using the python library SciPy (command scipy.spatial.distance.braycurtis) [77]. Eukaryotes with at least 50% breadth of coverage were considered present in a sample. MDS was performed on the resulting all-vs-all distance matrix using the python library sklearn (command sklearn.manifold. MDS) [78]. MDS was plotted using a custom function in Matplotlib [79]. Stress was calculated using sklearn. Open source code detailing this analysis is available at https://github.com/MrOlm/InfantEukaryotes.

We tested for significant associations between samples containing eukaryotes and various forms of metadata using the python SciPy package [77]. Included were six pieces of continuous metadata (DOL, infant birth weight, etc.), 23 pieces of categorical metadata (specific antibiotics given and specific NICU room locations), and the phyla-level abundance of all bacterial genomes (seven total phyla) (Additional file 7: Table S4). Bacterial phyla-level abundance was determined by summing the relative abundance of all bacterial genomes present in a sample. Bacterial genomes for previously sequenced samples are available in a previous publication [54], and bacterial genomes for newly sequenced genomes were binned using the same methods. Metadata was filtered such that between 20 and 80% of values were non-zero in both samples containing eukaryotes and samples not containing eukaryotes. This resulted in a total of 13 pieces of metadata for statistical testing (Additional file 7: Table S4).

In order to eliminate statistical bias introduced through sampling the same infant multiple times, one sample from each infant was chosen for statistical tests. If the infant was not colonized by a eukaryote, the sample was chosen at random. If the infant was colonized by a eukaryote, the sample with the highest eukaryotic abundance was chosen. Samples were considered to have a eukaryote present if the sum of the relative abundance of eukaryotes with at least 50% breadth was at least 0.1% relative abundance. Fisher's exact test was used for categorical metadata, and Wilcoxon rank-sum test was used for continuous data. Benjamini-Hochberg p value correction [80] was performed to account of multiple hypothesis testing. The results of all statistical tests are provided in Additional file 8: Table S5. Open source code detailing this statistical analysis is available at https://github.com/MrOlm/InfantEukaryotes.

A permutation test was performed to determine if fungi were distributed randomly among the infants. First, 100,000 trials were run where each trial consisted of randomly selecting 13 individuals with replacement from a total population of 161 individuals. The number of infants chosen was determined for each trial, and an empirical p value was determined based on how

many trials had 10 of less infants chosen. Open source code detailing this statistical analysis is available at https://github.com/MrOlm/InfantEukaryotes.

2.6.7 Ploidy, heterozygosity, and population microdiversity

In order to identify variants, reads from the sample in which a particular genome was binned from were mapped back to the de novo assembled genome using Bowtie 2 [75] with default parameters. The PicardTool (http://broadinstitute.github.io/picard/) functions "SortSam" and "MarkDuplicates" were used to sort the resulting sam file and remove duplicate reads. FreeBayes [81] was used to perform variant calling with the options "--pooled-continuous -F 0.01 -C 1." Variants were filtered downstream to include only those with support of at least 10% of total mapped reads in order to avoid false positives. Furthermore, to avoid including variants as a result of mismapping reads, variants were filtered to include only those with coverage depth within a range of the average genome coverage plus or minus half of the genome mean coverage. SNP read counts were calculated using the "AO" and "RO" fields in the FreeBayes vcf output file. Multiallelic sites were defined as sites with two or more non-reference alleles. Variants were called using the same methodology for both simulated read datasets and isolate genomes. Variants were used to determine ploidy, heterozygosity, and population microdiversity as described in the "Results" section. Source code with full implementation details is available at https://github.com/MrOlm/InfantEukaryotes.

To confirm that multiallelic sites are not the result of non-specifically mapped reads from the bacterial community, we fragmented with pIRS (https://github.com/galaxy001/pirs) a diploid *C. parapsilosis* genome into simulated reads and added these reads to an infant gut metagenome sample without *C. parapsilosis*. The resulting read dataset along with a separate dataset comprised of only the simulated reads were then mapped to the original *C. parapsilosis* genome. No additional variants were detected between the sample with metagenomic reads and the sample without, indicating non-specifically mapped reads from bacterial community members have a minimal effect.

In order to determine the effect of stochastic read coverage on variant frequencies, simulated haploid, diploid, and triploid genomes were generated using the pIRS (https://github.com/galaxy001/pirs) diploid command with the *C. albicans* P57072 reference genome. The command was used once to generate a diploid genome and twice to generate a triploid genome. Simulated reads were then generated for each genome using the pIRS simulate command at 10×, 50×, and 100× coverage. Assemblies and raw reads were downloaded for both *C. albicans* A48 and *C. parapsilosis* CDC317 from NCBI to be used as example isolate genomes for comparison. Based on this analysis, only the two genomes with at least 50× coverage were included in peak allele frequency analysis.

Genome aneuploidy was analyzed in two ways. First, reads from each sample were mapped back to genomes assembled from that sample. The coverage of each scaffold was determined in 10 kbp windows, and the coverage of all windows for each scaffold over 10 kbp was plotted. Plots were then analyzed for scaffolds with differing coverage, indicative of the presence of multiple copies of a subset of the chromosomes (Additional file 13: Figure S6). Second, reads from samples with genomes assembled from them were mapped to the closest available reference genome. The same procedure was then performed with these reference genomes in all cases where at least 80% of the

genome was covered by reads. This allowed the determination of aneuploidy on the whole-chromosome level (Additional file 14: Figure S7). Both methods agreed that in all cases, no aneuploidy was detected.

2.7 Figures

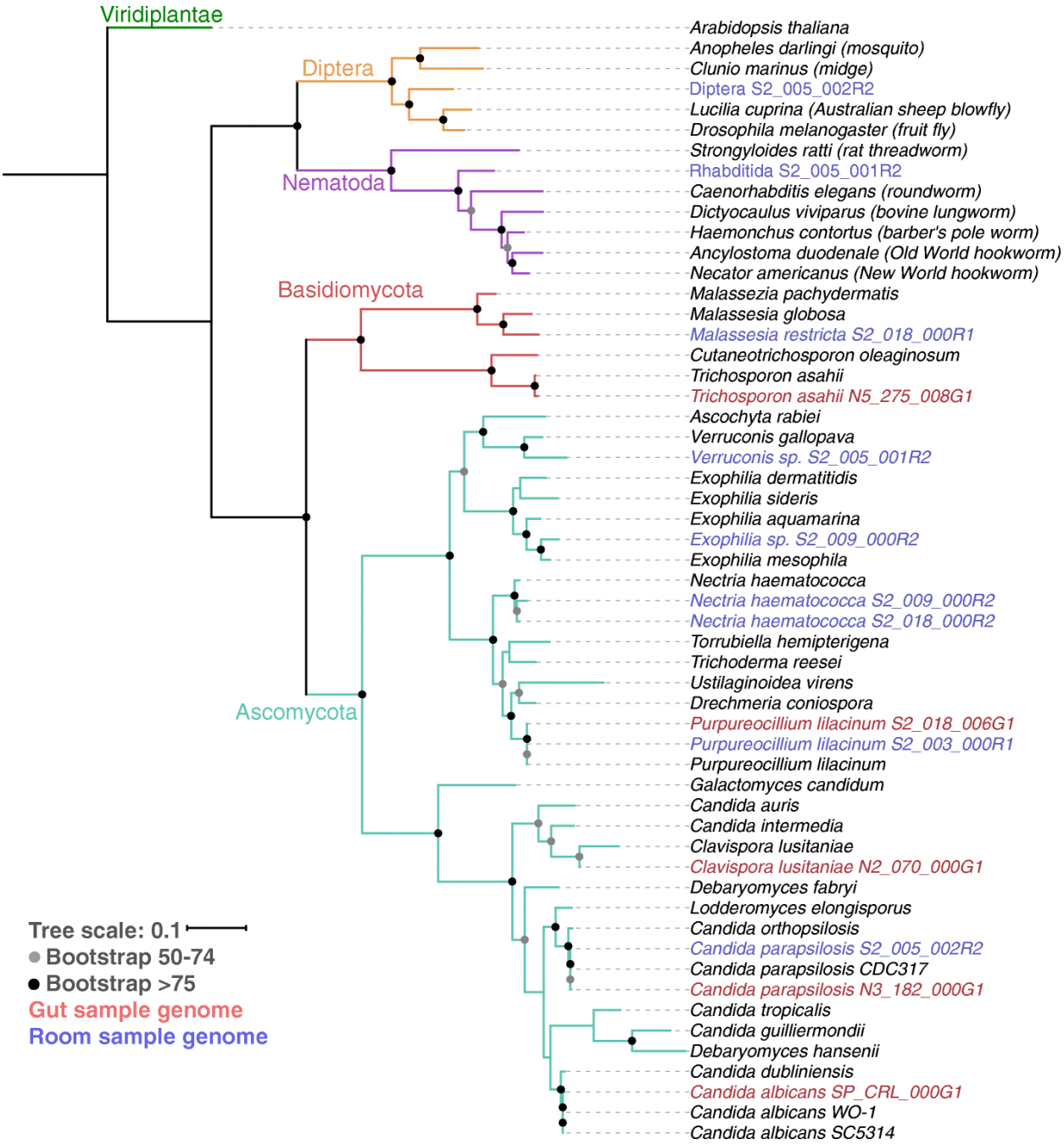

Figure 2.1. Phylogenetic tree of recovered eukaryote genomes. Genomes from infant-derived fecal samples (red) and NICU samples (blue) were classified using a phylogenetic tree based on the concatenation of the sequences of 16 ribosomal proteins (see the "Methods" section). Branches with greater than 50% bootstrap support are labeled with their bootstrap support range. Reference ribosomal protein sequences were obtained from NCBI [30] and the Candida Genome Database [30].

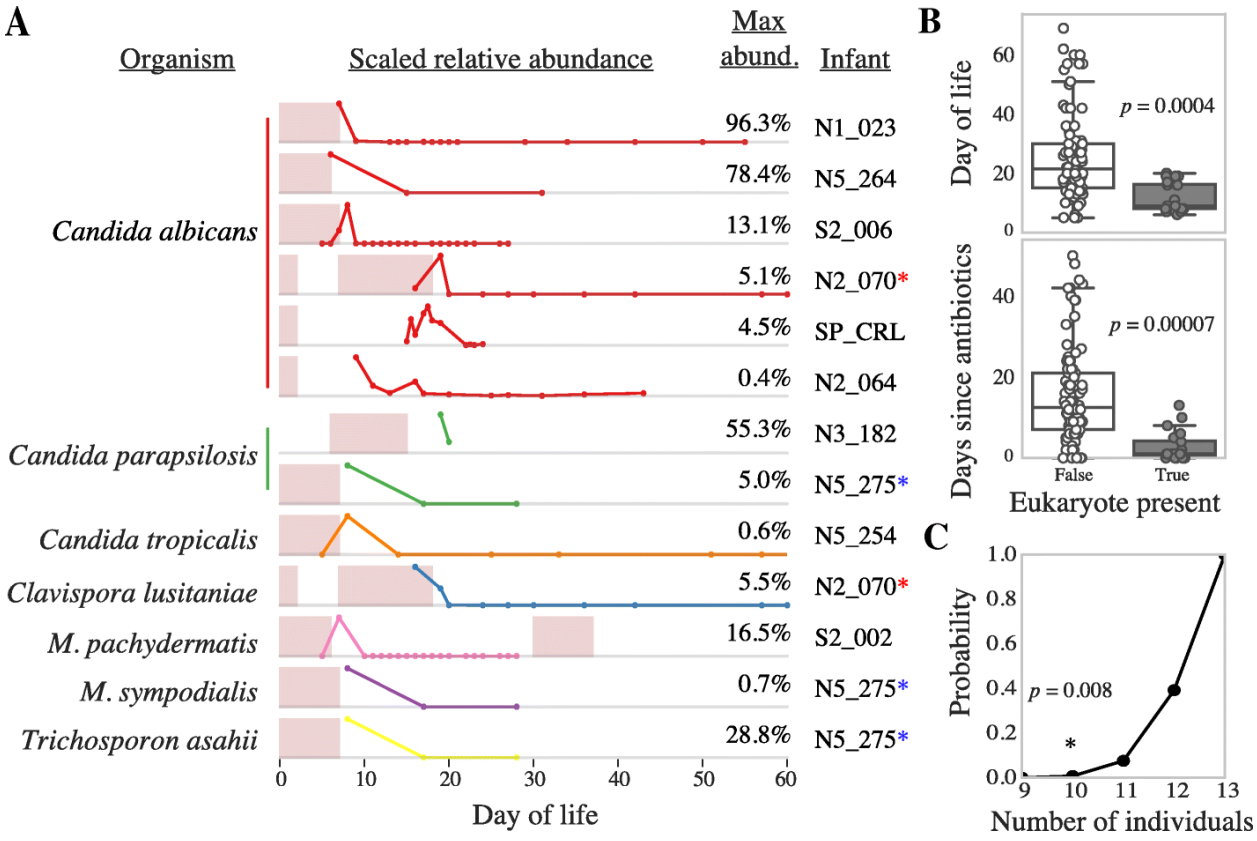

Figure 2.2. Abundance of eukaryotes colonizing infants. a The scaled relative abundance of each eukaryote colonizing an infant. Numbers on the right indicate the maximum relative abundance of the organism in that infant, and gray dividing lines indicate 0% relative abundance. Dots on the line-plots indicate days of life on which fecal samples were collected and sequenced. Infants colonized by multiple eukaryotes are marked with a colored asterisk. Pink bars indicate periods of antibiotic administration. b Metadata significantly associated with eukaryote abundance. The distribution of values for all samples in which eukaryotes are not present (left; white box plot) compared to values of samples in which eukaryotes are present (right; gray box plot). The p values were calculated using the Wilcoxon rank-sum test with Benjamini-Hochberg multiple testing p value correction. *P. lilacinum* was excluded from statistical tests due to its likely contaminant status. c Fungi are distributed among fewer individuals than expected by random chance. A permutation test was performed to determine the probability of observing 10 or less unique individuals colonized by 13 fungi from a population of 161 individuals. The number of unique individuals colonized is shown on the x-axis, and the empirical p value based on 100,000 trials is shown on the y-axis. An asterisk marks the true number of unique infants colonized in this study (10) and the associated p value.

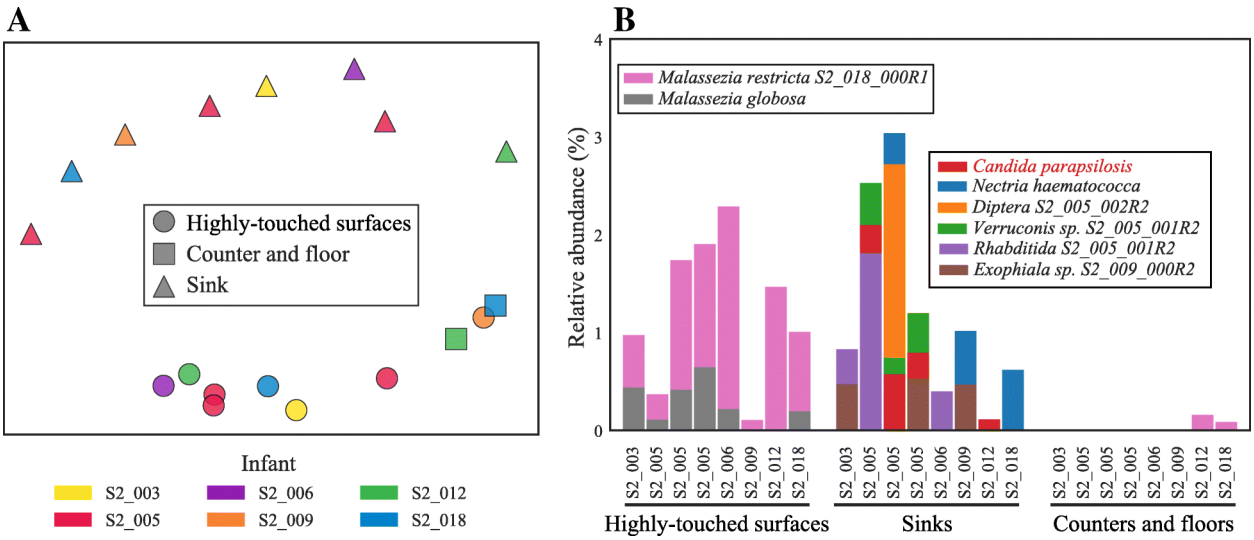

Figure 2.3. Eukaryotic microbiome of the neonatal intensive care unit (NICU). a Multidimensional scaling (MDS) of the Bray-Curtis dissimilarity between all NICU samples. Samples cluster by environment type rather than the room or occupant. The stress of the MDS was calculated to be 0.23. b Compositional profile of eukaryotic organisms detected in the NICU. Each colored box represents the percentage of reads mapping to an organism's genome, and the stacked boxes for each sample show the fraction of reads in that dataset accounted for by different eukaryotic genomes in each sample.

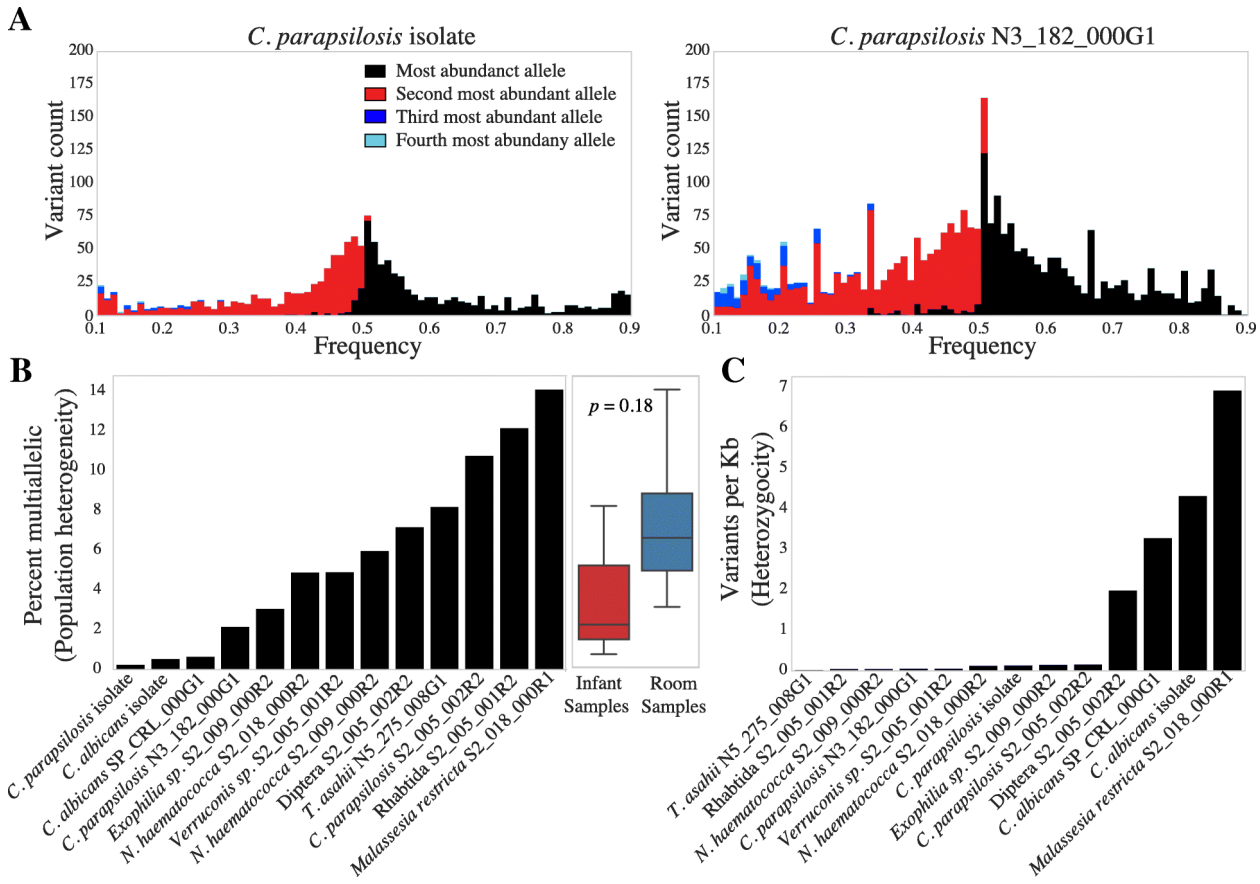

Figure 2.4. Ploidy, zygosity, and microdiversity of recovered eukaryotic genomes. a Histogram of the frequencies of the four most abundant variants at each variant site in an isolate genome of *C. parapsilosis* and in a genome of *C. parapsilosis* recovered in this study. Black, red, dark blue, and light blue bars indicate the abundances of the most abundant, second, third, and fourth most abundant variant, respectively. b For each genome, black bars indicate the percentage of variant sites that are multiallelic (contain more variants at a site than would be expected based upon ploidy alone). Haplotypes with more than two alleles are also considered to be multiallelic. A box plot compares the values from genomes originating from infant guts vs. the NICU room. c For each genome, black bars indicate the number of heterozygous variants per kb across the entire assembled genome.

Table 2.1. Description of de novo assembled eukaryotic genomes

Source	Genome	Completeness	Length (bp)	N50 (bp)	Coverage
Infant gut	*Purpureocillium lilacinum S2_018_006G1*	98.4	35,688,710	422,361	20×
Infant gut	*Clavispora lusitaniae N2_070_000G1*	95.8	11,907,650	89,311	18×
Infant gut	*Candida parapsilosis N3_182_000G1*	96.7	12,563,647	65,710	182×
Infant gut	*Trichosporon asahii N5_275_008G1*	90.1	23,419,590	32,912	13×
Infant gut	*Candida albicans SP_CRL_000G1*	91.1	12,561,678	22,840	30×
NICU room	*Purpureocillium lilacinum S2_003_000R1*	98.4	35,724,498	520,486	67×
NICU room	*Malassezia restricta S2_018_000R1*	72.6	6,457,898	4912	18×
NICU sink	*Nectria haematococca S2_018_000R2*	96.7	44,952,822	24,418	10×
NICU sink	*Candida parapsilosis S2_005_002R2*	92.8	11,573,959	14,507	9×
NICU sink	*Rhabditida S2_005_001R2*	74.9	50,505,025	8214	8×
NICU sink	*Nectria haematococca S2_009_000R2*	73.6	31,143,909	8000	7×
NICU sink	*Exophiala sp. S2_009_000R2*	75.9	24,670,482	7386	7×
NICU sink	*Diptera S2_005_002R2*	52.5	43,769,201	6834	10×
NICU sink	*Verruconis sp. S2_005_001R2*	52.8	15,639,153	5112	6×

Table 2.2. Description of detected fungal taxa

Taxa	Common habitats	Pathogenicity	Number of infants	Locations In NICU	Refs
Candida albicans	Warm blooded animals	Common nosocomial pathogen	6	Undetected	[1]
Candida parapsilosis	Warm blooded animals	Common nosocomial pathogen (especially neonates)	2	Sink	[82]
Candida tropicalis	Warm blooded animals	Common nosocomial pathogen	1	Undetected	[83]
Nectria haematococca	Soil, rhizosphere	Pathogen of immunocompromised patients	0	Sink	[84]
Malassezia sympodialis	Human skin	Opportunistic pathogen	1	Undetected	[85]
Malassezia globosa	Human skin	Common commensal; implicated in dandruff	0	Surfaces	[86]
Malassezia pachydermatis	Skin of mammals	Opportunistic pathogen	1	Undetected	[87]
Trichosporon asahii	Soil, human skin and GI tract	Rare opportunistic pathogen	1	Undetected	[88]
Verruconis	Soil, decaying vegetation	Verruconis includes black yeasts; human pathogens	0	Sink	[89]

For supplemental figures, tables, and information for Chapter 2, see https://www.biorxiv.org/content/10.1101/597468v1

3 Genetic and behavioral adaptation of *Candida parapsilosis* to the microbiome of hospitalized infants revealed by *in situ* genomics, transcriptomics and proteomics

Patrick T. West, Samantha L. Peters, Matthew R. Olm, Feiqiao B. Yu, Yue C. Lou, Brian A. Firek, Robyn Baker, Alexander D. Johnson, Michael J. Morowitz, Robert L. Hettich, Jillian F. Banfield

3.1 Abstract

3.1.1 Background

Candida parapsilosis is a common cause of invasive candidiasis, especially in newborn infants, and infections have been increasing over the past two decades. *C. parapsilosis* has been primarily studied in pure culture, leaving gaps in understanding of its function in microbiome context.

3.1.2 Results

Here, we reconstructed five unique *C. parapsilosis* genomes from premature infant fecal samples and analyzed their genome structure, population diversity and *in situ* activity relative to reference strains in pure culture. All five genomes contain hotspots of single nucleotide variants, some of which are shared by strains from multiple hospitals. A subset of environmental and hospital-derived genomes share variants within these hotspots suggesting derivation of that region from a common ancestor. Four of the newly reconstructed *C. parapsilosis* genomes have four to sixteen copies of the gene RTA3, which encodes a lipid translocase and is implicated in antifungal resistance, potentially indicating adaptation to hospital antifungal use. Time course metatranscriptomics and metaproteomics on fecal samples from a premature infant with a *C. parapsilosis* blood infection revealed highly variable *in situ* expression patterns that are distinct from those of similar strains in pure cultures. For example, biofilm formation genes were relatively less expressed *in situ*, whereas genes linked to oxygen utilization were more highly expressed, indicative of growth in a relatively aerobic environment. In gut microbiome samples, *C. parapsilosis* coexisted with Enterococcus faecalis that shifted in relative abundance over time, accompanied by changes in bacterial and fungal gene expression and proteome composition.

3.1.3 Conclusions

The results reveal potentially medically relevant differences in Candida function in gut vs. laboratory environments, and constrain evolutionary processes that could contribute to hospital strain persistence and transfer into premature infant microbiomes.

3.2 Introduction

Candida species are the most common cause of invasive fungal disease (Naglik et al. 2008; Silva et al. 2012). A variety of Candida species cause candidiasis and are recognized as a serious public health challenge, especially among immunocompromised and hospitalized patients (Clerihew et al. 2007, Bliss, 2015). Historically, *Candida albicans* most commonly has been recognized as the cause of candidiasis, and as a result, is the focus of the majority of Candida research (Kuhn et al. 2004; Trofa et al. 2008; Bliss, 2015). However, *Candida parapsilosis*, despite being considered less virulent than *C. albicans*, is the Candida species with the largest increase in incidence since 1990 (Trofa et al. 2008). Given important differences in the biology of *C. albicans* compared to non-albicans species, more research on non-albicans Candida species, especially the subset that poses a serious health risk, is needed (Bliss, 2015).

C. parapsilosis is often a commensal member of the gastrointestinal tract and skin (Trofa et al. 2008; Gonia et al. 2017). Passage from hospital workers' hands to immunocompromised patients is thought to be a common cause of opportunistic infection in hospital settings (Huang et al. 1998). *C. parapsilosis* infections of premature infants are of particular concern. Indeed, *C. parapsilosis* is the most frequently isolated fungal organism in many neonatal intensive care units (NICUs) in the UK (Clerihew et al. 2007) and is responsible for up to one-third of neonatal Candida bloodstream infections in North America (Fridkin et al. 2006). Adding to the concern is the limited number of antifungal drugs and the increasing prevalence of antifungal drug resistance in Candida species. An estimated 3-5% of *C. parapsilosis* are resistant to fluconazole, the most commonly applied antifungal (Whaley et al. 2017). The recent emergence of multidrug-resistant *Candida auris* with its resultant high mortality rate (Forsberg et al. 2019) serves as a warning regarding the potential for outbreaks of multidrug-resistant *C. parapsilosis*. Therefore, understanding behavior of *C. parapsilosis*, both as a commensal organism and opportunistic pathogen, is incredibly important.

A challenge that complicates understanding of the medically relevant behavior of Candida in the human microbiome is that the hosts used in model infection systems (e.g., rat or murine mucosa) are not natural hosts to Candida species. Study of Candida in these models relies on some form of predisposition of the animal by occlusion, immunosuppression, surgical alteration, or elimination of competing microbial flora (Naglik et al. 2008). Pure culture experiments, an alternative to model system studies, are often the most accessible way to study Candida. However, the lack of a microbial community context is a large caveat, considering bacteria could influence the nutrition, metabolism, development, and evolution of eukaryotes. Indeed, other microbial eukaryotes have been shown to be dramatically influenced by their surrounding microbial communities. Choanoflagellates, the closest known living relative of animals, live in aquatic environments and feed on bacteria by trapping them in their apical collar (Hibberd et al. 1975). The Choanoflagellate *Salpingoeca rosetta* is primarily a unicellular organism but formation of multicellular rosettes is induced by a sulphonolipd (RIF1) and inhibited by a

sulfonate-containing lipid, both produced by the bacterium *Algoriphagus machipongonensis* (Cantley et al. 2016). Furthermore, the bacterium *Vibrio fischeri* produces a chondroitinase, EroS, capable of inducing sexual reproduction in *S. rosetta* (Woznica et al. 2017). Together, these results demonstrate the influence that bacteria can exert on the morphology, development, and evolution of microbial eukaryotes.

There is more direct evidence motivating study of *C. parapsilosis* functioning *in situ*. For instance, *Caenorhabditis elegans* model of polymicrobial infection experiments showed that *C. albicans* exhibits complex interactions with *Enterococcus faecalis*, a bacterial human gut commensal and opportunistic pathogen. In this context, *C. albicans* and *E. faecalis* negatively impact one another's virulence (Cruz et al. 2013), suggesting a mechanism that promotes commensal behavior in a gut microbial community context. The decrease in *C. albicans* virulence was attributed to inhibition of hyphal morphogenesis and biofilm formation by proteases secreted by *E. faecalis* (Cruz et al. 2013) as well as *E. faecalis* capsular polysaccharide (Bachtiar et al. 2016). No research has investigated *C. parapsilosis* in a microbial community context.

An alternative to studying Candida species in animal models or laboratory cultures is to use an untargeted shotgun sequencing approach (genome-resolved metagenomics). DNA is extracted from fecal or other samples and sequenced. The subsequent DNA sequences are assembled, and metagenome-assembled genomes (MAGs) are reconstructed. Much work of this type has focused on the bacterial members of the human microbiome; however, recently developed methods such as EukRep (West et al. 2018) enable reconstruction of eukaryotic genomes from metagenomes with greater consistency, including genomes of Candida species (Olm et al. 2019). The availability of genomes enables evolutionary studies and the application of other 'omics' approaches, such as transcriptomics, proteomics, and metabolomics, making it possible to go beyond metabolic potential to study activity *in situ*. Although there are limitations related to establishing causality via experimentation, the approaches can provide insights into metabolism and changes in metabolism linked to shifts in community composition in human-relevant settings.

Here, we applied shotgun metagenomics, metatranscriptomics, and metaproteomics to investigate the behavior and evolution of Candida in the premature infant gut and hospital environment. Novel de novo assembled *C. parapsilosis* and *C. albicans* genomes were reconstructed and the metagenomic data analyzed in terms of heterozygosity and population diversity. Due to the substantially less prior research on *C. parapsilosis* and the availability of *C. parapsilosis*-containing samples suitable for transcriptomics and proteomics, we focused our analyses on *C. parapsilosis* and identified genes and genomic regions under diversifying selection. Notably, we also identified instances of copy number gain of a gene involved in fluconazole resistance, pointing to a mechanism for hospital adaptation (Whaley et al. 2016). *C.*

parapsilosis in situ transcriptomic and proteomic profiles were clearly distinct from profiles reported previously from culture settings. Substantial shifts in *C. parapsilosis* expression occurred with changes in microbiome composition over a few day period, suggesting the strong influence of bacterial community composition on *C. parapsilosis* behavior.

3.3 Results

3.3.1 Recovery of novel Candida strain genomes

Fecal samples were collected from 161 premature infants primarily during the first 30 days of life (DOL) (full range of DOL 5–121), with an average of 7 samples per infant. Samples of the Neonatal Intensive Care Unit (NICU) were taken from six patient rooms within the hospital housing the infants (Magee-Womens Hospital of UPMC, Pittsburgh, PA, USA). Candida genomes were assembled from samples containing >2 Mbp of predicted eukaryotic DNA using a EukRep-based pipeline (West et al. 2018; see the "Methods" section for details). Three of the Candida genomes (Olm et al. 2019) and the bacterial component (Olm et al. 2019) were analyzed previously (see the "Methods" section). Eleven unique Candida genomes were assembled in total (Table 3.1), six *C. albicans* genomes and five *C. parapsilosis* genomes. All genomes have an estimated completeness >85% except for *C. parapsilosis* L2_023 and NYC subway, which had low coverage (4x and 6x respectively) in their samples. Nine genomes were reconstructed from premature infant fecal samples; one genome was derived from a NICU room sample S2_005. For comparison, we analyzed a Candida genome that we reconstructed from a publicly available metagenome read dataset from the New York City subway (NYC_subway; Afshinnekoo et al. 2015), as well as four previously published *C. parapsilosis* and fifty-one *C. albicans* isolate genomes.

3.3.2 Candida genomic variability

To characterize genomic variability in the strains of *C. albicans* and *C. parapsilosis* represented by metagenome-derived genomes, we identified single-nucleotide variants (SNVs) by mapping reads against completed reference genomes (strain SC5314 for *C. albicans* and CDC317 for C. parapsilosis). *C. albicans* genomes ranged from 3.2-9.9 heterozygous SNVs per kb (heterozygosity), whereas *C. parapsilosis* genomes ranged from 0.12-0.38 heterozygous SNVs per kb. Thus, we infer that, compared to *C. albicans*, *C. parapsilosis* displays very low variability in its diploid chromosome pair, which can be indicative of low genetic variability in the hospital environment and primarily asexual reproduction (Magwene et al. 2011).

Low heterozygosity in *C. parapsilosis* genomes has been reported for previously sequenced genomes (Pryszcz et al. 2013). Interestingly, *C. parapsilosis* genomes derived from our fecal metagenomes showed even lower overall heterozygosity than pure culture reference genomes (Figure S3.1). In general, this would not be expected because within-sample population diversity due to sampling of a microbial community should inflate measures of genomic heterozygosity. Thus, the lower genomic heterozygosity may be reflective of infants being initially colonized by essentially a single Candida genotype.

Because multiple new strains were sequenced from the same hospital, the phylogenetic relationships of new and previously sequenced strains from the same hospital were of interest

from the perspectives of the persistence of Candida populations in the hospital environment and transfer from room to human. To place the hospital and gut-associated sequences in context, we first compared those genomes to available reference genomes from NCBI using pair-wise average nucleotide identity (ANI) and by construction of single nucleotide variant (SNV) trees (Figure 1A, Figure S1-2). L2_023 was not included due to low sequencing coverage. *C. albicans* strains were spread throughout the tree of known *C. albicans* diversity (Figure S3.2) whereas *C. parapsilosis* strains from infant gut and NICU samples were clustered on a single branch (Figure 3.1A) separate from other reference hospital and environmental strains. Further, the two infant gut strains, sampled years apart, were nearly identical (99.99% identity). We verified this with whole genome alignments of the hospital and gut sequences (Figure S3.1-S3.2). We thus infer that the hospital room and gut *C. parapsilosis* strains are very closely related.

Based on analysis of population structure of seven *C. parapsilosis* genomes (Figure S3.3), we predicted six distinct *C. parapsilosis* ancestral populations. The exception is the fecal strain N3_182, which appears to be a recombinant admixture of the ancestral populations NICU strain S2_005 and the fecal strain C1_006. Given that N3_182 was sequenced four years before C1_006, both parental strains must have both existed in the hospital environment prior to hybridization. The findings provide evidence for a clearly defined, distinct hospital associated C. parapsilosis strains, a hybrid of which colonized a premature infant.

3.3.3 C. parapsilosis SNV hotspots as indicators of genes under selection

To investigate whether genomes sampled from the hospital could provide evidence of evolutionary adaptation to this environment, we visualized the spatial distribution of *C. parapsilosis* genomic diversity in the newly reconstructed genomes by mapping reads from each genome to a reference sequence (CDC317, recovered from a clinical sample) and calling SNVs. We plotted the density of SNVs in 1.3 kbp sliding windows across the genome of each strain (Figure 3.1B). Both heterozygous and homozygous SNVs are largely evenly distributed throughout the genome, with the exception of a few small regions with highly elevated SNV counts (regions of elevated diversity) that we refer to as SNV hotspots (Figure 3.1B).

Interestingly, SNV hotspots show a high level of conservation between all strains (Figure 3.1C). The one exception is reference strain GA1 cultured from human blood (Pryszcz et al. 2013), which shares only ~10% of its SNV hotspots with any other given strain. Notably, the NY subway strain is fairly similar to the clinical reference strain (few and minor hotspots) whereas our hospital sequences share SNV hotspots with both of the CBS strains (one from an olive and the other from skin), consistent with genomic similarity of the hospital and CBS strains in those regions.

To provide a more complete view of variation hotspots, we also mapped the reads from each population to the three other reference genomes (environmental strains CBS1984 and CBS6318, and the GA1 blood isolate, Figure S3.4). The number of SNV hotspots ranged from 16-45, and the regions were 5 kb to 24.5 kb in length. Due to the large size of the SNV hotspots, each hotspot overlaps a number of individual genes with SNVs spread both within and between genes (Figure 3.1D). In total, 376 genes are present within a SNV hotspot in at least one strain. No particular KEGG family or PFAM domain was significantly enriched in SNV hotspots. This, combined with the fact that SNVs are spread both within and between genes may be indicative of

SNV hotspots being recombination hotspots, or locations where additional SNVs hitchhike along with SNVs under selection.

3.3.4 Multicopy RTA3 gene

Another explanation for SNV hotspots could be due to gene copy number variation, as recent duplications of a region acquire mutations yet reads from these duplications map back to a single location. Overall, when windowed genomic coverage is plotted alongside SNV density (Figure 3.2A), this is clearly not the case. However, across the entire genome two regions of high coverage (Figure 3.2A), indicating high copy number variation, were identified and neither correspond to SNV hotspots. The first high copy number region contains an estimated 17-28 copies of the 18S, 25S, 5S, and 5.5S rRNA genes (Table S3.1, Figure 3.2B). The variation in rRNA copy number may indicate a range of maximum growth rates (Roller et al. 2016). The second region, which corresponds to the lipid translocase RTA3 gene and flanking sequence, is present in 9-16 copies (Table S3.1) in strains C1_006, N3_182, L2_023, and NYC_subway but not the four reference genomes or hospital room genome (Figure 3.2B). The high copy number RTA3 genes also have no detectable SNVs and different boundaries in each strain, suggesting the duplications were very recent and independent events in each strain.

3.3.5 In situ metatranscriptomics and metaproteomics

Given most work with Candida species is performed in pure culture or in murine models, little is known about their behavior in the human gut. We hypothesized performing metatranscriptomics and metaproteomics on infant fecal samples with *C. parapsilosis* would reveal unique transcriptomic and proteomic profiles, indicative of differences in metabolism and behavior between culture and *in situ* settings. Two prospective infants were identified, infant 06 with a documented Candida blood infection (Figure 3.3) and infant 74 with a documented Candida lung infection. Both infants were treated with fluconazole shortly after detection of Candida infection (Figure 3.3, Table S3.2). Metagenomic, metatranscriptomic, and metaproteomic datasets were generated from fecal samples at five to six timepoints for each infant. In infant 74, no Candida species were detected in the generated datasets (Figure S3.4). However, in infant 06, metagenomic sequencing confirmed the presence of *C. parapsilosis* (strain C1_006) in the fecal samples. De novo gene prediction was performed on the metagenome-derived *C. parapsilosis* genome and the resulting gene models were used for mapping transcriptomic reads and proteomic peptides (Figure 3.3).

In addition to *C. parapsilosis*, genomes were recovered for three bacterial species in infant 06: *Enterococcus faecalis*, *Lactobacillus gasseri*, and *Staphylococcus epidermidis*. Interestingly, in every infant where a Candida genome was assembled or detected through read mapping, *E. faecali*s was also present (N=7). *C. parapsilosis* is highly abundant in the first 20 days of life before quickly being replaced or outnumbered, largely by *E. faecalis*. Similar abundance patterns have been observed previously for microbial eukaryotes in neonatal fecal samples (Olm et al. 2019). *C. parapsilosis* transcriptomic abundance shows a similar pattern to the DNA abundance but transcription remains detectable at later time points (Figure 3.3). In contrast, *C. parapsilosis* proteomic abundance remained relatively stable over all timepoints.

3.3.6 C. parapsilosis expression in situ vs. culture settings

Given most work with *C. parapsilosis* has been performed on pure cultures, we wondered if there are differences in behavior and metabolism that would be detectable by comparing transcriptomic datasets. For comparison, we downloaded raw sequencing reads from publicly available *C. parapsilosis* RNAseq experiments (Guida et al. 2011; Pryszcz et al. 2013), including datasets from multiple strains (CDC317, CBS1954, and CBS6318) and varying culture conditions, including different media, growth temperatures, and oxygen concentrations. A hierarchical clustering of expression of CDC317 transcripts reveals a clearly distinct transcriptomic profile between *in situ* and all culture settings (Figure 3.4A). Notably, *in situ* samples are also extremely variable; clustering as far apart from one another as from the culture samples (Figure 3.4A). We quantitatively identified differentially expressed transcripts between culture and *in situ* settings and found that 53% of transcripts were significantly differentially expressed; 23% up *in situ*, 30% down (Figure 3.4B), further highlighting the stark differences between *in situ* and culture settings.

In situ and culture transcriptome samples were differentiable in a principal component analysis (PCA), paralleling the hierarchical clustering of *C. parapsilosis* transcriptomes (Figure 3.5A). We performed a sparse Partial Least Squares Discriminant Analysis (sPLS-DA), treating each transcript as a variable, to try and identify important features able to discriminate between *in situ* and culture in a multivariate space (Figure 3.5B, Figure S3.5, Table S3.3). Important features were enriched for mitochondrial and aerobic respiration genes (9/50), uncharacterized genes (15/50), and a subset of ribosomal proteins (8/50; p=2.3x10-7).

Biofilm formation is an important virulence factor for Candida species; often contributing to the development of systemic infections (Nobile et al. 2012; Nobile et al. 2015). We were interested in whether the expression of virulence factors was enriched *in situ*, given the samples were obtained from an infant with a documented Candida blood infection. We obtained a list of well characterized biofilm formation genes from *C. albicans* (Nobile et al. 2015), identified orthologs in *C. parapsilosis* and compared their expression *in situ* to culture settings. Biofilm formation showed lower expression overall *in situ* (Figure 3.4C).

We were curious to see if the multicopy RTA3 gene in infant strain C1_006 (Figure 3.2B) showed increased expression as compared to the single copy RTA3 gene in reference strain CDC317. Indeed, the expression of the RTA3 in strain C1_006 is significantly higher (Figure 3.2C), suggesting a role of this gene duplication as a way to increase overall expression of RTA3. Interestingly, we did not see an increase in expression following fluconazole treatment (Figure S3.6), indicating RTA3 expression may be consistently higher in C1_006. However, it is worth noting we were unable to obtain samples until seven days after fluconazole treatment and any treatment effect on expression may have already passed.

3.3.7 C. parapsilosis impact on bacterial expression

E. faecalis, S. epidermidis and *L. gasseri* bacteria in infant 06 had transcripts sequenced at high depths at multiple time points (Figure 3.3) so it was possible to investigate whether the presence or absence of *C. parapsilosis* had a distinguishable effect on their transcriptomic profiles. We compared bacterial transcription in these samples to transcription patterns of bacteria in the absence of Candida using previously reported datasets (21 samples for *E. faecalis* and 20 samples

for *S. epidermidis*; Sher et al. 2020). The analysis was not possible for *L. gasseri* as this bacterium was not present in any of the metatranscriptomes used for comparison. The transcriptomes of both *E. faecalis* and *S. epidermidis* were distinguishable between the presence and absence of *C. parapsilosis*, and this effect appears to be independent of infant of origin and thus the bacterial strain variant type (Figure 3.5C-D). This result suggests *C. parapsilosis* has a large impact on the behavior and metabolism of other gut community members. In addition, the expression of *E. faecalis* genes previously shown to negatively impact *C. albicans* virulence (Cruz et al. 2013) showed no significant difference in expression between *C. parapsilosis* negative and positive samples.

Important features identified from a sPLS-DA on Candida-positive vs. Candida-negative samples included a subset of *E. faecalis* ribosomal proteins (Table S3.3, Figure S3.5). Additionally, ribosomal proteins all showed higher expression *in situ*, suggesting increased *E. faecalis* growth rate in the presence of *C. parapsilosis*. Other important features included mannitol specific phosphotransferase system (PTS) transporters upregulated in Candida-positive samples and downregulated mannose specific PTS transporters (Table S3.3). Furthermore, Mannitol-1-phosphate 5-dehydrogenase, an enzyme responsible for the conversion of D-mannitol to fructose, was upregulated in Candida-positive samples, indicating an increased capacity for degradation of mannitol in addition to import (Table S3.3). Important features in *S. epidermidis* were less clear, but again included a subset of ribosomal proteins as well as beta-lactamases, both with increased expression *in situ* (Table S3.3).

3.3.8 Transcriptomics enriched gene functions

Given the large differences in transcriptomes between culture and *in situ*, we looked for functions enriched in either setting (Figure 3.6, Table S3.4). DESeq2 identified groups of differentially expressed genes that were too large to be informative, so more restrictive cutoffs were used. Up *in situ* was defined as having >3 log2 expression *in situ* whereas down *in situ* was defined as <-3 log2 expression in situ. Up *in situ* was enriched for KEGG families for LSM 2-8 and 1-7 complexes, a family of proteins involved in mRNA metabolism highly conserved in eukaryotes (Beggs et al. 2005), as well as Cytochrome c oxidase and bc1 complex and proteins without an annotated KEGG family (Figure 3.6, Table S3.4). Down *in situ* is enriched for helicase and polysaccharide synthase PFAM domains. Additionally, proteins without an annotated KEGG family (unknown function) were enriched in both groups (Table S3.4).

3.3.9 Proteomics

As noted above, the metaproteomic abundances for infant 06 were relatively stable over time, with evidence of Candida core metabolic activities such as glycolysis, which indicate stability of this organism within the gut environment (Figure 3.3C). Comparison of the Candida *in situ* proteomics data with data from pure culture experiments was not possible as no pure culture proteomics datasets suitable for comparison have been published. Using the metagenome-derived *C. parapsilosis* genome as reference, we identified the most abundant proteins and found that this subset included ribosomal and F-Type ATPase proteins (Figure 3.6) and was significantly enriched for HSP70 and actin PFAM domains (Table S4). Also, among proteins found with the most peptide evidence were proteins related to protection of the organism from oxidative stress, such as superoxide dismutase. The high abundance protein set included some of the genes contained within SNV hotspots but there was no significant association. We also examined the most abundant

proteins in the bacterial species. In *E. faecalis* and *S. epidermidis*, Lac genes were some of the most abundant suggesting lactose may be an important substrate for these community members. Finally, among human proteins detected, there was ample evidence of neutrophil degranulation, which indicates an active host immune response. Neutrophils use oxidative mechanisms to promote fungal clearance (Desai et al., 2018), which suggests Candida is employing oxidative protection in response to this host defense mechanism.

3.4 Discussion

Fungal pathogens are known to have hospital reservoirs. For example, the water supply system of a paediatric institute was shown to be a reservoir for *Fusarium solani* (Mesquita-Rocha et al., 2013). A NICU outbreak of *Malassezia pachydermatis* was linked to the dog of a healthcare worker (Chang et al., 1998), although persistence via long-term carriage by a healthcare worker vs. continual passage between infants and rooms (or a combination of these) could not be resolved. However, much remains to be learned about where reservoirs of hospital-associated fungi are and how long strains persist in them. In contrast to previous studies of *C. parapsilosis* utilizing pure culture and model systems, we applied genome-resolved metagenomics, metatranscriptomics, and metaproteomics to study *C. parapsilosis* in the context of the infant gut and hospital rooms of a neonatal intensive care unit. We detected novel, near identical *C. parapsilosis* genomes sequenced years apart in separate infants, suggesting transmission of members of a fungal population from reservoir to infant or infant to reservoir to infant. It is worth noting that although the strains are near-identical, the multicopy RTA3 locus in each strain had different boundaries and different copy numbers. This observation suggests that these two strains are very closely related members of a somewhat more diverse hospital adapted population.

Population genomic analyses of reconstructed genomes revealed multiple, independent instances of copy number gain of the RTA3 gene. RTA3, a lipid translocase, has been implicated in resistance to azole class antifungal drugs such as fluconazole in *C. albicans* (Whaley et al. 2016). The RTA3 gene is frequently overexpressed in resistant isolates and increased expression of RTA3 increases resistance to fluconazole whereas deletion of the RTA3 gene results in increased azole susceptibility (Whaley et al. 2016). Copy gain of this gene in *C. parapsilosis* strains may represent a mechanism for rapid adaptation to fluconazole, the most widely used antifungal in most hospitals (Whaley et al. 2016), as a means by which to increase its expression and thus its resistance. Similar gene copy number gains have been reported for the human amylase gene, hypothesized to be in response to increases in starch consumption (Pajic et al. 2019). Indeed RTA3 expression *in situ* from strain C1_006, which has RTA3 in multicopy, was significantly increased as compared to single copy strain CDC317 in culture (Guida et al. 2011; Figure 2C). The high likelihood that the copy number gain occurred independently in multiple strains suggests selection for this particular genomic feature. Identifying mechanisms of antifungal resistance is of particular importance given 3-5% of *C. parapsilosis* strains are already resistant to fluconazole (Whaley et al. 2017) and our relative inability to deal with infections of drug-resistant fungi.

Examining the genomic distribution of SNVs within the genomes of each *C. parapsilosis* strain revealed the presence of SNV hotspots. Many of these SNV hotspots are shared between strains,

some of which are specific to the hospital and infant gut strains. Unlike *C. albicans, C. parapsilosis* is not an obligate commensal of mammals (Trofa et al. 2008). Consequently, some regions of the *C. parapsilosis* genome may be under selection for adaptation to the hospital, in addition to the gut environment. Further supporting the idea that some genomic innovation is associated with adaptation to the built environment, the NICU strain clustered the most closely to the NYC subway strain based on SNV hotspot overlap (Figure 3.1C). These two strains are geographically and phylogenetically distinct but the shared regions of diversification may be related to their common need to adapt to the built environment.

Metatranscriptomics of infant fecal samples revealed *C. parapsilosis* transcriptomes that are both highly variable and distinct from those of culture samples. Interestingly, the degree of variance exhibited by transcriptomes of the same population in the same infant over a few day period was greater than that observed between *C. albicans* white and opaque phenotypes (Figure S7; Tuch et al. 2010). The *C. albicans* white and opaque phenotypes differ in their appearance (Slutsky et al. 1987), mating style (Miller et al. 2002), and environmental conditions they are best adapted to (Ramirez-Zavala et al. 2008, Huang et al. 2009), and represent two exceptionally distinct Candida phenotypes. The high variability in *C. parapsilosis* is likely the result of changing conditions presented in the gut, including microbial community composition as well as the developing physiology of the host. Varying stages of infection and/or response to antifungal treatment may also have had an effect, but more dense time-series and additional infants would be required to elucidate these effects.

In contrast to the large changes in *C. parapsilosis* RNA and DNA relative abundances over time, *C. parapsilosis* peptide relative abundance remained stable over the study period. It is not uncommon to see different signals from transcripts and proteins (Haider et al. 2013), in part because proteins can persist for relatively long periods of time compared to transcripts. The most abundant proteins in the proteomics dataset have a HSP70 domain found in heat shock proteins (HSP). In *C. albicans*, HSP have been documented to help control virulence by interacting with regulatory systems, and to enable drug resistance (Gong et al. 2017).

The presence of *C. parapsilosis* within infant gut samples may impact the transcriptomes of bacterial gut community members. Important features for separating Candida-positive and Candida-negative samples included a suite of upregulated mannitol transporters and downregulated mannose transporters in *E. faecalis* (Table S3). *C. parapsilosis* strain SK26.001 is documented as producing mannitol (Meng et al. 2017) and mannose, in the form of the polysaccharide mannan, which can be an important component of extracellular polysaccharides produced by Candida (Dominguez et al. 2019). Interestingly, a characteristic of *E. faecalis* is its ability to grow by fermenting mannitol (Quiloan et al. 2012). Given the potential for interaction between *E. faecalis* and *C. parapsilosis*, its possible the disappearance of *C. parapsilosis* induced a substrate switch in *E. faecalis*.

Interestingly, statistical tests detected a subset of ribosomal proteins as important features for separating transcriptome patterns of *C. parapsilosis in situ* from those reported from culture studies, as well as for separating Candida-positive from Candida-negative samples for both *E. faecalis* and *S. epidermidis* (Table S3.3). In recent years, ribosomal heterogeneity, in which ribosomal protein subunits are swapped out or missing from individual ribosomes, has gained

traction as a way for organisms to regulate translation (Guimaraes et al. 2016, Shi et al., 2017, Genuth et al. 2018). Ribosomal heterogeneity may be being utilized as an additional regulatory measure to adapt to the rapidly changing gut microbial context. Alternatively, fluctuations in ribosomal subunit abundance could be to maintain ribosomal homeostasis (Cruz et al. 2017), or individual ribosomal subunits could be performing functions unrelated to protein synthesis (Zhou et al. 2015).

Biofilm formation is an important virulence factor of Candida infections (Cavalheiro et al. 2018). Infant 06 had a documented Candida blood infection, and such infections are commonly systemic (Mavor et al. 2005). Interestingly, despite infection, Candida biofilm formation genes were relatively less expressed *in situ* in the gut of Infant 06 as compared to expression levels previously reported over a range of culture conditions. Similarly, genes with a PFAM domain for polysaccharide synthase, genes potentially important for the generation of Candida biofilm matrices (Dominguez et al. 2019), were less expressed *in situ* than in cultures. Thus, biofilm formation may not be an important component of every infection.

Genes linked to oxygen utilization, such as cytochrome c oxidase subunits, were more highly expressed *in situ* than over the range of culture conditions, suggesting growth in a relatively aerobic environment. This may be reflective of the higher oxygen levels in the gut during early life (Chong et al. 2018).

The prevalence of transcripts of uncharacterized genes in the *in situ* transcriptomes (Figure 3.5B; Table S3.3) is particularly interesting. *C. parapsilosis* and other Candida species are rarely studied in a microbial community context, leaving gaps in understanding of genes required for organism-organism interactions. We suspect that some of the highly expressed genes are important for Candida interactions with bacteria and other community members. Thus, they represent important targets for future co-culture-based investigations.

3.5 Conclusions

We applied genome-resolved metagenomics, metatranscriptomics, and metaproteomics to recover genomes for, and study the behavior of, *C. parapsilosis in situ*. We showed *C. parapsilosis* has a highly distinct transcriptomic profile *in situ* vs in culture. Further, the extreme variability in the *in situ* transcriptome data indicates the considerable effect the gut microbial community and human host may have on *C. parapsilosis* behavior and metabolism. Overall, these results demonstrate that *in situ* study of *C. parapsilosis* and other Candida species is not only possible but necessary for a more holistic understanding of their biology.

3.6 Methods

3.6.1 Metagenomic sampling and sequencing

This study made use of previously published infant datasets: NIH1 (Brown et al. 2018), NIH2 (Brooks et al. 2017), NIH3 (Raveh-Sadka et al. 2015), NIH4 (Rahman et al. 2018), Sloan2 (Brooks et al. 2017), and SP_CRL (Sharon et al. 2013), as well as several new datasets including multiple timepoints from infant 06 and infant 74, and samples L2_023, S3_003, and S3_016.

For newly generated metagenomic sequencing from infant 06 and infant 74, total genomic DNA and total RNA were extracted from fecal samples using Qiagen's AllPrep PowerFecal DNA/RNA kit (Qiagen) and subsequently split into DNA and RNA portions. The aliquot used for metagenomic sample preparation was treated with RNase A. DNA quality and concentration were verified with Qubit (Thermofisher) and Fragment Analyzer (Agilent). Illumina libraries with an average insert size of 300 bps were constructed from purified genomic DNA using the Nextera XT kit (Illumina) and sequenced on Illumina's NovaSeq platform in a paired end 140 bp read configuration, resulting in at least 130 million paired end reads from each library.

NICU metagenomic sampling was described and published previously (Brooks et al. 2017). All samples were collected from the same NICU at UPMC Magee-Womens Hospital (Pittsburgh, PA). In order to generate enough DNA for metagenomic sequencing, DNA was collected from multiple sites in the NICU and combined into three separate pools for sequencing. Highly-touched surfaces included samples originating from the isolette handrail, isolette knobs, nurses hands, in-room phone, chair armrest, computer mouse, computer monitor, and computer keyboard. Sink samples included samples from the bottom of the sink basin and drain. Counters and floors consisted of the room floor and surface of the isolette. See previous publication for details (Brooks et al. 2017; Brooks et al. 2018).

3.6.2 Eukaryotic genome binning and gene prediction

For each sample, sequencing reads were assembled independently with IDBA-UD (Peng et al. 2012). Additionally, for each infant, reads from every time point were concatenated together. A co-assembly was then performed on the pooled reads for each infant with IDBA-UD in order to assemble sequences from low abundance organisms. The Eukaryotic porton of each sample assembly was predicted with EukRep (West et al. 2018) and putative eukaryotic bins were generated by running CONCOCT (Alneberg et al. 2014) with default settings on the output of EukRep. To reduce computational load, resulting eukaryotic bins shorter than 2.5 mbp in length were not included in further analyses. GeneMark-ES (Ter-Hovhannisyan et al. 2008) and AUGUSTUS (Stanke et al. 2006) trained with BUSCO (Simão et al. 2015) were used to perform gene prediction on each bin using the MAKER (Cantarel et al. 2008) pipeline. In addition, a second homology-based gene prediction step was performed. Each bin was identified as either C. parapsilosis or C. albicans and reference gene sets from C. parapsilosis CDC317 and C. albicans SC5314 were used for homology evidence respectively in a second-pass gene prediction step with AUGUSTUS (Stanke et al. 2006), as implemented in MAKER (Cantarel et al. 2008).

3.6.3 SNV calling and detection of SNV hotspots

In order to call variants in each Candida genome, reads from the sample in which a particular genome was binned from, or the publicly available reads from SRA, were mapped back to the de novo assembled genome using Bowtie 2 (Langmead et al. 2012) with default parameters. The PicardTool (http://broadinstitute.github.io/picard/) functions “SortSam” and “MarkDuplicates” were used to sort the resulting sam file and remove duplicate reads. FreeBayes (Garrison et al. 2012) was used to perform variant calling with the options “--pooled-continuous -F 0.01 -C 1.” Variants were filtered downstream to include only those with support of at least 10% of total

mapped reads in order to avoid false positives. SNV read counts were calculated using the "AO" and "RO" fields in the FreeBayes vcf output file.

SNV density was visualized across the CDC317 reference genome using a custom python script. SNV hotspots were quantitatively defined with 5 kbp windows with a slide of 500 bp across the genome, flagging windows with a SNV density at least three standard deviations above the genomic average SNV density, and merging overlapping flagged windows. Genes located within SNV hotspots as well as overlapping SNV hotspots between strains were identified with intersectBed (Quinlan et al. 2010).

3.6.4 Candida phylogenetics and population structure

For generation of a SNP tree for both *C. parapsilosis* and *C. albicans*, all publicly available genomic sequencing reads for both species were downloaded from NCBI's short read archive (SRA), including isolate *C. parapsilosis* read sets and *C. albicans* sets. SNVs were called for each isolate read set using the same pipeline used for metagenome-derived genomes, as described above. A SNP tree was generated for *C. parapsilosis* and *C. albicans* using SNPhylo (Lee et al. 2014) with settings '-r -M 0.5 -l 2' and '-r -M 0.5 -l 0.8' respectively and visualized using FigTree (https://github.com/rambaut/figtree/). For genomic average nucleotide identity (ANI) comparisons, *C. parapsilosis* and *C. albicans* reference genomes were downloaded from NCBI. Subsequently, dRep (Olm et al. 2017) in the 'compare_wf' setting was used to generate ANI comparisons for each genome. For inferring *C. parapsilosis* population structure, FreeBayes vcf files were converted to PLINK bed format with PLINK (Pucell et al. 2007) and used as input for ADMIXTURE (Alexander et al. 2011). The predicted number of ancestral populations, K, was selected using ADMIXTURE's cross-validation procedure for values 1-8.

3.6.5 Detection of copy number variation

Genomic copy number variation within the *C. parapsilosis* strains was searched for by mapping reads from the sample the genome was derived from to the *C. parapsilosis* CDC317 reference genome. Windowed coverage was then calculated across the genome in 100bp sliding windows using pipeCoverage (https://github.com/MrOlm/pipeCoverage) and visualized with Integrated Genomics Viewer (IGV) (Robinson et al. 2017). Copy numbers for multicopy regions were estimated by dividing the average coverage of the windows located within the multicopy region by the average genomic coverage.

3.6.6 Transcriptomic sequencing and analysis

Total RNA was extracted from fecal samples using the AllPrep PowerFecal DNA/RNA kit (Qiagen) and subsequently treated with DNase. Purified RNA quality and concentration were measured using the Fragment Analyzer (Agilent). Illumina sequencing libraries were constructed with the ScriptSeq Complete Gold Kit (Illumina) without performing the rRNA removal step, resulting in library molecules with an average insert size of 150 bp. Sequencing was performed on Illumina's NextSeq platform in a paired end 75 bp configuration, resulting in an average of 54 million paired end reads per sample.

Transcriptomic reads from studies (Guida et al. 2011, Pryszcz et al. 2013) were downloaded from the SRA. Transcriptomic reads from each dataset were then mapped to *C. parapsilosis*

reference strain CDC317 gene models with Kallisto (Bray et al. 2016) and transcript per million (TPM) values were used to compare expression levels across samples. Differentially expressed transcripts were identified using raw read counts with the R package DESeq2 (Love et al. 2014). Rlog transformation was applied to transcript read counts from each sample prior to generation of transcriptome PCAs. PCA plots were generated with DESeq2. Important features for separating *C. parapsilosis in situ* and culture as well as *E. faecalis* and *S. epidermidis* Candida-positive and Candida-negative samples were identified through the use of a sparse Partial Least Squares Discriminant Analysis (sPLS-DA) as implemented in the MixOmics package (Rohart et al. 2017) on rlog transformed transcript read counts. MixOmics cross-validation (tune.splsda) was used with settings fold = 3 and nrepeat = 50 to estimate the optimal number of components (features) for separating each pair of sample types.

Genes were annotated with KEGG KOs and PFAM domains using HMMER with KOfam (Aramaki et al. (2019) and Pfam-A (El-Gebali et al. 2019) HMM databases. Subsets of genes of interest (described in results) were then searched for significantly enriched KEGG families or PFAM domains with a hypergeometric distribution test as part of the R 'stats' package (R Core Team, 2013).

3.6.7 Generation of Proteomic Datasets

Lysates were prepared from ~50mg of fecal material by bead beating in SDS buffer (4% SDS, 100 mM Tris-HCl, pH 8.0) using 0.15-mm diameter zirconium oxide beads. Cell debris was cleared by centrifugation (21,000 x g for 10 min). Pre-cleared protein lysates were adjusted to 25mM dithiothreitol and incubated at 85°C for 10 min to further denature proteins and to reduce disulfide bonds. Cysteine residues were alkylated with 75 mM iodoacetamide, followed by a 20-minute incubation at room temperature in the dark. After incubation, proteins were isolated by chloroform-methanol extraction. Protein pellets were washed with methanol, air-dried, and resolubilized in 4% sodium deoxycholate (SDC) in 100 mM ammonium bicarbonate (ABC) buffer, pH 8.0. Protein samples were quantified by BCA assay (Pierce) and transferred to a 10 kDa MWCO spin filter (Vivaspin 500; Sartorius) before centrifugation at 12,000 x g to collect denatured and reduced proteins atop the filter membrane. The concentrated proteins were washed with 100 mM ABC (2x the initial sample volume) followed by centrifugation. Proteins were resuspended in a 1x volume of ABC before proteolytic digestion. Protein samples were digested *in situ* using sequencing-grade trypsin (G-Biosciences) at a 1:75 (wt/wt) ratio and incubated at 37°C overnight. Samples were diluted with a 1x volume of 100 mM ABC, supplied with another 1:75 (wt/wt) aliquot of trypsin, and incubated at 37°C for an additional 3 hours. Tryptic peptides were then spin-filtered through the MWCO membrane and acidified to 1% formic acid to precipitate the residual SDC. The SDC precipitate was removed from the peptide solution with water-saturated ethyl acetate extraction. Samples were concentrated via SpeedVac (Thermo Fisher), and peptides were quantified by BCA assay (Pierce) before LC-MS/MS analysis.

12ug of each peptide sample was analyzed by automated 2D LC-MS/MS using a Vanquish UHPLC with autosampler plumbed directly in-line with a Q Exactive Plus mass spectrometer (Thermo Scientific). A 100 µm inner diameter (ID) triphasic back column [RP-SCX-RP; reversed-phase (5 µm Kinetex C18) and strong-cation exchange (5 µm Luna SCX) chromatographic resins; Phenomenex] was coupled to an in-house pulled, 75 µm ID nanospray emitter packed with 30 cm Kinetex C18 resin. Peptides were autoloaded, desalted, separated, and

analyzed across four successive salt cuts of ammonium acetate (35, 50, 100, and 500 mM), each followed by a 105-minute organic gradient. Mass spectra were acquired in a data-dependent mode with the following parameters: a mass range of 400 to 1,500 m/z; MS and MS/MS resolution of 35K and 17.5K, respectively; isolation window = 2.2 m/z with a 0.5m/z isolation offset; unassigned charges and charge states of +1, + 5, +6, +7 and +8 were excluded; dynamic exclusion was enabled with a mass exclusion window of 10 ppm and an exclusion duration of 45seconds.

MS/MS spectra were searched against custom-built databases composed of the concatenated sequenced metagenome derived predicted proteomes from all time-points, the human reference proteome from UniProt, common protein contaminants, and reversed-decoy sequences using Proteome Discover 2.2 (Thermo Scientific), employing the CharmeRT workflow (Dorfer et al., 2018). Peptide spectrum matches (PSMs) were required to be fully tryptic with two miscleavages, a static modification of 57.0214 Da on cysteine (carbamidomethylated) residues, and a dynamic modification of 15.9949 Da on methionine (oxidized) residues. False-discovery rates (FDRs), as assessed by matches to decoy sequences, were initially controlled at 1% at the peptide level. To alleviate the ambiguity associated with shared peptides, proteins were clustered into protein groups by 100% identity for microbial proteins and 90% amino acid sequence identity for human proteins using USEARCH (Edgar et al., 2010). FDR-controlled peptides were then quantified according to the chromatographic area under the curve (AUC) and mapped to their respective proteins. Peptide intensities were summed to estimate protein-level abundance based on peptides that uniquely mapped to one protein group. Protein abundance distributions were then normalized across samples using InfernoRDN (Polpitiya et al, 2008), and missing values were imputed to simulate the mass spectrometer's limit of detection using Perseus (Tyanova et al., 2016).

3.7 Figures

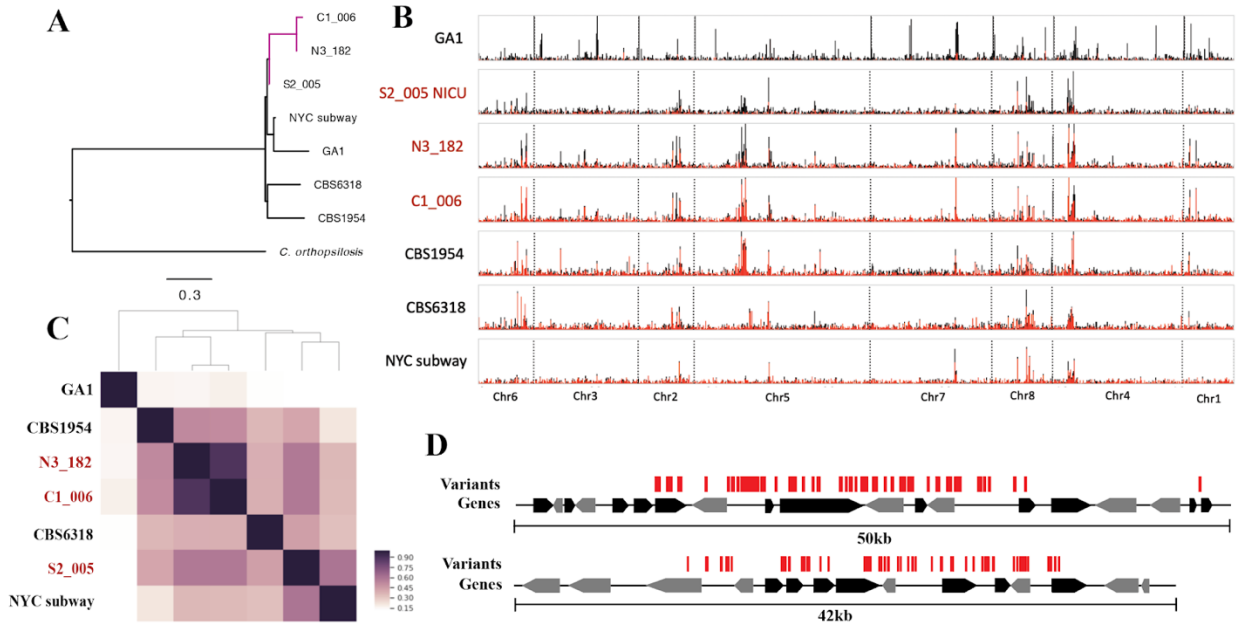

Figure 3.1. Analysis of *C. parapsilosis* genomic variability reveals a potential hospital associated population and the presence of SNV hotspots. (A) A phylogenetic tree of *C. parapsilosis* strains constructed from concatenated SNVs. Metagenome derived hospital strains from this study demarcated as the purple clade. ANI comparisons and a *C. albicans* SNV tree are also available in Figures S1-S2. (B) Whole genome SNV density plots for each *C. parapsilosis* strain. Strain names in red are strains assembled from samples from infants or the NICU from Magee-Women's Hospital. SNV density plotted in 1.3kb sliding windows. Window size was selected based on ease of visualization. Chromosomes are separated with dashed lines. Total bar height represents total SNV density and homozygous SNV proportion is labeled in red whereas heterozygous is black. (C) Depiction of SNV hotspot overlap between each strain. Pairwise overlap was calculated between each strain and plotted. Strain names in red are strains assembled from samples from infants or the NICU from Magee-Women's Hospital. (D) Two example SNV hotspots. Individual SNVs are represented with red bars.

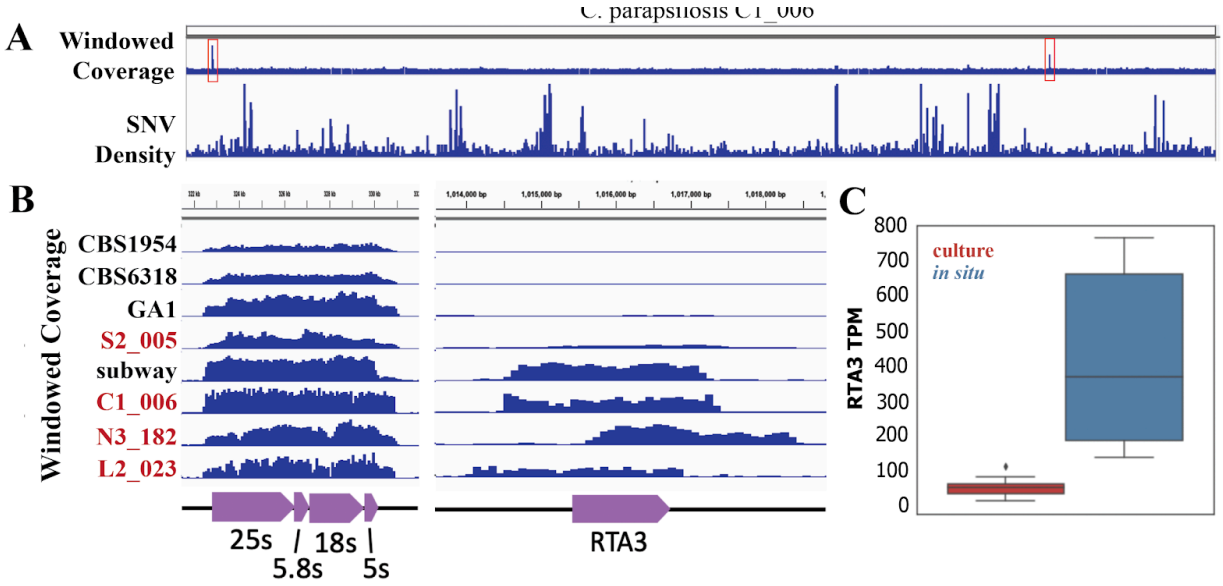

Figure 3.2. *C. parapsilosis* strains have high copy number rRNA and RTA3 loci. (A) Whole genome windowed coverage of SNP density for *C. parapsilosis* strain C1_006. High copy number regions of interest are highlighted with red boxes. (B) An expanded view of highlighted high copy number regions from (A). Windowed coverage is plotted as 100bp sliding windows. Metagenome-derived hospital strains from this study labeled in red. (C) Boxplot of expression of the RTA3 gene from multicopy strain C1_006 *in situ* (red) and strain CDC317 in culture (blue). Expression represented as Transcripts Per Million (TPM).

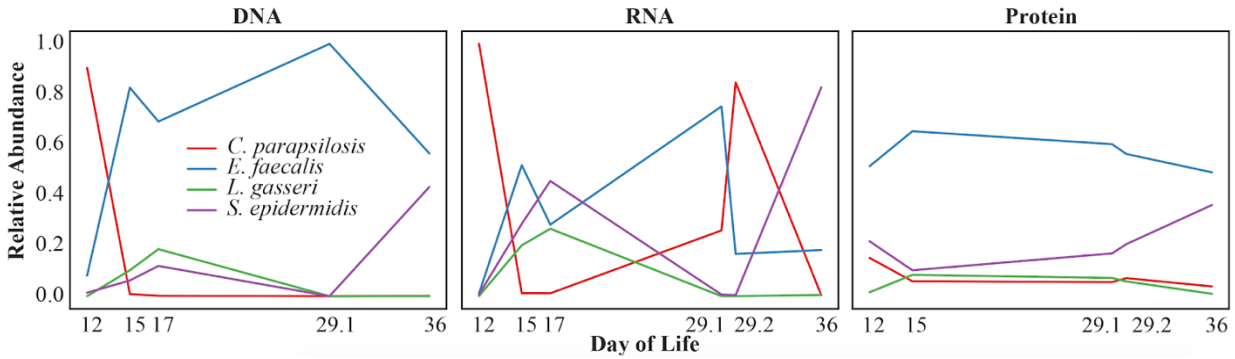

Figure 3.3. *In situ* metagenomics metatranscriptomics, and metaproteomics of infant 06. Plotted are the relative DNA, RNA, and peptide abundances for each detected organism after human removal. Plotted on the x axis are the Days Of Life (DOL) samples were taken.

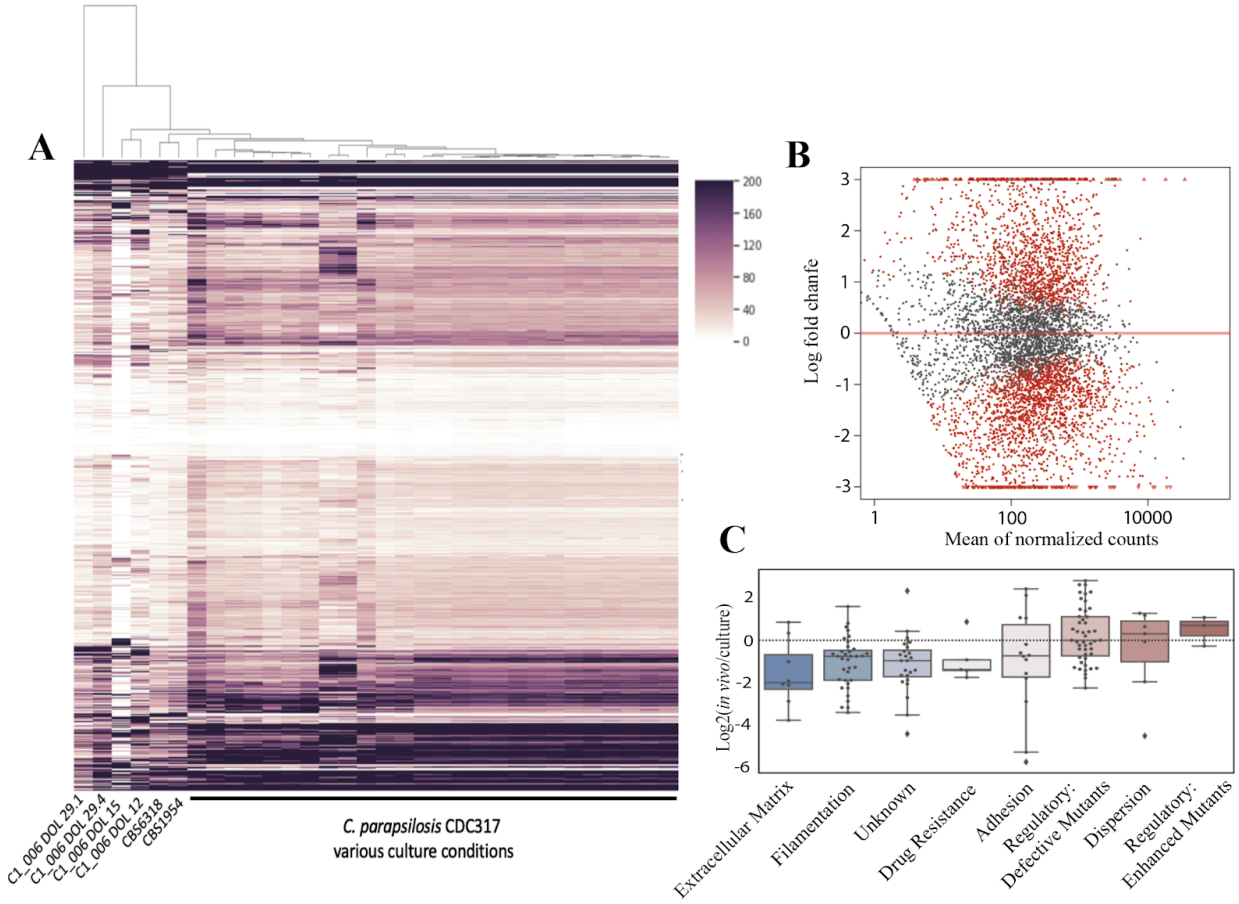

Figure 3.4. *C. parapsilosis* displays distinct and highly variable *in situ* transcriptomic profiles. (A) Hierarchical clustering of *C. parapsilosis* TPM values for C1_006 in *in situ* samples and pure culture samples under a variety of conditions. (B) Average log2 fold change *in situ* vs culture plotted against the mean of normalized counts for each transcript. Transcripts in red were identified as being significantly differentially expressed by DESeq2. (C) Boxplots of expression of categories of genes involved in biofilm formation. Regulatory defective mutants refers to regulatory genes that inhibited biofilm formation when mutated.

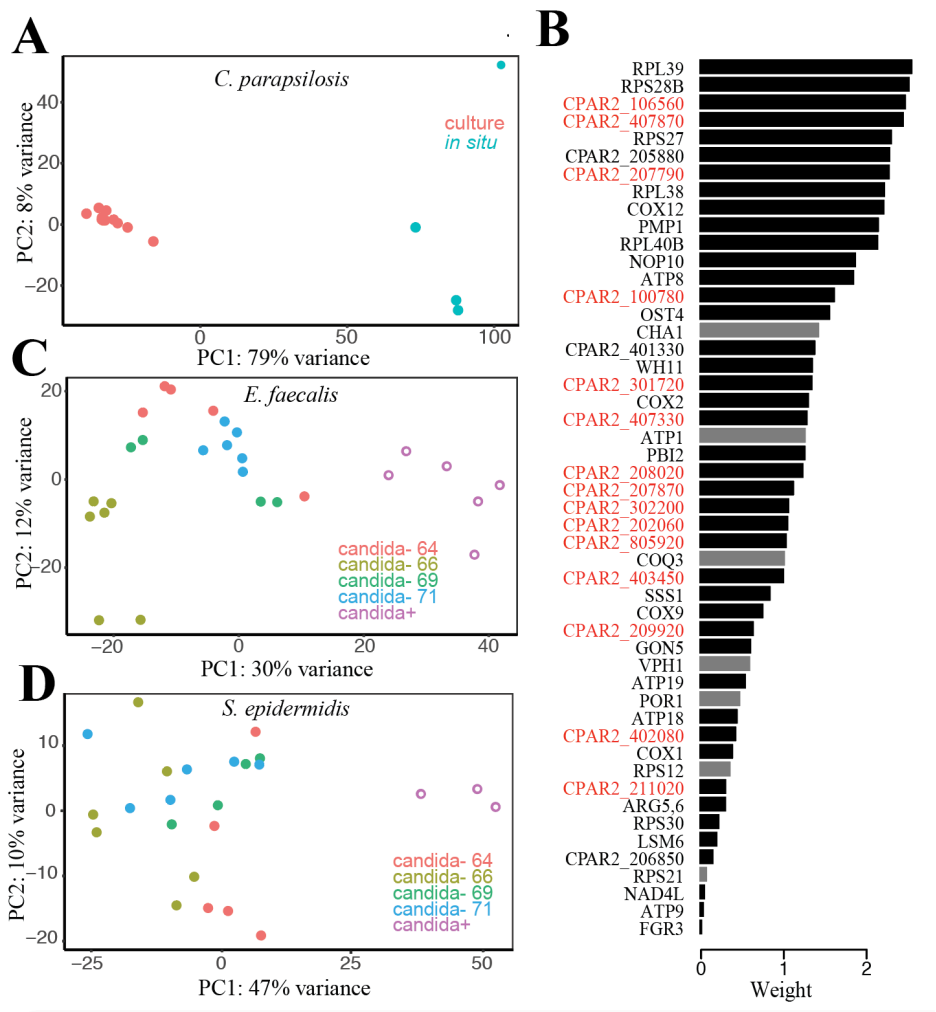

Figure 3.5. Presence of *C. parapsilosis* affects bacterial community member's expression. (A) PCA of C. parapsilosis *in situ* and pure culture transcriptomes. (B) Depiction of features identified by sPLS-DA for separating *C. parapsilosis in situ* and pure culture transcriptomes. Plotted are the feature weights. Black bars are genes that exhibited higher expression on average *in situ* whereas grey had higher average expression in culture. Genes labeled in red correspond to proteins of unknown function. (C-D) PCAs of *E. faecalis* (C) and *S. epidermidis* (D) transcriptomes from infant microbiomes both with and without detected *C. parapsilosis*. Candida-negative transcriptomes were from four different infants (published previously; Sher et al. 2020) denoted as 64, 66, 69, and 71.

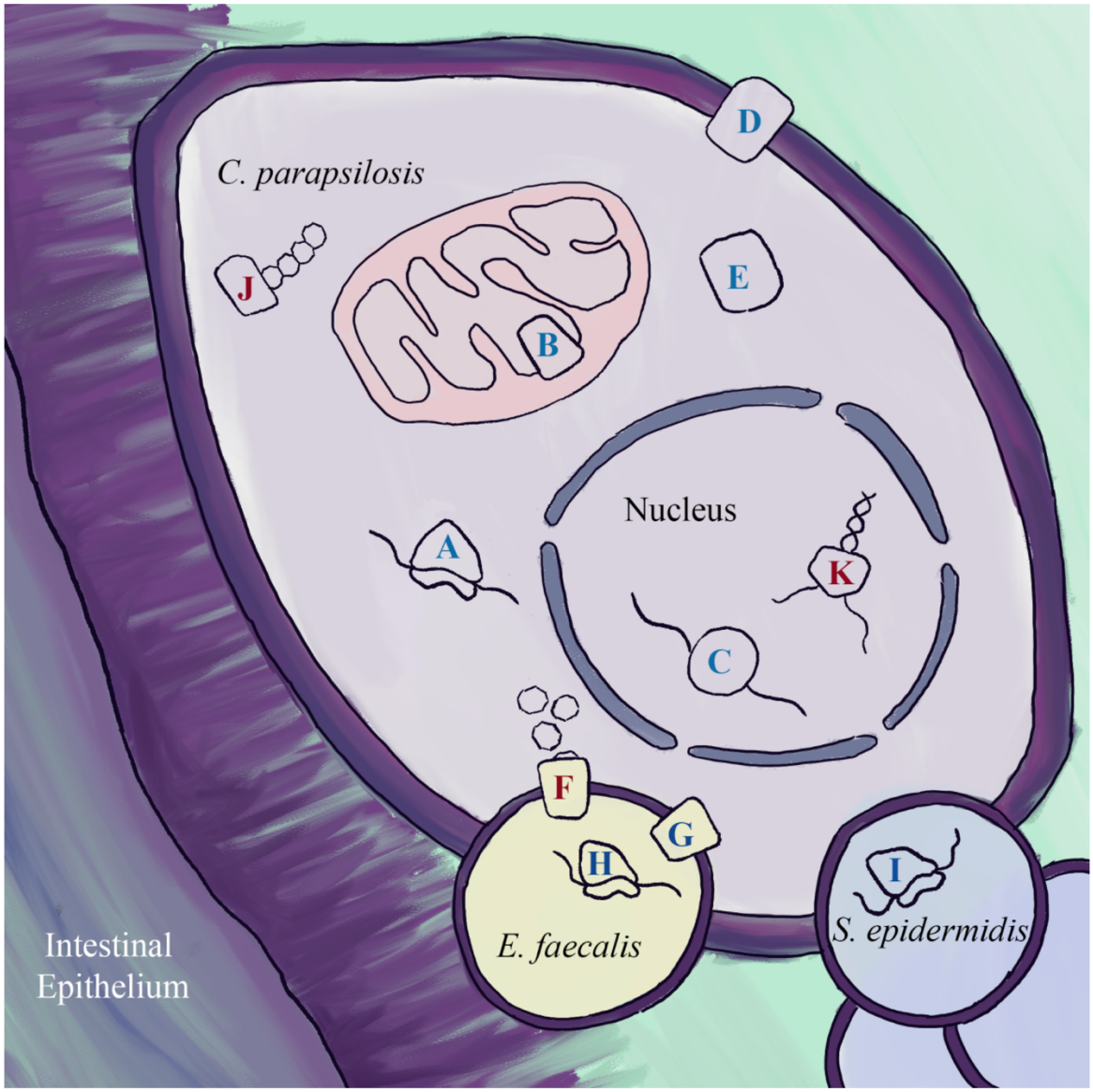

Figure 3.6. *in situ* enriched gene categories. Diagram depicting *C. parapsilosis* in the context of the infant gut, highlighting gene categories or families that were significantly enriched in differentially expressed genes between *in situ* and culture. Blue letters represent functions with higher expression *in situ*, while red represent functions with lower expression *in situ*. See Table S5 for details. (A) ribosomal proteins (B) cytochrome c oxidase subunits (C) LSM complexes (D) proton antiporters (F) *E. faecalis* mannose transporters (G) *E. faecalis* mannitol transporters (H) *E. faecalis* subset of ribosomal proteins (I) *S. epidermidis* subset of ribosomal proteins (J) *C. parapsilosis* polysaccharide synthases (downregulated *in situ*) (K) *C. parapsilosis* helicases (downregulated *in situ*).

Table 3.1. Overview of Candida strain genomes used in this study

Genome	Genus	Species	Length	# Scaffolds	N50	Completeness	Sample Type	Reference
C1_006	Candida	parapsilosis	11852211	191	108686	92	Infant fecal metagenome	This study
N3_182	Candida	parapsilosis	12563647	342	65710	97	Infant fecal metagenome	Olm et al. 2019
S2_005	Candida	parapsilosis	11573959	1051	14507	93	NICU metagenome	Olm et al. 2019
NYC Subway	Candida	parapsilosis	7420453	1285	6417	62	NYC subway metagenome	This study
L2_023	Candida	parapsilosis	4870205	2906	1700	35	Infant fecal metagenome	This study
CDC317	Candida	parapsilosis	13030174	9	2091826	97	Clinical skin isolate	Butler et al. 2009
GA1	Candida	parapsilosis	13025060	39	1114083	97	Clinical human blood isolate	Pyrszcz et al. 2013
CBS1984	Candida	parapsilosis	13044404	25	962200	96	Olive fruit isolate	Pyrszcz et al. 2013
CBS6318	Candida	parapsilosis	13050515	28	1691491	97	Healthy skin isolate	Pyrszcz et al. 2013
N1_023	Candida	albicans	13456346	1675	15180	94	Infant fecal metagenome	This study
N2_070	Candida	albicans	13540857	1614	14761	93	Infant fecal metagenome	This study
N5_264	Candida	albicans	11647081	746	27434	85	Infant fecal metagenome	This study
S3_003	Candida	albicans	11972257	1049	14710	87	Infant mouth metagenome	This study
S3_016	Candida	albicans	10068784	802	19749	86	Infant mouth, skin, and gut metagenome coassembly	This study
SP_CRL	Candida	albicans	12561678	897	22840	91	Infant fecal metagenome	Olm et al. 2019

For supplementary tables, figures, and information see:
https://www.biorxiv.org/content/10.1101/2020.03.23.004093v1.full.pdf+html

4 Evidence for interdependence and pathways of interaction between the alkaline-adapted Choanoflagellate Salpingoeca monosierra and its enclosed bacterial community

Patrick T. West, Kayley H. Hake, Alexander Crits-Christoph, Nicole King, Jillian F. Banfield*

4.1 Abstract

Bacteria can substantially impact the development and metabolism of microbial eukaryotes. A new species of choanoflagellate, *Salpingoeca monosierra*, forms multicellular rosettes that enclose a complex bacterial community within the hollow center. Here, we used shotgun metagenomic sequencing to reconstruct a 42 Mbp draft genome of this microbial eukaryote, along with genomes of 24 coexisting bacteria from a complex co-culture derived from Mono Lake, California. In contrast to the two previously sequenced choanoflagellates from marine environments, this organism inhabits a highly alkaline ecosystem. Potentially as an adaptation to high pH, the S. monosierra predicted cell surface and secreted proteins have significantly higher isoelectric points compared to those of the previously studied Choanoflagellates. The bacterial community associated with *S. monosierra* is dominated by species of Oceanospirillales, Ectothiorhodospiraceae, Flavobacteria, and Sphingobacteria. From genomic functional predictions, the bacterial community is largely composed of heterotrophs. Notably, however, three members have the capacity for CO2 fixation via the Calvin-Benson pathway and one highly abundant bacterium is capable of N2 fixation. This suggests that the Choanoflagellates benefit from carbon and nitrogen currencies produced by the bacteria, in return for existence in a protected environment. Both *S. monosierra* and an Ectothiorhodospiraceae species genomes encode prominent biosynthetic gene clusters predicted to produce hybrid polyketide-nonribosomal peptides that may play roles in eukaryotic-bacterial interactions. Our results expand understanding of the genomic and functional diversity of choanoflagellates and uncover potential mechanisms for their interaction with closely associated microbial communities.

4.2 Introduction

Microbial eukaryotes are important pathogens, microbial predators, photosynthesizers, environmental quality indicators, markers of past environmental change, and critical for understanding our own evolutionary history. However, as demonstrated by barcode sequencing, the vast majority of microbial eukaryotic diversity is currently uncultivable or not present in culture collections (Pawlowski et al. 2012), making research on these organisms exceedingly difficult. Studying microbial eukaryotes in their environmental or microbial community context is an even greater challenge due to the complexity of natural consortia. Microbial community context is likely of high significance, given that interactions between bacteria and eukaryotes influence development, metabolism, and evolution of all types of eukaryotes, ranging from animals such as sponges (Pita et al. 2018) and humans (Gilbert et al. 2018) to unicellular ciliated protozoa (Gong et al. 2016). In sponges, associated bacterial and archaeal species exhibit complex interactions with the host; sometimes serving as food sources, functioning as pathogens, or living as symbionts (Pita et al. 2018). Metabolic interactions in particular have been well documented in this system. Sponges are often host to photosynthetic cyanobacteria, and maintain cyanobacterial cell counts proportional to their own, suggesting control of their growth through stealing of photosynthates or consumption of excess photosymbionts (Taylor et al. 2007; Thacker et al. 2012). Similarly, nitrogenase activity (and therefore nitrogen fixation) has been detected in multiple cyanobacteria-containing sponges, and several heterotrophic nitrogen fixing bacteria have been isolated from sponges (Taylor et al. 2007). These observations suggest that symbiotic microbial nitrogen fixation may be an important source of nitrogen for macroscopic sponges.

Choanoflagellates, the closest known living organisms to animals, are unicellular heterotrophs that subsist largely on bacteria. However, interactions between choanoflagellates and bacteria extend beyond this metabolic relationship. In the choanoflagellate *Salpingoeca rosetta*, bacteria influence multicellular rosette formation (Alegado et al., 2012; Woznica et al., 2016) and sexual reproduction (Woznica et al., 2017) through the release of diffusible small molecules including lipids and a chondroitinase (Cantley et al. 2016; Woznica et al., 2017). Interestingly, these documented interactions are all with bacteria used for feeding choanoflagellates, not bacteria they would commonly encounter in their natural habitat. Little is known about the bacteria closely associated with Choanoflagellates in the environment, and their potential interactions.

Recently, the choanoflagellate *S. monosierra*, although unable to be isolated, was brought into co-culture with bacteria as part of a complex community derived from Mono Lake, California, offering an opportunity to study a choanoflagellate in a more natural community context. Like the marine choanoflagellate species, *Salpingoeca rosetta*, *S. monosierra* forms multicellular rosettes (Hake et al. 2019), but unlike *S. rosetta*, *S. monosierra* rosettes are hollow and enclose a microbial community, suggesting inter-domain interactions. However, it is unclear whether these associations are mutualistic and if so, how the various partners benefit. The inability to achieve a pure culture of *S. monosierra* and the limitations of barcode sequencing to study multi-Domain consortia necessitates new whole community-centric approaches to investigate bacterial composition, metabolic traits/capacities as well as potential modes of inter-Domain action. Untargeted shotgun sequencing of complex communities and environmental samples, in which all DNA from a sample is sequenced regardless of its organismal source or genetic context, can address these challenges. Most importantly, recently developed methods have made it relatively

straightforward to reconstruct the choanoflagellate genome from shotgun metagenomic samples without isolation (West et al. 2018).

Here, using shotgun metagenomic sequencing, we genomically describe the new species of choanoflagellate, S. monosierra, as well as its associated microbial community of diverse bacterial species that form a stable microbiome within the hollow *S. monosierra* rosette (Hake et al., 2019). Fluorescent in situ hybridization (FISH) studies that leveraged the 16S rRNA gene sequences from this dataset revealed that a subset of the bacterial community associated with *S. monosierra* (Hake et al., 2019) grows inside the rosettes. Here, we analyzed the full metagenomic dataset to describe the bacterial community at the species/strain level and report a variety of bacterial metabolisms present, including carbon fixation, nitrogen fixation, broad heterotrophy, and an abundance of sulfur and arsenic detoxifying enzymes. The genome of *S. monosierra* shows adaptation to the alkaline environment of Mono Lake. In addition, biosynthetic gene clusters (BGCs) were identified in both a bacterial species and *S. monosierra*, pointing to a potential method of interaction between these community members.

4.3 Results

4.3.1 Community structure

In order to elucidate the bacterial diversity, identify potential functions, and interactions within this microbiome, shotgun metagenomic sequencing was performed on complex co-cultures. *S. monosierra* co-cultures were grown in two different media; artificial Mono Lake water (see methods) with Media E as a carbon source, and artificial Mono Lake water treated with the antibiotic gentamicin, with cereal grass as a carbon source. Antibiotic treatment was applied to select against bacterial cells that were not enclosed within rosettes. Both co-cultures were centrifuged to concentrate *S. monosierra* rosettes. Pellets and supernatant were then sequenced separately, resulting in four metagenomic datasets (Figure 4.1) comprising 113.1 Gbp of sequencing data. 16S rRNA sequences recovered from the assembled metagenomic data suggest the presence of 10 bacterial species growing within the rosettes and 24 in total associated with *S. monosierra*. Across all four samples, 24 non-redundant bacterial genomes with >90% completeness were recovered. Additionally, using EukRep (West et al. 2018; see methods), the 47.2 Mbp genome choanoflagellate *S. monosierra* (Table 4.1) was recovered. 98.1% of reads mapped back to the non-redundant combined Choanoflagellate and bacterial genome set, indicating the binned genomes represent essentially all of the diversity present in the co-cultures.

Across all four samples and both media settings, the bacterial community is remarkably similar; being dominated by Oceanospirillales, Ectothiorhodospiraceae (purple sulfur bacteria), and Bacteroidetes (Figure 4.1A,B, Supplemental Figure S1). Interestingly, Oceanospirillale_55_89, the species present in nearly every rosette based on fluorescent in situ hybridization (Hake et al. 2019) was not present in rosette-depleted samples, suggesting it may only be present in rosettes, whereas Oceanospirillale_52_91 and other common rosette residents such as Ectothiorhodospiraceae_64_283 (Hake et al. 2019) are highly abundant both in rosette enriched samples and rosette depleted samples, indicating rosette residents may be a mix of obligate and non-obligate members.

Of the 24 bacterial species identified growing within the co-cultures, 10 species were previously demonstrated to be growing within rosettes via fluorescent in situ hybridization (FISH) of species specific 16s sequences (Hake et al. 2019). However, connecting a metagenomically binned genome to a 16S rRNA gene sequence is a challenge because this gene often does not assemble and because it is usually present in multiple copies within a genome and may not bin with the genome due to higher coverage. The 16S rRNA gene sequences for Ectothiorhodospiracea_64_283 and Oceanospirillales_55_89, both bacteria demonstrated to be present within rosettes, were successfully reassembled into genomes with metaSPAdes (see methods). A few other genomes show strong abundance correlations with rRNA genes but could not be directly linked via assembly.

Bacterial replication rates were estimated with iRep (Brown et al. 2016) for each community member and compared across conditions (Figure 4.1B). Replication rates were relatively consistent across all samples, with the exception of the rosette depleted gentamicin treated sample, in which replication rates were significantly higher. Increased replication rates are often observed in antibiotic treated samples (Brown et al. 2016) and the lack of increased replication rate in in the rosette enriched gentamicin treated sample suggests that the antibiotic may not be reaching inside the rosettes or is entering at a diminished rate. Whether there is regulation of small molecules into *S. monosierra* rosettes is an open question. However, rosettes are impermeable to beads the size of bacterial cells and bacterial community members were never completely removed from colonies, even after greater than 6 weeks of treatment with multiple antibiotics (Hake et al. 2019).

4.3.2 Bacterial community metabolism

The nature of the relationship between *S. monosierra* and the bacteria located within its rosette is currently unknown. Whether the intra-rosette bacteria are performing a distinct role or function advantageous to *S. monosierra* is of interest. One possibility is that the bacteria provide resources such as fixed nitrogen that are useful for *S. monosierra*. Indeed, one bacterium, Oceanospirillales 52_91, is capable of nitrogen fixation and is the most abundant bacterium in three of the four metagenomic samples. The majority of bacteria from rosette enriched samples, however, appear to be primarily heterotrophic, with a large suite of CAZymes, a complete or near complete TCA cycle, and at least one form of a cytochrome c oxidase (Figure 4.2, Table S4.1), suggesting a broad capacity for heterotrophy and aerobic respiration. A number of organisms also have the capacity for nitrate reduction (2/24), nitrite reduction (4/24), or both (6/24) indicating nitrate may be an important electron acceptor for this community in addition to oxygen, and that these members of the community may be facultative anaerobes. Three bacteria are predicted to have portions of the denitrification pathway up to reduction of nitrite to nitric oxide (Table S4.1). This may be interesting, as nitric oxide is a known signaling molecule in mammals. 23 community members contain sulfur dioxygenases (Figure 4.2), which have been broadly found in heterotrophic bacteria in the past (Liu et al. 2014) and are most commonly used for the detoxification of sulfide. 20 bacterial organisms have capacity for arsenate reduction through arsC, also possibly for detoxification via arsenate reduction and export (Martin et al. 2001). Four organisms are predicted to be capable of CO oxidation (Table S4.2) and three of these (Rhodobacterales_64_110, Oceanospirillales_55_89, and Ectothiorhodospiraceae_64_108) have capacity for nitrate reduction

and cytochrome c oxidases. Interestingly at least one bacterium isolated from Mono Lake can oxidize CO with oxygen in aerobic conditions and nitrate in anaerobic conditions (King 2015).

Despite the presence of numerous ectothiorhodospiraceae, a clade typically known for its anaerobic, photosynthetic members, we found no complete pathways for photosynthesis. However, two Ectothiorhodospiracea and an Oceanosprillales have the full Calvin cycle, including RuBisCO. The three RuBisCO proteins were identified with HMMs and their forms determined by placement in a phylogenetic tree (Figure S4.2). The form I RuBisCOs (with small and large subunits) belong to Ectothiorhodospiraceae_67_18 and Ectothiorhodospiraceae_64_283 and the Form II RuBisCO belongs to Oceanospirillales_52_91. Based on the Calvin cycle genes and RuBisCO forms, we conclude that these bacteria can fix CO_2. The two Ectothiorhodospiraceae appear to have similar metabolic capacity, as both contain a full pathway for the sequential oxidation of methanol to CO_2 through formaldehyde and formate (Supplemental table S4.1). Methylotrophic autotrophy has been observed before in proteobacteria (Dedysh et al. 2005). The third potential autotroph, Oceanospirillales_52_91 is the most abundant organism in three of the four samples and has the capacity for nitrogen fixation (Figure 4.2, Table S4.1).

4.3.3 S. monosierra Genome

An *S. monosierra* genome, of comparable completeness to the two genomes of previously sequenced choanoflagellates, was reconstructed from both rosette enriched samples (Table 4.1). Assembling merged reads from both samples did not result in a higher quality assembly (Table S2), therefore the genome from the rosette enriched Media E sample was selected as the representative genome. Although the estimated completeness is very similar to those of the previously sequenced *S. rosetta* and *M brevicollis* genomes, the overall assembly is more fragmented with an N50 of 0.029 Mbp and L50 of 503. This should not be an inherent issue of metagenomic assemblies for microbial eukaryotes, as less fragmented fungal genomes have been previously assembled from metagenomic datasets (West et al. 2018). Given low genomic heterozygosity (i.e., low incidence of divergent alleles that can fragment assemblies; Figure S4.3), we attribute genome fragmentation to repeats. For example, the genome has a large fraction of non-coding regions, which frequently contain repetitive elements capable of fragmenting eukaryote genome assemblies (Tørresen et al. 2019). In particular, the genome has ~11,000 occurrences of long strings of solely T nucleotides (20 bp or greater in length) that we will refer to as polyT tracts (Figure 4.3E). The distribution of *S. monosierra* polyT tract lengths shows a sharp peak at a length of 100 bp (Figure 4.3D) and a maximum length of 100 bp, suggesting 150 bp reads were unable to span the full length of the tracts. Supporting the expectation that they should contribute to fragmentation of the *S. monosierra* genome, we note that 15% of scaffold breaks end in a polyT tract.

Interestingly, we noticed in *S. monosierra*, the polyT tracts were relatively concentrated in intronic spaces of the genome (Figure 4.3F-G). They are polyT rather than polyA tracts as they occur in the direction of the gene sequence 90% of the time (Figure 4.3G). To determine whether such long polyT tracts are unusual, we scanned over 600 eukaryotic genomes for polyT tracts $\geq$ 20 bp in length and found three other genomes with similar phenomena (Figure 4.3D-G): *D. discoideum*, *Orpinomyces sp.*, and *R. microsporus*. Despite all four organisms belonging to the Opisthokonta, there does not appear to be a clear phylogenetic relationship between them. *D. discoideum* and *Orpinomyces sp*. had a high number of mostly very short polyT tracts, randomly distributed

throughout their genome. *R. microsporus* had very long polyT tracts ranging in length upwards of 600 bp that also appeared to be randomly distributed throughout its genome.

4.3.4 S. monosierra genetic diversity

To get an idea of the genetic diversity present in *S. monosierra* populations, Single Nucleotide Variants (SNVs) were called across the *S. monosierra* genome. However, for accurate interpretation and calling of SNV patterns, it is important to estimate genome ploidy. Plots of the density of allele frequencies across the entire genome suggest the *S. monosierra* genome is triploid (Figure 4.3C) with allele frequency peaks at roughly one third and two thirds in both rosette enriched samples (Figure S4.3). Thus, at each allele position, it is inferred there are three alleles, two of which are the same and one that differs. However, the linkage of allele variants within the three chromosomes cannot be determined. With that noted, the heterozygosity of the genome is remarkably low, indicating very low genetic diversity, with 0.26 heterozygous SNVs per kbp. The relatively infrequent SNVs are also evenly distributed throughout the assembly, indicating a relatively clonal *S. monosierra* population present in the co-culture (Figure S4.3).

4.3.5 Alkaline adaptation in S. monosierra

Mono lake, an alkaline, hypersaline lake, applies strong selective pressures on microbial organisms. In highly acidic environments, it has been observed that bacteria and archaea have shifted average proteomic isoelectric points to adapt to the low extracellular pH (Bardavid et al. 2012). However, adaptation to such environments has rarely been examined in microbial eukaryotes. For signs of adaptation, we compared the amino acid composition and isoelectric points of proteins in *S. monosierra* relative to *S. rosetta* and *M. brevicollis* (Figure 4.3A). *S. monosierra* exhibits a significantly higher average isoelectric point in secreted proteins compared to both *S. rosetta* ($p=6.5\times10^{-6}$) and *M. brevicollis* ($p=1.4\times10^{-6}$). The same trend was not present in non-secreted proteins, emphasizing the change in isolelectric points is likely an adaptation to the alkaline environment as the pH of the intracellular environment is generally tightly regulated (Boron 2004). Each amino acid has varying individual isoelectric points and it is conceivable that overall protein isoelectric point could be modified by substituting amino acids with similar amino acids with higher individual isoelectric points. However, by examining differences in amino acid frequencies between *S. monosierra* and other choanoflagellates (Figure 4.3B), it is not obvious what amino acid compositional differences may be driving the total difference in total isoelectric point.

4.3.6 Biosynthetic Gene Clusters

One of the most highly abundant bacterial bins, Ectothiorhodospiraceae_64_283, contains a 90 kb novel biosynthetic gene cluster (BGC) with three large ORFs that encode a trans-AT polyketide synthase (PKS) and a nonribosomal peptide synthetase (NRPS) hybrid (Figure 4.4A). The gene cluster ends with two thioesterase domains and the single nonribosomal peptide subunit with a condensation domain and adenylation domain predicted by antiSMASH (Blin et al. 2019) and PRISM (Skinnider et a. 2017) to be specific for arginine. Given previous documented interactions between bacteria and choanoflagellates involving small molecules (Woznica et al 2017), it is possible this cluster may represent a mode of interaction. Surprisingly, the *S. monosierra* genome contains biosynthetic genes of its own in the form of two polyketide synthetase genes (PKS) spread across two separate contigs (Figure 4.4B). One of the ORFs contains a Malonyl-CoA decarboxylase while the other contains an NRPS condensation domain with no corresponding

adenylation domain, and neither contains an acyltransferase domain. Due to the absence of both acyltransferase and adenylation domains, it is likely that these genes are a portion of a larger gene cluster. Despite the biosynthetic genes being spread across multiple contigs, the genes were assembled similarly in the two replicate genomes and are the first polyketide biosynthesis genes reported in the Choanozoa.

4.4 Discussion

Choanoflagellates are the closest known living relatives of animals, and given their unique phylogenetic position, the nature of interactions between bacterial community members and *S. monosierra* is of intense interest. Analysis of metabolic potential identified bacteria with pathways for non-photosynthetic carbon fixation through the Calvin-Benson pathway, including Oceanospirillales_55_89, a species consistently identified within the *S. monosierra* rosettes (Hake et al. 2019), as well as nitrogen fixation. However, the majority of bacterial species appear to be heterotrophic, based on their metabolic potential, including some of the species most commonly identified inside rosettes (Hake et al. 2019). Thus, it is possible the bacteria use these pathways and provide carbon and nitrogen compounds to the choanoflagellate host, similar to relationships observed between photosynthetic bacteria and algae in lichens (Armaleo et al. 2019), sponges (Taylor et al. 2007; Thacker et al. 2012), and corals (Muller-Parker et al. 2015). It is also possible the bacteria provide functioning similar to a proto-gut, degrading complex carbon compounds and giving a portion of the products to the Choanoflagellate host. A wide diversity of bacteria contain sulfur dioxygenases, commonly used for sulfide detoxification in many heterotrophic bacteria as well as almost all animals (Liu et al. 2014). Thus, bacterial community members could assist *S. monosierra* to detoxify sulfide, augmenting the functioning of the *S. monosierra* sulfur dioxygenase (Table S4.1). Alternatively, it is possible the relationship is entirely predatory and the choanoflagellate hosts consume intra-rosette bacterial cells at leisure. However, *S. monosierra* maintains a bacterial population within its rosettes for an extended period of time, suggesting it has at least some control over the growth of intra-rosette bacterial populations.

In exchange for resources provided to the Choanoflagellate, bacteria may benefit from localization within rosettes because their metabolism is enhanced by a potentially anaerobic environment provided by the *S. monosierra* rosettes (Boyd et al. 2013). Alternatively, the rosettes may simply provide an environment protected from predation or with otherwise modified biogeochemical characteristics (e.g., pH, inorganic ion concentrations, nutrients - e.g., nitrogen compounds) for the bacteria to grow in. Thus, overall, the bacterial community members and *S. monosierra* help one another adapt to the harsh environment of Mono Lake.

It is interesting, however, that some bacterial functions may be deleterious to the host. For example, Mono Lake is a site with high arsenic concentrations and well-studied from the perspective of bacterial arsenic metabolism. Bacterial community members display wide capacity for arsenate detoxification via arsenate reduction. However, detoxification via arsC converts As5+ to the more toxic As3+.

Secondary metabolites have previously been demonstrated to cause distinct changes in choanoflagellate morphology and behavior (Cantley et al. 2016). Curiously, the presence of Ectothiorhodospiraceae_64_283, the bacteria with a hybrid polyketide-nonribosomal peptide BGC, was strongly correlated with a unique *S. monosierra* rosette phenotype in which the rosette 'bursts' as bacterial growth appears to outgrow the confines of the rosette interior (Hake et al. 2019). It is possible that inter-domain chemical communication plays a role in causing this unique phenotype. In this particular case, early evidence suggests this interaction is not beneficial to *S. monosierra* and may be pathogenic.

Despite the overall low heterozygosity of the *S. monosierra* genome, aggregate allele frequencies suggest the genome is triploid. This conclusion is supported by allele frequencies from two independent sequencing samples and by the fact that the SNVs are spread throughout the entire assembly. However, it has previously been determined *S. rosetta* undergoes a ploidy shift between haploid and diploid as part of its sexual cycle (Levin et al. 2013). Measuring allele frequencies at a single time point with the low number of SNVs present in the *S. monosierra* genome may not have the resolution to detect multiple ploidy levels within a single population. *S. monosierra* displayed remarkably low heterozygosity, with a mere 0.26 SNVs per kbp, that could indicate that it is an asexual or selfing species in contrast to *S. rosetta* which undergoes mating (Levin et al. 2013). It could also suggest a recent population bottleneck, possibly linked to laboratory cultivation. Interestingly, other choanoflagellates also have overall low heterozygosity (King et al. 2008) and are highly resistant to genetic perturbation. Low heterozygosity may be emblematic of their apparent lack of genomic plasticity.

The availability of a complex co-culture presents a rare opportunity to study a choanoflagellate with closely associated bacteria cultured from the same environment. Due to the unique phylogenetic position of choanoflagellates, the genomic and community compositional results reported here provide a foundation for further study to better understand the types of bacterial-eukaryotic interactions that influenced early animal development and evolution.

4.5 Methods

4.5.1 Sequencing and assembly

Raw sequencing reads were processed with bbtools (http://jgi.doe.gov/data-and-tools/bbtools/) and quality-filtered with SICKLE with default parameters (version 1.21; https://github.com/najoshi/sickle). IBDA_UD (Peng et al. 2012) was used to assemble and scaffold filtered reads. IDBA_UD was chosen as it is a widely used, publicly available program designed for metagenomic assemblies. Unlike almost all other such assemblers, it includes a scaffolding step. This is important because longer sequences can be more robustly binned. Scaffolding errors were corrected using MISS (I Sharon, BC Thomas, JF Banfield, unpubl.), a tool that searches and fixes gaps in the assembly based on mapped reads that exhibit inconsistencies between raw reads and assembly.

4.5.2 Targeted reassembly with SPAdes

Assembled 16S sequences and the split contigs of a repeat-rich secondary metabolite cluster were used as trusted contigs (--trusted-contigs option) in a run of the SPAdes assembler (Nurk et al. 2013). In doing so, a subset of 16S sequences were incorporated into larger scaffolds that could then be associated with a bin. The secondary metabolite cluster was also successfully assembled into a single scaffold, enabling structural prediction of the metabolite and confidence in having the entire cluster assembled.

4.5.3 Prokaryotic binning and annotations

Prodigal with the -meta option (Hyatt et al. 2012) was used to predict protein-coding genes on each assembled metagenomic sample. Ribosomal RNAs were predicted with Rfam (Nowrocki et al, 2015). Predicted proteins were given functional annotations by aligning to UniProt (UniProt, 2010), UniRef90 (Suzek et al. 2007) and KEGG (Kanehisa et al. 2016). Prokaryotic draft genomes were binned with CONCOCT (ref). Bins were then refined through the use of ggKbase (ggkbase.berkeley.edu) by manually checking GC, coverage, and the phylogenetic profile of a given bins constitutive contigs. A non-redundant set of bins across all four samples was then generated using dRep (Olm et al. 2017).

4.5.4 S. monosierra binning and annotation

The *S. monosierra* genome was binned as described in detail in West et al. 2018. Briefly, EukRep (West et al. 2018) was run on each individual sample to separate eukaryotic and prokaryotic contigs. Predicted eukaryotic contigs were then binned into putative bins using CONCOCT (Alneberg et al. 2014). For each bin, protein-coding genes were predicted using the MAKER2 pipeline (Holt and Yandell, 2011) with default parameters, self-trained GeneMark-ES (Ter-Hovhannisyan et al. 2008), AUGUSTUS (Stanke et al. 2006) trained with BUSCO (Simao et al. 2015), and the reference proteomes of *S. rosetta* and *M. brevicollis* as homology evidence. Genome assembly stats were generated with a custom ruby script.

4.5.5 Phylogenetic analysis

Bacterial protein sets obtained from NCBI and JGI's genome portal (ref) For each protein set, 16 ribosomal proteins (L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24, S3, S8, S10, S17, and S19) were identified by BLASTing a reference set of 16 ribosomal proteins obtained from a variety of protistan organisms against the protein sets. BLAST hits were filtered to a minimum e-value of $1.0 \times 10{-}5$ and minimum target coverage of 25%. The 16 ribosomal protein data sets were aligned with MUSCLE (v. 3.8.31) (Edgar 2004) and trimmed by removing columns containing 90% or greater gaps. The alignments were then concatenated. A maximum likelihood tree was constructed using RAxML (v. 8.2.10) (Stamatakis 2014), on the CIPRES web server (Miller et al. 2010), with the LG plus gamma model of evolution (PROTGAMMALG) and with the number of bootstraps automatically determined with the MRE-based bootstopping criterion.

4.5.6 SNV analysis

In order to identify variants in the S. monosierra genome, reads from samples A and C (rosette enriched samples) were mapped back to the de novo assembled *S. monosierra* genome with Bowtie 2 (Langmead 2012). The PicardTool (http://broadinstitute.github.io/picard/) functions "SortSam" and "MarkDuplicates" were used to sort the resulting sam file and remove duplicate reads. FreeBayes (Garrison and Marth, 2012) was used to perform variant calling with the options "--

pooled-continuous -F 0.01 -C 1." Variants were filtered downstream to include only those with support of at least 10% of total mapped reads in order to avoid false positives. Multiallelic sites were defined as sites with two or more non-reference alleles. To determine the ploidy of the genome, the allele frequency for each allele at each variant site was plotted as a histogram. The distribution of SNPs across the genome was visualized with a custom matplotlib function.

4.5.7 Isoelectric point analysis

Reference protein sets for *S. rosetta* and *M. brevicollis* were obtained from NCBI. Predicted protein sets for all three choanoflagellates were run through SignalP (Pretersen et al. 2011) to identify putative transmembrane and secreted proteins. Secreted and non-secreted proteins were then grouped separately and isoelectric points were calculated for each protein using the IPC software (Kozlowski 2016). Significant differences between secreted and non-secreted protein sets and between genomes were calculated using the rank-sums test (Pedregosa et al. 2011). The amino acid frequencies from *S. monosierra* predicted proteins were then compared to the average amino acid frequencies of *S. rosetta* and *M. brevicollis*.

4.5.8 polyT tracts

PolyT tracts were identified by searching for strings of either A's or T's throughout the genome in question. Only strings longer than 19 nucleotides in length were considered to be polyT tracts. The S. monosierra genome, *S. rosetta* and *M. brevicollis* genomes were searched along with over 600 eukaryotic genomes obtained from NCBI and JGI's Mycocosm (Grigoriev et al. 2011). PolyT tracts overlapping genes, introns, and other genomic features were identified using the pyBedTools suite (Quinlan et al. 2010) "intersect" function. Relative distance plots were generated using a custom matplotlib function.

4.6 Figures

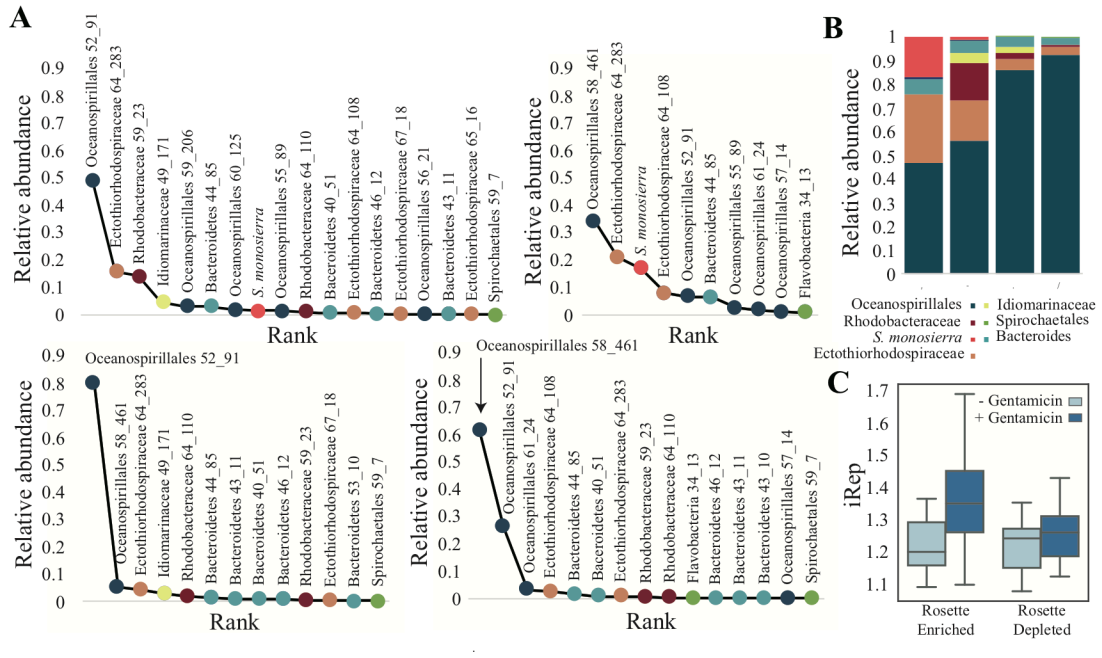

Figure 4.1. Community composition of complex co-culture samples derived from Mono Lake. (A) Rank abundance curves for samples A, B, C, and D respectively. Points are colored by organism family. (B) Phylogenetic composition of each of the four metagenomic samples. Samples ordered as A, C, B, and D. (C) Estimated replication rates (iRep values) for bacterial genomes, separated by sample.

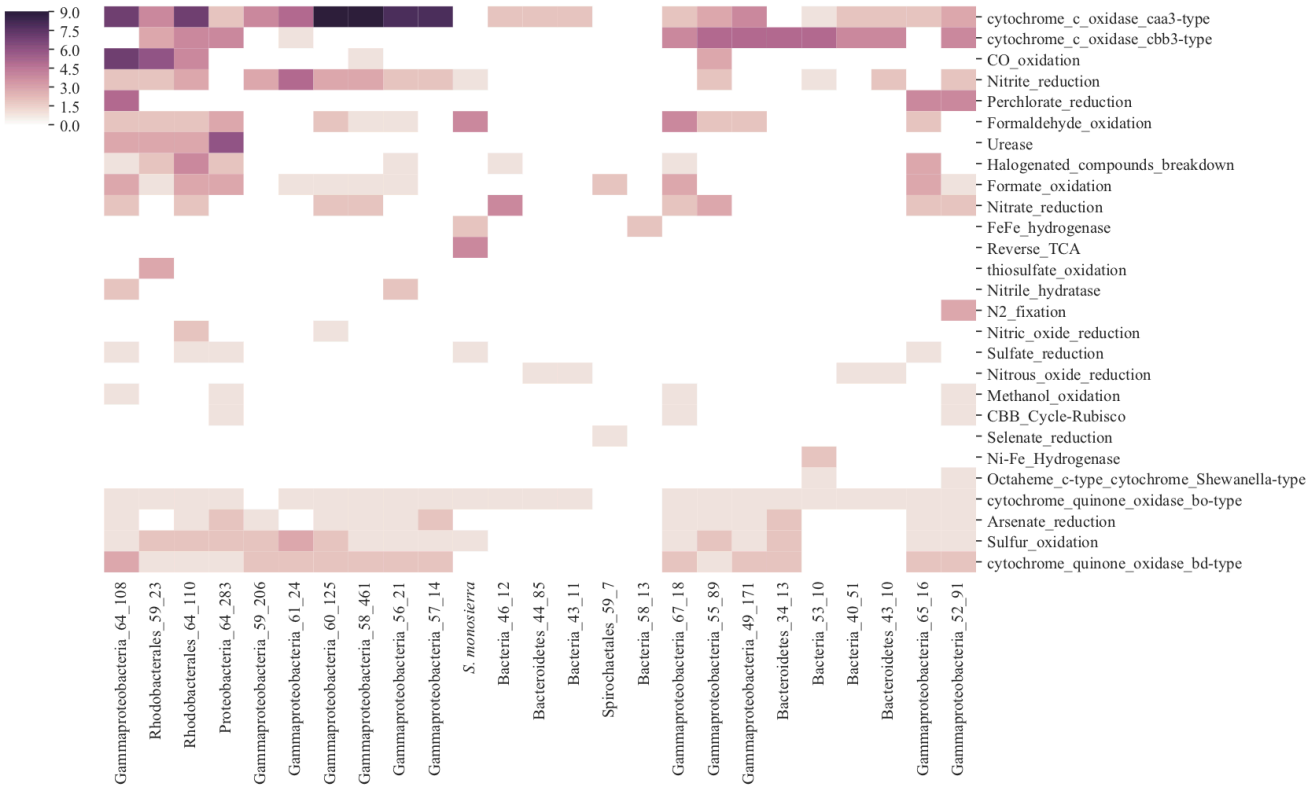

Figure 4.2. Summary of metabolisms present in reconstructed genomes. Bacterial genomes are represented as a non-redundant set. Metabolism presence is determined by presence or absence of marker genes unique to a particular metabolic pathway. Darker colors represent multiple copies of the marker gene being present in a given genome.

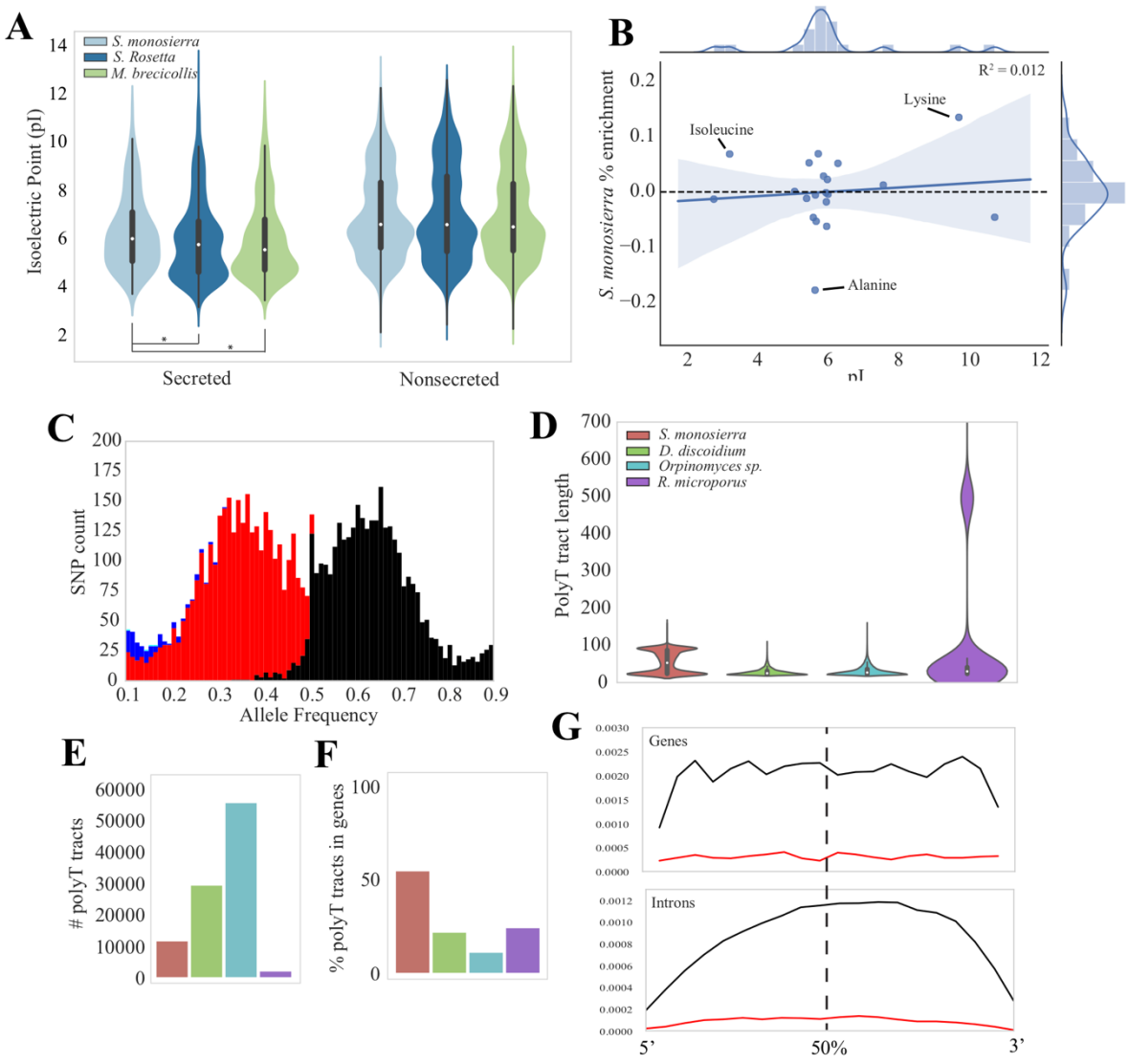

Figure 4.3. Characteristics of the *S. monosierra* genome. (A) Violin plots of protein isoelectric points for proteins from *S. monosierra*, *S. rosetta*, and *M. brevicollis*. Asterisks denote significant differences between distributions. (B) The enrichment of individual amino acids in *S. monosierra* secreted proteins as compared to *S. rosetta* and *M. brevicollis* proteins plotted against the isoelectric point of the individual amino acids. (C) Allele frequency density plot for SNVs called within the *S. monosierra* assembly. Black denotes the frequency of the highest frequency allele, red the second highest, and blue the third highest if the site has more than two alleles. (D) Distriubution of polyT tract lengths. (E) Total number of polyT tracts. (F) The percent of polyT tracts located within genes in each respective genome. (G) Relative distance plots of the presence of polyT tracts averaged across all *S. monosierra* genes and introns.

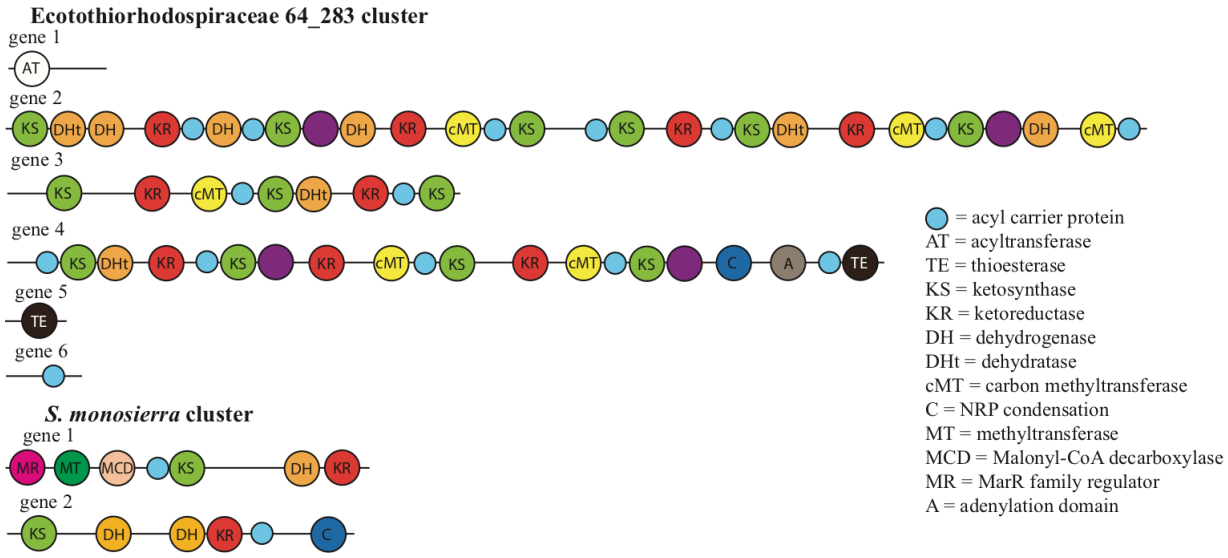

Figure 4.4. Visual representation of the biosynthetic gene clusters in Ectothiorhodospiraceae and *S. monosierra*.

Name	***S. monosierra***	***M. brevicollis***	***S. rosetta***
Size (mbp)	47.2	41.6	55.5
# genes	10895	9203	11731
# scaffolds	2107	218	125
Longest Scaffold (Mbp)	0.163944	3.607471	4.624088
N50 (Mbp)	0.029177	1.073601	1.890124
L50	503	13	10

Table 4.1. *S. monosierra* genome assembly quality.

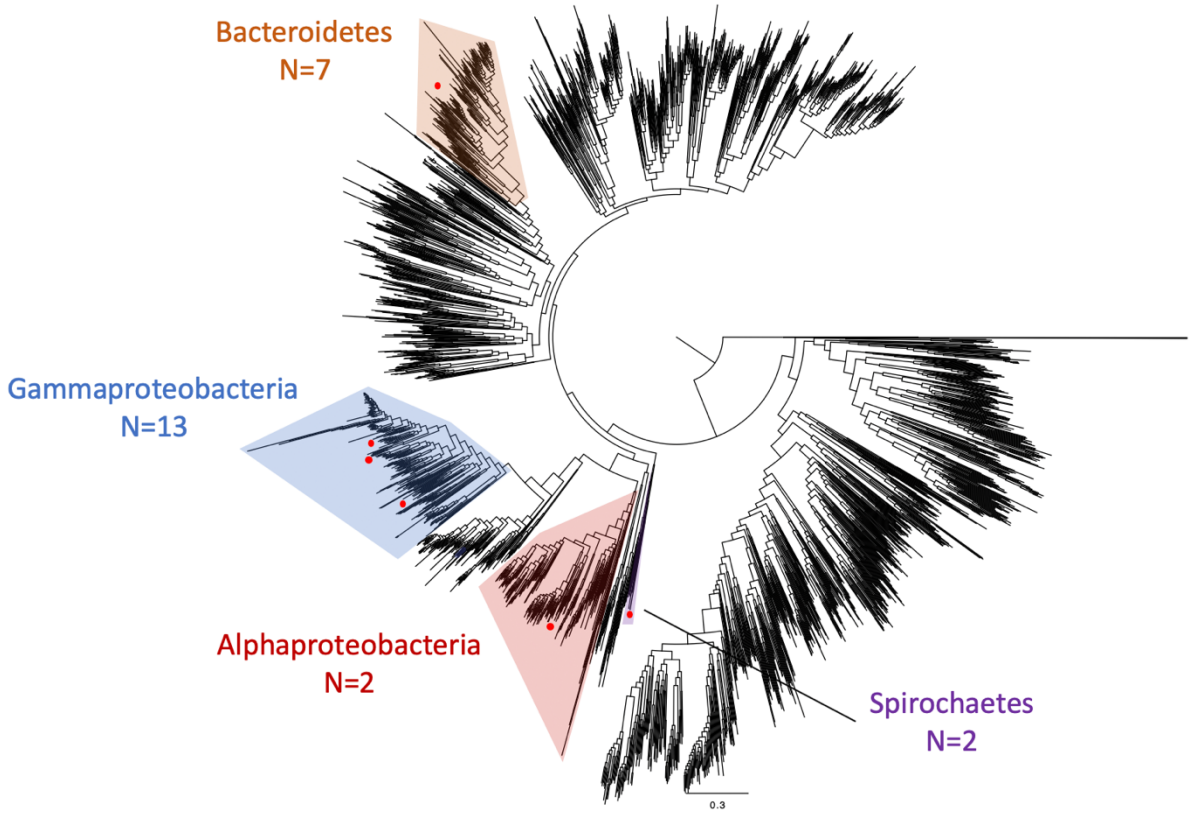

Figure S4.1. Concatenated 16S ribosomal tree of bacterial organisms from complex co-culture samples. Binned organisms are represented by red dots on the tree. Most binned organism belonged to the Bacteroidetes (Sphingobacteria) and Gammaproteobacteria (Oceanospirillales and Ectothiorhodospiraceae).

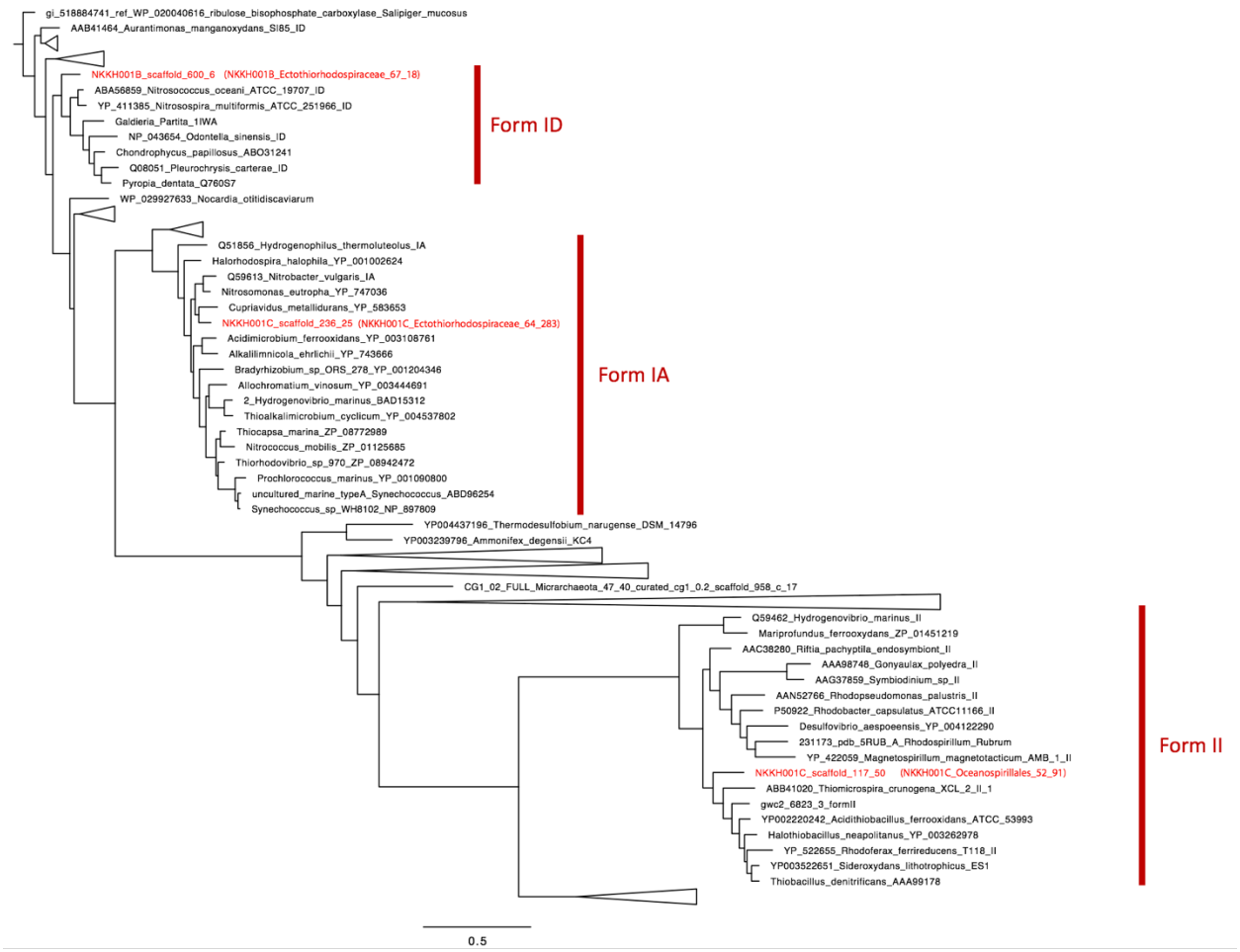

Figure S4.2. RuBisCO large subunit tree. Reference RuBisCO proteins were collected and a phylogenetic tree was constructed including RuBisCO proteins identified from complex co-culture samples (highlighted in red). Novel RuBisCO's were determined to be of Form I and Form II.

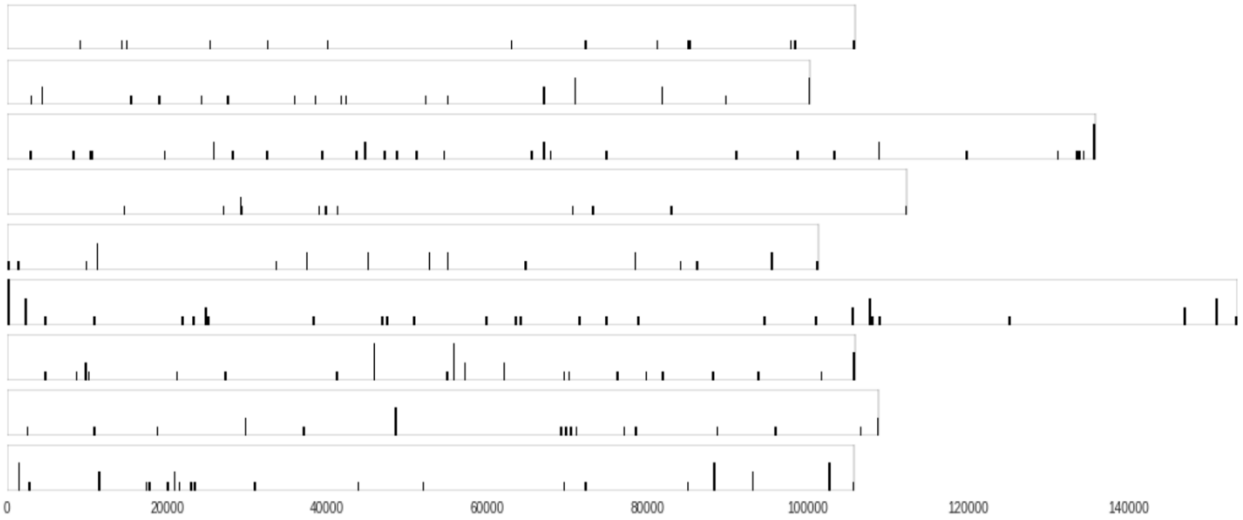

Figure S4.3. SNV distribution across the *S. monosierra* genome. The nine longest S. monosierra genome assembly scaffolds are included, where each row represents a single scaffold. Black bars represent the total number of SNVs present within 100 bp sliding windows across the *S. monosierra* assembly.

HMM	Gammaproteobacteria_57_14.	Gammaproteobacteria_58_461	Gammaproteobacteria_61_24	Gammaproteobacteria_64_108	Proteobacteria_64_283	Bacteria_43_11	Bacteria_46_12	Bacteria_58_13
Nitrite_reduction	2	3	5	2	0	0	0	0
cytochrome_c_oxidase_cbb3-type	0	0	1	0	4	0	0	0
Formaldehyde_oxidation	0	1	0	2	3	0	0	0
CO_oxidation	0	1	0	7	0	0	0	0
Nitrate_reduction	0	2	0	2	0	0	4	0
Sulfur_oxidation	1	1	3	1	2	0	0	0
cytochrome_c_oxidase_caa3-type	8	9	5	7	2	2	2	0
cytochrome_quinone_oxidase_bo-type	1	1	1	1	1	1	1	0
cytochrome_quinone_oxidase_bd-type	2	2	2	3	1	0	0	0
Arsenate_reduction	2	1	0	1	2	0	0	0
Halogenated_compounds_breakdown	0	0	0	1	2	0	1	0
Formate_oxidation	0	1	1	3	3	0	0	0
Nitrile_hydratase	0	0	0	2	0	0	0	0
Nitric_oxide_reduction	0	0	0	0	0	0	0	0
Perchlorate_reduction	0	0	0	5	0	0	0	0
Sulfate_reduction	0	0	0	1	1	0	0	0
Selenate_reduction	0	0	0	0	0	0	0	0
Nitrous_oxide_reduction	0	0	0	0	0	1	0	0
Ni-Fe_Hydrogenase	0	0	0	0	0	0	0	0
Octaheme_c-type_cytochrome	0	0	0	0	0	0	0	0
Methanol_oxidation	0	0	0	1	1	0	0	0
CBB_Cycle-Rubisco	0	0	0	0	1	0	0	0
Urease	0	0	0	3	6	0	0	0
thiosulfate_oxidation	0	0	0	0	0	0	0	0
FeFe_hydrogenase	0	0	0	0	0	0	0	2
Reverse_TCA	0	0	0	0	0	0	0	0
N2_fixation	0	0	0	0	0	0	0	0

HMM	Bacteroidetes_44_85	Gammaproteobacteria_49_171	Gammaproteobacteria_67_18	Rhodobacterales_59_23	Rhodobacterales_64_110	Bacteroidetes_34_13	S. monosierra	Gammaproteobacteria_52_91
Nitrite_reduction	0	0	0	2	3	0	1	2
cytochrome_c_oxidase_cbb3-type	0	5	4	3	4	5	0	4
Formaldehyde_oxidation	0	2	4	2	2	0	4	0
CO_oxidation	0	0	0	6	4	0	0	0
Nitrate_reduction	0	0	2	0	2	0	0	2
Sulfur_oxidation	0	1	1	2	2	2	1	1
cytochrome_c_oxidase_caa3-type	2	4	2	4	7	0	0	3
cytochrome_quinone_oxidase_bo-type	1	1	1	1	1	1	1	1
cytochrome_quinone_oxidase_bd-type	0	2	2	1	1	2	0	2
Arsenate_reduction	0	1	1	0	1	2	0	1
Halogenated_compounds_breakdown	0	0	1	2	4	0	0	0
Formate_oxidation	0	0	3	1	3	0	0	1
Nitrile_hydratase	0	0	0	0	0	0	0	0
Nitric_oxide_reduction	0	0	0	0	2	0	0	0
Perchlorate_reduction	0	0	0	0	0	0	0	4
Sulfate_reduction	0	0	0	0	1	0	1	0
Selenate_reduction	0	0	0	0	0	0	0	0
Nitrous_oxide_reduction	1	0	0	0	0	0	0	0
Ni-Fe_Hydrogenase	0	0	0	0	0	0	0	0
Octaheme_c-type_cytochrome	0	0	0	0	0	0	0	1
Methanol_oxidation	0	0	1	0	0	0	0	1
CBB_Cycle-Rubisco	0	0	1	0	0	0	0	1
Urease	0	0	0	3	3	0	0	0
thiosulfate_oxidation	0	0	0	3	0	0	0	0
FeFe_hydrogenase	0	0	0	0	0	0	2	0
Reverse_TCA	0	0	0	0	0	0	4	0
N2_fixation	0	0	0	0	0	0	0	3

HMM	Gammaproteobacteria_55_89	Gammaproteobacteria_56_21	Gammaproteobacteria_59_206	Gammaproteobacteria_60_125	Gammaproteobacteria_65_16	Spirochaetales_59_7	Bacteria_40_51	Bacteria_53_10
Nitrite_reduction	2	2	3	3	0	0	0	1
cytochrome_c_oxidase_cbb3-type	5	0	0	0	0	0	4	5
Formaldehyde_oxidation	2	1	0	2	2	0	0	0
CO_oxidation	3	0	0	0	0	0	0	0
Nitrate_reduction	3	0	0	2	2	0	0	0
Sulfur_oxidation	2	1	2	2	1	0	0	0
cytochrome_c_oxidase_caa3-type	3	8	4	9	2	0	2	1
cytochrome_quinone_oxidase_bo-type	1	1	0	1	1	0	1	1
cytochrome_quinone_oxidase_bd-type	1	2	2	2	2	0	0	0
Arsenate_reduction	1	1	1	1	1	0	0	0
Halogenated_compounds_breakdown	0	1	0	0	3	0	0	0
Formate_oxidation	0	1	0	1	3	2	0	0
Nitrile_hydratase	0	2	0	0	0	0	0	0
Nitric_oxide_reduction	0	0	0	1	0	0	0	0
Perchlorate_reduction	0	0	0	0	4	0	0	0
Sulfate_reduction	0	0	0	0	1	0	0	0
Selenate_reduction	0	0	0	0	0	1	0	0
Nitrous_oxide_reduction	0	0	0	0	0	0	1	0
Ni-Fe_Hydrogenase	0	0	0	0	0	0	0	2
Octaheme_c-type_cytochrome	0	0	0	0	0	0	0	1
Methanol_oxidation	0	0	0	0	0	0	0	0
CBB_Cycle-Rubisco	0	0	0	0	0	0	0	0
Urease	0	0	0	0	0	0	0	0
thiosulfate_oxidation	0	0	0	0	0	0	0	0
FeFe_hydrogenase	0	0	0	0	0	0	0	0
Reverse_TCA	0	0	0	0	0	0	0	0
N2_fixation	0	0	0	0	0	0	0	0

Table S4.1. Summary of metabolism marker genes, detected with HMMs, present within reconstructed microbial genomes.

Conclusions

Microbial eukaryotes are important community members that have often been neglected in the field of microbial ecology. In order to ascertain a holistic understanding of our ecosystems and of our own evolutionary history, they cannot continue to be ignored. Foremost among the reasons for neglect, is the difficulty in studying microbial eukaryotes. In my work, I have focused on employing and improving culture-independent methods to study microbial eukaryotes in their natural community context, and hope to show that studying them in this way is critical for our understanding of their biology and the biology of their surrounding communities.

A primary conclusion of this work is that it is now possible to use genome-resolved metagenomic methods to assemble and bin eukaryotic genomes from complex communities. In Chapter 1, I present EukRep, a machine-learning based tool for classification of assembled scaffolds as either eukaryotic or prokaryotic. This is beneficial because it allows the use of eukaryotic gene predictors on eukaryotic scaffolds prior to binning, a limitation that previously severely impacted the ability to bin eukaryotic genomes. We then present a broader pipeline that incorporates EukRep and other publicly available tools to generate high quality eukaryotic gene predictions, and subsequently, eukaryotic genome bins.

Importantly, we've shown with this pipeline, it is possible to reconstruct a diversity of eukaryotic genomes from a broad range of environments, including from samples that had previously been analyzed and eukaryotic genomes missed. From Crystal Geyser groundwater samples, we assembled high quality, complete, fungal and arthropod genomes. From a thiocyanate reactor, we assembled a genomes for a Rhizarium and a Nucleariida, phylogenetically novel protists from phylums with poor genome representation and no close sequenced relatives. Among others, we also reconstructed genomes for yeasts, fungi, arthropods, nematodes, and protists from the infant gut, hospital room environment, and a complex co-culture of organisms from Mono Lake, demonstrating the broad versatility of our approach. Our pipeline represents a first attempt to systematically bin eukaryotes, in part to show its possible but also to inspire future improvements. In the time since its publication, additional tools have been published that incorporate EukRep and iterate on the idea of a Eukaryote binning pipeline, such as MetaEuk (Karin et al. 2020).

Our work characterized the broad presence of microbial eukaryotes in the infant gut and hospital environment. Fungal pathogens are known to have hospital reservoirs, however, much remains to be learned about where reservoirs of hospital-associated fungi are and how long strains persist in them. We detected novel, near identical *C. parapsilosis* genomes sequenced years apart in separate infants, suggesting transmission of members of a fungal population from reservoir to infant or infant to reservoir to infant. This result suggests the presence of a potentially diverse, hospital associated population of *C. parapsilosis* given the high incidence of *C. parapsilosis* infection in immune-compromised individuals. Hospital sinks were found to host a surprisingly diverse and variable eukaryotic community, and recovered genomes included an arthropod from the Diptera (true flies) and a nematode who's most closely related sequenced relatives included a bovine lungworm and *Caenorhabditis elegans.*

What prior work has been done with microbial eukaryotes has generally necessitated lab culture work, and thus, has not considered the impact of a surrounding microbial community on behavior and metabolism. Metagenomic techniques, combined with methods such as metatranscriptomics, and metaproteomics, afforded us the opportunity to study the behavior and metabolism of microbial eukaryotes in their natural community context. Metatranscriptomics of infant fecal samples containing *C. parapsilosis*, when compared to transcriptomic datasets from culture settings, supported our hypothesis that the microbiome context has a significant impact on the metabolism and behavior of *C. parapsilosis*. In particular, *C. parapsilosis* expression patterns were both highly divergent and highly variable over time compared to culture settings.

It is important to consider microbial eukaryotes in a community context not only for insight into their own behavior, but also for a more holistic understanding of microbial community behavior. In Chapter 3, we show expression patterns of *E. faecalis* and *S. epidermidis* were clearly distinguishable between infant fecal samples with and without *C. parapsilosis* present across multiple infants, indicating *C. parapsilosis* may have a particularly strong effect on the behavior of these bacteria in the human microbiome. In Chapter 4, the unique physiology of the choanoflagellate *S. monosierra,* ie. enveloping bacterial cells within its multicellular rosettes, may drive unique community dynamics. Whether the interactions between *S. monosierra* and bacterial community members are mutualistic is currently unknown, however bacterial genomes binned from metagenomic samples contained metabolisms such as carbon fixation, nitrogen fixation, nitrate reduction, arsenate reduction, and sulfide oxidation; all metabolisms potentially useful to *S. monosierra*. In addition, both a bacterial community member and *S. monosierra* itself contained biosynthetic gene clusters, potentially involved in inter-domain communication.

The work I present here represents first steps in incorporating Eukaryotes into broader, whole community microbial ecology studies and studying microbial eukaryote biology and behavior in a community context using genome-resolved methods; however, there are clear directions in which to continue. On the technical side, existing *ab initio* Eukaryotic gene prediction algorithms, such as GeneMark-ES and AUGUSTUS, have been designed for isolate sequencing experiments. This comes with evident caveats for use in a metagenomics context. For one, they require training on an individual genome, in part due to the difficulty of predicting eukaryotic genes without transcript evidence. Ideally, a gene prediction algorithm could be developed that functions similarly to meta-prodigal, and can predict genes on both eukaryotic and prokaryotic scaffolds without a need for genome-specific training.

An unexpected outcome of this work is that Eukaryotes appear to be relatively rare in shotgun metagenomes, more rare than barcode sequencing would suggest. I believe this sparsity is owed largely to technical limitations, as Eukaryotic genomes may be particularly difficult to assemble from metagenomes. Eukaryotic genomes can be extremely large, though the size varies greatly, relative to those of bacteria and archaea, thus requiring a much greater sequencing depth for reasonable genome coverage. In addition, eukaryotic cells are much larger than bacterial cells, and so eukaryotic genome copy number will be fewer for the same amount of biomass in a given sample, exacerbating the sequencing depth issue. Finally, eukaryotic genomes are often repeat rich, resulting in highly fragmented assemblies when reliant solely on short read sequencing. However, these challenges are reasons I remain excited about the current and future potential for

studying eukaryotes with genome-resolved metagenomic methods. Continued advances in sequencing throughput, long read sequencing, and DNA extraction methods will naturally address these limitations. I look forward to seeing the future results of applying eukaryotic genome resolved metagenomics and similar techniques to study lichens, sponges, and numerous other unexplored systems.

References

Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., … Mase, C. E. (2015). Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. Cell Syst 1(1):97-97 e3 https://doi.org/10.1016/j.cels.2015.07.006

Aguileta, G., Lengelle, J., Marthey, S., Chiapello, H., Rodolphe, F., Gendrault, A., … Giraud, T. (2010). Finding candidate genes under positive selection in Non-model species: Examples of genes involved in host specialization in pathogens. Molecular Ecology, 19(2), 292–306. https://doi.org/10.1111/j.1365-294X.2009.04454.x

Alegado RA, Brown LW, Cao S, Dermenjian RK, Zuzow R, Fairclough SR, Clardy J, King N. (2012). A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. eLife 1: e00013

Alexander, D. H., Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics, 12: 246

Aliaga S, Clark RH, Laughon M, Walsh TJ, Hope WW, Benjamin DK, et al. Changes in the incidence of candidiasis in neonatal intensive care units. Pediatrics. 2014;133:236–42.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomics contigs by coverage and composition. Nat Methods 11: 1144–1146.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, et al. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature Commun 7: 13219.

Anvar SY, Khachatryan L, Vermaat M, van Galen M, Pulyakhina I, Ariyurek Y, Kraaijeveld K, den Dunnen JT, de Knijff P, ‘t Hoen PAC, et al. 2014. Determining the quality and complexity of next-generation sequencing data without a reference genome. Genome Biol 15: 555.

Aramaki T., Blanc-Mathieu R., Endo H., Ohkubo K., Kanehisa M., Goto S., Ogata H. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics, btz859

Bachtiar, E. W., Dewiyani, S., Akbar, S. M. S., Bachtiar, B. M. (2016). Inhibition of Candida albicans biofilm development by unencapsulated Enterococcus faecalis cps2. Journal of Dental Sciences 11(3): 323-330 https://doi.org/10.1016/j.jds.2016.03.012

Baley JE, Kliegman RM, Boxerbaum B, Fanaroft AA. Fungal colonization in the very low birth weight infant. Pediatr. 1986;78:225–32.

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. Institute of Mathematical Statistics. 2001;29:1165–88.

Bennett RJ, Johnson AD. Completion of a parasexual cycle in Candida albicans by induced chromosome loss in tetraploid strains. EMBO J. 2003; 22(10):2505–15.

Bennett, R. J. (2015). The parasexual lifestyle of Candida albicans. Current Opinion in Microbiology, 28, 10–17. https://doi.org/10.1016/j.mib.2015.06.017
Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Research 47(W1): W81-W87
Bliss, J. M. (2015). Candida parapsilosis: An emerging pathogen developing its own identity. Virulence 6(2): 109-111. https://doi.org/10.1080/21505594.2015.1008897
Bokulich NA, Mills DA, Underwood MA. Surface microbes in the neonatal intensive care unit: changes with routine cleaning and over time. J Clin Microbiol. 2013;51:2617–24.
Boyd ES, Peters JW. (2013). New insights into the evolutionary history of biological nitrogen fixation. Front. Microbiol. 4(201)
Bray, N. L., Pimental, H., Melsted, P., Pachtor, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34: 525-527
Brooks B, Firek BA, Miller CS, Sharon I, Thomas BC, Baker R, et al. Microbes in the neonatal intensive care unit resemble those found in the gut of premature infants. Microbiome. 2014;2:1.
Brooks B, Olm MR, Firek BA, Baker R, Geller-McGrath D, Reimer SR, et al. The developing premature infant gut microbiome is a major factor shaping the microbiome of neonatal intensive care unit rooms. bioRxiv. 2018:315689 Available from: https://www.biorxiv.org/content/early/2018/05/07/315689. Cited 9 May 2018.
Brooks B, Olm MR, Firek BA, Baker R, Thomas BC, Morowitz MJ, et al. Strainresolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. Nat Commun. 2017;8:1814.
Brooks, B., Olm, M. R., Firek, B. A., Baker, R., Thomas, B. C., Morowitz, M. J., Banfield, J. F. (2017). Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. Nat Commun. 8:1814.
Brooks, B., Olm, M.R., Firek, B.A., Baker, R., Geller-McGrath, D., Reimer, S.R., Soenjoyo, K.R., Yip, J.S., Dahan, D., Thomas, B.C., et al. (2018). The developing premature infant gut microbiome is a major factor shaping the microbiome of neonatal intensive care unit rooms Microbiome, 6: 112
Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of replication rates in microbial communities. Nat Biotechnol 34: 1256–1263.
Brown CT, Xiong W, Olm MR, Thomas BC, Baker R, Firek B, et al. Hospitalized premature infants are colonized by related bacterial strains with distinct proteomic profiles. MBio, Am Soc Microbiol. 2018:9 Available from: https:// doi.org/10.1128/mBio.00441-18.
Brown, C.T., Xiong, W., Olm, M.R., Thomas, B.C., Baker, R., Firek, B., Morowitz, M.J., Hettich, R.L., and Banfield, J.F. (2018). Hospitalized Premature Infants Are Colonized by Related Bacterial Strains with Distinct Proteomic Profiles. MBio 9
Brown CT, Olm MR, Thomas BC, Banfield JF. (2016). Measurement of bacterial replication rates in microbial communities. Nat Biotechnol 34(12): 1256-1263
Burki F. (2014). The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb Perspect Biol 6(5): a016147
Bush RK, Portnoy JM. The role and abatement of fungal allergens in allergic diseases. J Allergy Clin Immunol. 2001;107:S430–40.
Butler G, Rasmussen MD, Lin MF, Santos MAS, Sakthikumar S, Munro CA, et al. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature. 2009;459:657–62.

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A. S., Sakthikumar, S., Munro, C. A., … Cuomo, C. A. (2009). Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature, 459(7247), 657–662. https://doi.org/10.1038/nature08064

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37: D233–D238.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. genome.cshlp.org. 2008;18:188–96.

Cantley, A. M., Woznica, A., Beemelmanns, C., King, N., Clardy, J. (2016). Isolation and Synthesis of a Bacterially Produced Inhibitor of Rosette Development in Choanoflagellates. J Am Chem Soc 138(13): 4326-4329 https://doi.org/10.1021/jacs.6b01190

Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. 2008. Protists are microbes too: a perspective. ISME J 3: 4–12.

Cavalheiro, M., Teixeira, M. C. (2018). Candida Biofilms: Threats, Challenges, and Promising Strategies. Front Med (Lausanne), 13(5): 28 https://doi.org/10.3389/fmed.2018.00028

Chang HJ, Miller HL, Watkins N, Arduino MJ, Ashford DA, Midgley G, et al. An epidemic of Malassezia pachydermatis in an intensive care nursery associated with colonization of health care workers' pet dogs. N Engl J Med. Mass Medical Soc. 1998;338:706–11.

Chang, H. J., Miller, H. L., Watkins, N., Arduino, M. J., Ashford, D. A., Midgley, G., … Jarvis, W. R. (1998). An epidemic of Malassezia pachydermatis in an intensive care nursery associated with colonization of helath care workers' pet dogs. N Engl J Med 338(11): 706-11 https://doi.org/10.1056/NEJM199803123381102

Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, et al. Geography and location are the primary drivers of office microbiome composition. Gilbert JA, editor. mSystems. 2016;1 Available from: http://msystems.asm.org/content/1/2/e00022-16.abstract.

Chen B, Snider RJ, Snider RM. 1996. Food consumption by Collembola from northern Michigan deciduous forest. Pedobiologia 40: 149–161.

Chen IA, Markowitz VM, Che K, Palaniappan K, Szeta E, Pillay M, Ratner A, Huang J, Anderson E, Huntemann M, et al. 2016. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res 45: D507–D516.

Chen T-A, Hill PB. The biology of Malassezia organisms and their ability to induce immune responses and skin disease. Vet Dermatol. 2005;16:4–26.

Chong, C. Y. L., Bloomfield, F. H., O'Sullivan, J. M. (2018). Factors Affecting Gastrointestinal Microbiome Development in Neonates. Nutrients, 10(3), 274 https://doi.org/10.3390/nu10030274

Clerihew, L., Lamagni, T. L., Brocklehurst, P., & McGuire, W. (2007). Candida parapsilosis infection in very low birthweight infants. Archives of Disease in Childhood: Fetal and Neonatal Edition, 92(2), 127–129. https://doi.org/10.1136/fnn.2006.097758

Cruz, M. R., Graham, C. E., Gagliano, B. C., Lorenz, M. C., & Garsin, D. A. (2013). Enterococcus faecalis Inhibits Hyphal Morphogenesis and Virulence of Candida albicans. Infection and Immunity 81(1), 189–200. https://doi.org/10.1128/IAI.00914-12

Cuvelier ML, Allen AE, Monier A, McCrow JP, Messié M, Tringe SG, Woyke T, Welsh RM, Ishoey T, Lee JH, et al. 2010. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. Proc Natl Acad Sci 107: 14679–14684.

Dawson TL Jr. Malassezia globosa and restricta: breakthrough understanding of the etiology and

treatment of dandruff and seborrheic dermatitis through whole-genome analysis. J Investig Dermatol Symp Proc. 2007;12:15–9.
Dedysh SN, Smirnova KV, Khmelenina VN, Suzina NE, Liesack W, Trotsenka YA. (2005). Methylotrophic autotrophy in Beijerinckia mobilis. J Bacreriol 187: 3884-3888
De la Cruz, J., Gómez-Herreros, F., Rodríguez-Galán, O., Begley, V., de la Cruz Muñoz-Centeno, M., Chávez, S. (2018). Feedback regulation of ribosome assembly. Curr Genet, 64(2): 393-404 https://doi.org/10.1007/s00294-017-0764-x
Desai, J. V., Lionakis, M. S. (2018). The role of neutrophils in host defense against invasive fungal infections. Curr Clin Microbiol Rep. 5(3): 181-189 https://dx.doi.org/10.1007%2Fs40588-018-0098-6
Dominguez, E. G., Zarnowski, R., Choy, H. L., Zhao, M., Sanchez, H., Nett, J. E., Andes, D. R. (2019). Conserved Role for Biofilm Matrix Polysaccharides in Candida auris Drug Resistance. mSphere 4(1): e00680-18 https://doi.org/10.1128/mSphereDirect.00680-18
Dröge J, Gregor I, McHardy AC. 2015. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. Bioinformatics 31: 817–824.
Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14: 755–763.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461.
El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The PFAM protein families database in 2019. Nucleic Acids Research 10.1093/nar/gky995
Emerson JB, Thomas BC, Alvarez W, Banfield JF. 2015. Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. Environ Microbiol 18: 1686–1703.
Faddeeva-Vakhrusheva A, Derks MF, Anvar SY, Agamennone V, Suring W, Smit S, van Straalen NM, Roelofs D. 2016. Gene family evolution reflects adaptation to soil environmental stressors in the genome of the collembolan Orchesella cincta. Genome Biol Evol. 8: 2106–2117.
Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39 Available from: https://doi. org/10.1093/nar/gkr367.
Fischer M, Pleiss J. 2003. The Lipase Engineering Database: a navigation and analysis tool for protein families. Nucleic Acids Res 31: 319–321.
Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martinez AT, Otillar R, Spatafora JW, Yadav JS, et al. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science 336: 1715–1719.
Forsberg, K., Woodworth, K., Walters, M., Berkow, E. L., Jackson, B., Chiller, T., Vallabhaneni S. (2019). Candida auris: The recent emergence of a multidrug-resistant fungal pathogen. Med Mycol 57(1): 1-12 https://doi.org/10.1093/mmy/myy054
Fox, E. P., Bui, C. K., Nett, J. E., Hartooni, N., Mui, M. C., Andes, D. R., … Johnson, A. D. (2015). An expanded regulatory network temporally controls Candida albicans biofilm formation. Molecular Microbiology, 96(6), 1226–1239. https://doi.org/10.1111/mmi.13002
Fridkin SK, Jarvis WR. Epidemiology of nosocomial fungal infections. Clin Microbiol Rev. 1996;9:499–511.

Fridkin, S. K., Kaufman, D., Edwards, J. R., Shetty, S., & Horan, T. (2006). Changing incidence of Candida bloodstream infections among NICU patients in the United States: 1995-2004. Pediatrics, 117(5), 1680–1687. https://doi.org/10.1542/peds.2005-1996

Fudal, I. (2012). Genome Structure and Reproductive Behaviour Influence the Evolutionary Potential of a Fungal Phytopathogen, 8(11). https://doi.org/10.1371/journal.ppat.1003020

Fujimura KE, Johnson CC, Ownby DR, Cox MJ, Brodie EL, Havstad SL, et al. Man's bestfriend? The effect of pet ownership on house dust microbial communities. J Allergy Clin Immunol. 2010;126:410–2 412.e1–3.

Fujimura KE, Sitarik AR, Havstad S, Lin DL, Levan S, Fadrosh D, et al. Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. Nat Med. 2016; Available from: http://www.nature. com/doifinder/10.1038/nm.4176.

Gaitanis G, Magiatis P, Hantschke M, Bassukas ID, Velegraki A. The Malassezia genus in skin and systemic diseases. Clin Microbiol Rev. 2012;25:106–41.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, et al. 2003. The genome sequence of the filamentous fungus Neurospora crassa. Nature 422: 859–868.

Garrison, E., Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907

Genuth, N. R., Barna, M. (2018). The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. Molecular Cell, 71(3), 364-374. https://doi.org/10.1016/j.molcel.2018.07.018Get

Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. (2018). Current understanding of the human microbiome. Nature Medicine 24: 392-400

Giraldo A, Sutton DA, Samerpitak K, de Hoog GS, Wiederhold NP, Guarro J, et al. Occurrence of Ochroconis and Verruconis species in clinical specimens from the United States. J Clin Microbiol. 2014;52:4189–201.

Glöckner G, Hülsmann N, Schleicher M, Noegel AA, Eichinger L, Gallinger C, Pawlowski J, Sierra R, Eureneuer U, Pillet L, et al. 2014. The genome of the foraminiferan Reticulomyxa filosa. Curr Biol 24: 11–18.

Gong, Y., Li, T., Yu, C., Sun, S. (2017). Candida albicans Heat Shock Proteins and Hsps-Associated Signaling Pathways as Potential Antifungal Targets. Frontiers in Cellular and Infection Microbiology 7, 520. https://doi.org/10.3389/fcimb.2017.00520

Gong J, Qing Y, Zou S, Fu R, Se L, Zhang X, Zhang Q. (2016). Protist-Bacteria Associations: Gammaproteobacteria and Alphaproteobacteria Are Prevalent as Digestion-Resistant Bacteria in Ciliated Protozoa. Fronteiers in Microbiology 7: 1664-302X

Gonia, S., Archambault, L., Shevik, M., Altendahl, M., Fellows, E., Bliss, J. M., Wheeler, R. T., Gale, C. A. (2017). Candida parapsilosis Protects Premature Intestinal Epithelial Cells from Invasion and Damage by Candida albicans. Front Pediatr 5: 54 https://doi.org/10.3389/fped.2017.00054

Grigoriev IV, Cullen D, Goodwin SB, Hibbett D, Jeffries TW, Kubicek CP, Kuske C, Magnuson JK, Martin F, Spatafora JW, et al. 2011. Fueling the future with fungal genomics. Mycology 2: 192–209.

Grünwald, N. J., McDonald, B. A., & Milgroom, M. G. (2016). Population Genomics of Fungal and Oomycete Pathogens. Annual Review of Phytopathology, 54(1), 323–346. https://doi.org/10.1146/annurev-phyto-080614-115913

Guida, A., Lindstädt, C., Maguire, S. L., Ding, C., Higgins, D. G., Corton, N. J., Berriman, M.,

Butler, G. (2011). Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast Candida parapsilosis. BMC Genomics 12: 628 https://doi.org/10.1186/1471-2164-12-628
Guimaraes, J. C., Zavolan, M. (2016). Patterns of ribosomal protein expression specify normal and malignant human cells. Genome Biology 17: 236 https://doi.org/10.1186/s13059-016-1104-z
Haider, S., & Pal, R. (2013). Integrated Analysis of Transcriptomic and Proteomic Data, 91–110.
Hake KH. (2019). The microbiome of a colonial choanoflagellates from Mono Lae, CA. UC Berkeley. ProQuest ID: Hake_berkeley_0028E_18830. Merritt ID: ark:/13030/m5b61nrn
Hewitt KM, Mannino FL, Gonzalez A, Chase JH, Caporaso JG, Knight R, et al. Bacterial diversity in two neonatal intensive care units (NICUs). Ravel J, editor. PLoS One. 2013;8:e54703.
Hibberd, D. J. (1975). Observations on the ultrastructure of the choanoflagellate Codosiga botrytis (Ehr.) Saville-Kent with special reference to the flagellar apparatus. J Cell Sci. 17(1): 191-219
Hirakawa MP, Martinez DA, Sakthikumar S, Anderson MZ, Berlin A, Gujja S, et al. Genetic and phenotypic intra-species variation in Candida albicans. Genome Res. 2015;25:413–25.
Hoffman MT, Arnold AE. 2010. Diverse bacteria inhabit living hyphae of phylogenetically diverse fungal endophytes. Appl Environ Microbiol 76: 4063–4075.
Holland, L. M., Schröder, M. S., Turner, S. A., Taff, H., Andes, D., Grózer, Z., … Butler, G. (2014). Comparative Phenotypic Analysis of the Major Fungal Pathogens Candida parapsilosis and Candida albicans. PLoS Pathogens, 10(9). https://doi.org/10.1371/journal.ppat.1004365
Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491.
Huang Y-C, Li C-C, Lin T-Y, Lien R-I, Chou Y-H, Wu J-L, et al. Association of fungal colonization and invasive disease in very low birth weight infants. Pediatr Infect Dis J. 1998;17:819–22.
Huang Y-C, Lin T-Y, Lien R-I, Chou Y-H, Kuo C-Y, Yang P-H, et al. Candidaemia in special care nurseries: comparison of Albicans and Parapsilosis infection. J Infect. 2000;40:171–5.
Huang, G., Srikantha, T., Sahni, N., Yi, S., Soll, D. R. (2009). CO2 Regulates White-to-Opaque Switching in Candida albicans. Curr Biol, 19(4): 330-334 https://dx.doi.org/10.1016%2Fj.cub.2009.01.018
Huang, Y. C., Li, C. C., Lin, T. Y., Chou, Y. H., Wu, J. L., Hsueh, C. (1998). Association of fungal colonization and invasive disease in very low birth weight infants. Pediatr Infect Dis J. 17(9): 819-22 https://doi.org/10.1097/00006454-199809000-00014
Huffnagle GB, Noverr MC. The emerging world of the fungal microbiome. Trends Microbiol. 2013;21:334–41.
Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life. Nature Microbiol 1: 16048.
Hull, C. M., Raisner, R. M., Johnson, A. D. (2000). Evidence for mating of the “asexual” yeast Candida albicans in a mammalian host. Science. American Association for the Advancement of Science. 289:307–10.
Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9:90–5.
Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation site

prediction in metagenomic sequences. Bioinformatics 28: 2223–2230.
Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Rohwer FL, Mylnikov AP, Keeling PJ. A New Lineage of Eukaryotes Illuminates Early Mitochondrial Genome Reduction. Curr Biol 27(23): 3717-3724
Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, et al. The diploid genome sequence of Candida albicans. Proc Natl Acad Sci U S A. National Acad Sciences. 2004;101:7329–34.
Jones E, Oliphant T, Peterson P. SciPy: open source scientific tools for Python. URL http://scipy. org. 2001.
Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44: D457–D462.
Kantor RS, Huddy RJ, Iyer R, Thomas BC, Brown CT, Anantharaman K, Tringe S, Hettich RL, Harrison STL, Banfield JF. 2017. Genome-resolved meta-omics ties microbial dynamics to process performance in biotechnology for thiocyanate degradation. Environ Sci Technol 51: 2944–2953.
Kantor RS, van Zyl AW, van Hille RP, Thomas BC, Harrison STL, Banfield JF. 2015. Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unraveled with genome-resolved metagenomics. Environ Microbiol 17: 4929–4941.
Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, Thomas BC, Banfield JF. 2013. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. mBio 4: e00708-13.
Karin EL, Mirdita M, Söding J. (2020). MetaEuk – Sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome 8: 48
Keeling PJ, Burki F. (2019). Progress towards the Tree of Eukaryotes. Current Biology. 29(16): PR808-R817
Kim J, Sudbery P. 2011. Candida albicans, a major human fungal pathogen. J Microbial 49: 171–177.
Kim KT, Jeon J, Choi J, Cheong K, Song H, Choi G, Kang S, Lee YH. 2016. Kingdom-wide analysis of fungal small secreted proteins (SSPs) reveals their potential role in host association. Front Plant Sci 7: 186.
King N, Westbrook MJ, Young SL, Kuo A, Abedin M, … Rokhsar D. (2008). The genome of the choanoflagellate Monosiga brevicollis and the origin of the metazoans. Nature 451: 783-788
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5: 59.
Kothavade RJ, Kura MM, Valand AG, Panthaki MH. Candida tropicalis: its prevalence, pathogenicity and increasing resistance to fluconazole. J Med Microbiol. 2010;59:873–80.
Kozlowski LP. (2016). IPC – Isoelectric Point Calculator. Biol Direct 11(1): 55
Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567–580.
Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
Kuhn, D.M., Mikherjee P.K., Clark, T.A., Pujol, C., Chandra, J., Hajjeh R. A., Warnock, D. W., Soil, D. R., Ghannoum M. A. (2004). Candida parapsilosis characterization in an outbreak

setting. Emerg Infect Dis 10(6): 1074-81 https://doi.org/10.3201/eid1006.030873
Lachke SA, Lockhart SR, Daniels KJ, Soll DR. Skin facilitates Candida albicansmating. Infect Immun. 2003;71:4970–6.
Laforest-Lapointe I, Arrieta M-C. Microbial eukaryotes: a missing link in gut microbiome studies. mSystems. 2018;3:e00201–17.
Langmead, B., Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4): 357-359
LaTuga MS, Ellis JC, Cotton CM, Goldberg RN, Wynn JL, Jackson RB, et al. Beyond bacteria: a study of the enteric microbial consortium in extremely low birth weight infants. Driks A, editor. PLoS One. 2011;6:e27858.
Lax S, Sangwan N, Smith D, Larsen P, Handley KM, Richardson M, et al. Bacterial colonization and succession in a newly opened hospital. Sci Transl Med. 2017;9 Available from: http://stm.sciencemag.org/content/9/391/eaah6500.abstract.
Lax G, Eglit Y, Eme L, Bertrand EM, Roger AJ, Simpson AGB. (2018) Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. Nature 564: 410-414
Lee, T. H., Guo, H., Wang, X., Kim, C., Paterson, A. H. (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics, 15(1)
Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. academic.oup.com. 2007;23:127–8.
Levin TC, King N. (2013). Evidence for sex and recombination in the choanoflagellate Salpingoeca rosetta. Curr Biol 23(21): 2176-80
Liu H, Xin Y, Xun L. (2014). Distribution, Diversity, and Activities of Sulfur Dioxygenases in Heterotrophic Bacteria. Applied and Environmental Microbiology 80(5): 1799-1806
Love, M. I., Huber, W., Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15: 550
Luangsa-ard J, Houbraken J, van Doorn T, Hong S-B, Borman AM, HywelJones NL, et al. Purpureocillium, a new genus for the medically important Paecilomyces lilacinus: Purpureocillium, a new fungal genus for P. lilacinus. FEMS Microbiol Lett. 2011;321:141–9.
Luikart, G. (2014). Recent novel approaches for population genomics data analysis, 1661–1667.
Madden AA, Epps MJ, Fukami T, Irwin RE, Sheppard J, Sorger DM, et al. The ecology of insect-yeast relationships and its relevance to human industry. Proc Biol Sci. 2018;285 Available from: https://doi.org/10.1098/rspb.2017. 2733. rspb.royalsocietypublishing.org.
Magwene, P. M., Kayıkçı, Ö., Granek, J. A., Reininga, J. M., Scholl, Z., Murray, D. (2011). Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in Saccharomyces cerevisiae. PNAS 108(5) 1987-1992 https://doi.org/10.1073/pnas.1012544108
Mangot J, Logares R, Sánchez P, Latorre F, Seeleuthner Y, Mondy S, Sieracki ME, Jaillon O, Wincker P, Vargas C, et al. 2017. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. Sci Rep 7: 41498.
Manzoni P, Mostert M, Castagnola E. Update on the management of Candida infections in preterm neonates. Arch Dis Child Fetal Neonatal Ed. fn.bmj.com. 2015;100:F454–9.
Margarido GRA, Heckerman D. 2015. ConPADE: genome assembly ploidy estimation from next-generation sequencing data. PLoS Comput Biol 11: e1004229.
Mason, K. L., Downward, R. E., Falkowski, N. R., Young, V. B., Kao, J. Y., & Huffnagle, G. B. (2012). Interplay between the Gastric Bacterial Microbiota and Candida albicans during

Postantibiotic Recolonization and Gastritis, 150–158. https://doi.org/10.1128/IAI.05162-11
Mavor, A. L., Thewes, S., Hube, B. (2005). Systemic fungal infections caused by Candida species: epidemiology, infection process and virulence attributes. Curr Drug Targets 6(8): 863-74 https://doi.org/10.2174/138945005774912735
Meng, Q., Zhang, T., Wei, W., Mu, W., Miao, M. (2017). Production of Mannitol from a High Concentration of Glucose by Candida parapsilosis SK26.001. Appl Biochem Biotechnol 181(1):391-406 http://doi.org/10.1007/s12010-016-2219-0
Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-laylin LK, Maréchal-Drouard L, et al. 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318: 245–250.
Mesquita-Rocha S, Godoy-Martinez PC, Gonçalves SS, Urrutia MD, Carlesse F, Seber A, et al. The water supply system as a potential source of fungal infection in paediatric haematopoietic stem cell units. BMC Infect Dis. bmcinfectdis.biomedcentral.com. 2013;13:289.
Mesquite-Rocha, S., Godoy-Martinez, P. C., Gonçalves, S. S., Urrutia, M. D., Carlesse,F., Seber, A., … Colombo, A. L. (2013). The water supply system as a potential source of fungal infection in paediatric haematopoietic stem cell units. BMC Infec Dis 13:289 https://dx.doi.org/10.1186%2F1471-2334-13-289
Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In Proceedings of the Gateway Computing Environments Workshop (GCE), pp. 1–8, New Orleans, LA.
Miller, M. G., Johnson, A. D. (2002). White-opaque switching in Candida albicans is controlled by mating-type locus homeodomain proteins and allows efficient mating. Cell 110(3): 293-302 https://doi.org/10.1016/s0092-8674(02)00837-1
Min B, Park JH, Park H, Shin HD, Choi IG. 2017. Genome analysis of a zygomycete fungus Choanephora cucurbitarum elucidates necrotrophic features including bacterial genes related to plant colonization. Sci Rep 7: 40432.
Mistry J, Bateman A, Finn RD. 2007. Predicting active site residue annotations in the Pfam database. BMC Bioinformatics 8: 298.
Monier A, Welsh RM, Gentemann C, Weinstock G, Sodergren E, Armbrust EV, Eisen JA, Worden AZ. 2012. Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. Environ Microbiol 14: 162–176.
Morales-Cruz A, Amrine KC, Blanco-Ulate B, Lawrence DP, Travadon R, Rolshausen PE, Baumgartner K, Cantu D. 2015. Distinctive expansion of gene families associated with plant cell wall degradation, secondary metabolism, and nutrient uptake in the genomes of grapevine trunk pathogens. BMC Genomics 16: 469.
Mosier AC, Miller CS, Frischkorn KR, Ohm RA, Li Z, LaButti K, Lapidus A, Lipzen A, Chen C, Johnson J, et al. 2016. Fungi contribute critical but spatially varying roles in nitrogen and carbon cycling in acid mine drainage. Front Microbiol 7: 238.
Naglik, J. R., Fidel Jr., P. L., Odds, F. C. (2008). Animal models of mucosal Candida infection. FEMS Microbiology Letters 283(2): 129-139 https://doi.org/10.1111/j.1574-6968.2008.01160.x
Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. 2015. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 43(Database issue): D130–D137.
Nawrocki EP. 2009. “Structural RNA homology search and alignment using covariance models.”

PhD dissertation, Washington University, St. Louis, MO.
NCBI Resource Coordinators. 2017. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 45: D12–D17. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, Condon BJ, Copeland AC, Dhillon B, Glaser F. 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen dothideomycetes fungi. PLoS Pathogens 8: e1003037.
NCBI Resource Coordinators. Database resources of the national center for biotechnology information. Nucleic Acids Res. ncbi.nlm.nih.gov. 2017;45:D12–7.
Nobile, C. J., & Johnson, A. D. (2015). Candida albicans Biofilms and Human Disease. Annu Rev Microbiol 69: 71–92. https://doi.org/10.1146/annurev-micro-091014-104330.Candida
Nobile, C. J., Fox, E. P., Nett, J. E., Sorrells, T. R., Mitrovich, Q. M., Herndav, A. D., … Johnson, A. D. (2012). A recently evolved transcriptional network controls biofilm development in Candida albicans. Cell, 148(1–2), 126–138. https://doi.org/10.1016/j.cell.2011.10.048
Oberauner L, Zachow C, Lackner S, Högenauer C, Smolle K-H, Berg G. The ignored diversity: complex bacterial communities in intensive care units revealed by 16S pyrosequencing. Sci Rep. 2013;3:1413.
Oh J, Byrd AL, Deming C, Conlan S, Barnabas B, Blakesley R, et al. Biogeography and individuality shape function in the human skin metagenome. Nature. 2014;514:59–64.
Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genome de-replication that enables tracking of microbial genotypes and improved genome recovery from metagenomes. ISME J 11: 2864–2868.
Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017; Available from: https:// doi.org/10.1038/ismej.2017.126.
Olm MR, Brown CT, Brooks B, Firek B, Baker R, Burstein D, et al. Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. Genome Res. 2017; 27(4):601–12.
Olm MR, Butterfield CN, Copeland A, Boles TC, Thomas BC, Banfield JF. Thesource and evolutionary history of a microbial contaminant identifiedthrough soil metagenomic analysis. Brown CT, Newman DK, editors. MBio.2017;8:e01969–16.
Olm, M. R., Bhattacharya, N., Crits-Christoph, A., Firek, B. A., Baker, R., Song, Y. S., Morowitz, M. J., Banfield, J. F. (2019). Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria. Science Advances 5(12): eaax5727 https://doi.org/10.1126/sciadv.aax5727
Olm, M. R., Butterfield, C. N., Copeland, A., Boles, T. C., Thomas, B. C., Banfield, J. F. (2017). The source and evolutionary history of a microbial contaminant identified through soil metagenomic analysis. Brown CT, Newman DK, editors. MBio, 8:e01969–16.
Olm, M. R., West, P. T., Brooks, B., Firek B. A., Baker, R., Morowitz, M. J., Banfield, J. F. (2019). Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. Microbiome 7(1): 26 https://doi.org/10.1186/s40168-019-0638-1
Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17 Available from: http://genomebiology. biomedcentral.com/articles/10.1186/s13059-016-0997-x.

Ott SJ, Kühbacher T, Musfeldt M, Rosenstiel P, Hellmig S, Rehman A, et al. Fungi and inflammatory bowel diseases: alterations of composition and diversity. Scand J Gastroenterol. 2008;43:831–41.
Ovchinnikov S, Park H, Varghese N, Huang P, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides N, Baker D. 2017. Protein structure determination using metagenome sequence data. Science 355: 294–298.
Pajic, P., Pavlidis, P, Dean, K., Neznanova, L., Romano, R., Garneau, D., … Ruhl, S. (2019). Independent amylase gene copy number bursts correlate with dietary preferences in mammals. eLife 8: e44628 https://doi.org/10.7554/eLife.44628
Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Marrison HG, Sogin ML, Patterson DJ, Katz LA. 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst Biol 59: 518–533.
Parfrey, L. W., Walters, W. A., & Knight, R. (2011). Microbial eukaryotes in the human microbiome : ecology , evolution , and future directions, 2(July), 1–6. https://doi.org/10.3389/fmicb.2011.00153
Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle J, Dunthorn M, et al. 2012. CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal plant and fungal kingdoms. PLoS Biol 10: e1001419.
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in python. JMLR 12: 2825–2830.
Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomics sequence data with highly uneven depth. Bioinformatics 28: 1420–1428.
Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.
Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature. nature.com. 2018;556:339–44.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8: 785–786.
Pfaller MA. Nosocomial candidiasis: emerging species, reservoirs, and modes of transmission. Clin Infect Dis. academic.oup.com. 1996;22(Suppl 2):S89–94.
Pita L, RixL, Slaby BM, Franke A, Hentschel U. (2018). The sponge holobiont in a changing ocean: from microbes to ecosystems. Microbiome 6: 46
Plissonneau, C., & Stürchler, A. (2016). The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat, 7(5). https://doi.org/10.1128/mBio.01231-16.Editor
Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartmann, F. E., & Croll, D. (2017). Using Population and Comparative Genomics to Understand the Genetic Basis of Effector-Driven Fungal Pathogen Evolution. Frontiers in Plant Science, 8(February), 1–15. https://doi.org/10.3389/fpls.2017.00119
Porteous NB, Grooters AM, Redding SW, Thompson EH, Rinaldi MG, De Hoog GS, et al. Identification of Exophiala mesophila isolated from treated dental unit waterlines. J Clin Microbiol. 2003;41:3885–9.
Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, Hug LA, Burstein D,

Emerson JB, Thomas BC, et al. 2016. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations. Environ Microbiol 19: 459–474.
Probst AJ, Ladd B, Jarett JK, Sieber CMK, Emerson JB, Thomas BC, Stieglemier M, Kling A, Woyke T, Ryan MC, et al. 2018. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nat Microbiol 3: 328–336.
Probst AJ, Weinmaier T, Raymann K, Perras A, Emerson JB, Rattea T, Wanner G, Klingl A, Berg IA, Yoshinaga M. 2014. Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. Nat Commun 5: 5497.
Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 2001;29:137–40.
Pryszcz, L. P., Németh, T., Gácser, A., Gabaldón, T. (2013). Unexpected Genomic Variability in Clinical and Environmental Strains of the Pathogenic Yeast Candida parapsilosis. Genome Biol Evol. 5(12): 2382-2392 https://doi.org/10.1093/gbe/evt185
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. American J of Human Genetics, 81
Quandt CA, Kohler A, Hesse CN, Sharpton TJ, Martin F, Spatafora JW. 2015. Metagenome sequence of Elaphomyces granulatus from sporocarp tissue reveals Ascomycota ectomycorrhizal fingerprints of genome expansion and a Proteobacteria-rich microbiome. Environ Microbiol 17: 2952–2968.
Quiloan, M. L. G., Vu, J., Carvalho, J. (2012). Enterococcus faecalis can be distinguished from Enterococcus faecium via differential susceptibility to antibiotics and growth and fermentation characteristics on mannitol salt agar. Frontiers in Biology, 7 167-177. https://doi.org/10.1007/s11515-012-1183-5
Quinlan, A. R., Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6): 841-842
R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Rahman SF, Olm MR, Morowitz MJ, Banfield JF. Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. mSystems. 2018;3:e00123–17.
Rahman, S.F., Olm, M.R., Morowitz, M.J., and Banfield, J.F. (2018). Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. mSystems 3: e00123–17
Ramirez-Zavala, B., Reuss, O., Park, Y. N., Ohlsen, K., Morschhäuser, J. (2008). Environmental induction of white-opaque switching in Candida albicans. PLoS Pathog 4(6): e1000089 https://doi.org/10.1371/journal.ppat.1000089
Ratnasingham S, Hebert PDN. BOLD: the barcode of life data system. Mol Ecol Notes. 2007;7:355–64 Wiley Online Library. (http://www.barcodinglife.org).
Raveh-Sadka T, Firek B, Sharon I, Baker R, Brown CT, Thomas BC, et al. Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants. ISME J. 2016; Available from: http://www.nature.com/doifinder/10.1038/ismej.2016.83.
Raveh-Sadka T, Firek B, Sharon I, Beker R, Brown CT, Thomas BC, Morowitz MJ, Banfield JF.

2016. Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants. ISME J 10: 2817–2830.
Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, Sharon I, Baker R, Good M, Morowitz MJ, et al. 2015. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. eLife 4: e05477.
Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, et al. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. Kolter R, editor. Elife. 2015;4:e05477.
Rawlings ND, Barrett AJ, Finn RD. 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 44: D343–D350.
Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A., Mesirov, J. P. (2017). Variant Review with the Integrative Genomics Viewer (IGV). Cancer Research, 77(21): 31-34
Rohart, F., Gautier, B., Singh, A., Cao, K. L. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Computational Biology, https://doi.org/10.1371/journal.pcbi.1005752
Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22: 1337–1344.
Rosenberg, S. M. (2011). Stress-Induced Loss of Heterozygosity in Candida : a Possible Missing Link in the Ability to Evolve, 2(5), 3–6. https://doi.org/10.1128/mBio.00200-11.Copyright
Roy RS, Price DC, Schliep A, Cai G, Korobeynikov A, Yoon HS, Yang EC, Bhattacharya D. 2014. Single cell genome analysis of an uncultured heterotrophic stramenopile. Sci Rep 4: 4780.
Ruan S-Y, Chien J-Y, Hsueh P-R. Invasive trichosporonosis caused by Trichosporon asahii and other unusual Trichosporon species at a medical center in Taiwan. Clin Infect Dis. 2009;49:e11–7.
Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.
Sanchez V, Vazquez JA, Barth-Jones D, Dembry L, Sobel JD, Zervos MJ. Epidemiology of nosocomial acquisition of Candida lusitaniae. J Clin Microbiol. 1992;30:3005–8.
Schulze J, Sonnenborn U. Yeasts in the gut: from commensals to infectious agents. Dtsch Ärztebl Int. 2009;106:837.
Scorzoni, L., de Paula e Silva, A. C. A., Marcos, C. M., Assato, P. A., de Melo, W. C. M. A., de Oliveira, H. C., … Fusco-Almeida, A. M. (2017). Antifungal therapy: New advances in the understanding and treatment of mycosis. Frontiers in Microbiology, 8(JAN), 1–23. https://doi.org/10.3389/fmicb.2017.00036
Seenivasan R, Sausen N, Medlin LK, Melkonian M. (2013). Picomonas judraskeda Gen. Et Sp. Nov.: The First Identified Member of the Picozoa Phylum Nov., a Widespread Group of Picoeukaryotes, Formerly Known as 'Picobiliphytes'. PLoS ONE 8(3): e59565
Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res 23: 111–120.
Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013;23:111–20.
Sher, Y., Olm, M. R., Raveh-Sadka, T., Brown C. T., Sher, R., Firek, B., Baker, R., Morowitz,

M. J., Banfield, J. F. (2020). Combined analysis of microbial metagenomic and metatranscriptomic sequencing data to assess in situ physiological conditions in the premature infant gut. PLoS ONE 15(3): e0229537 https://doi.org/10.1371/journal.pone.0229537

Shi, Z., Fujii, K., Kovary, K. M., Genuth N. R., Röst, H. L., Teruel, M. N., Barna, M. (2017). Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide. Mol Cell 67(1): 71-83 https://doi.org/10.1016/j.molcel.2017.05.021

Shin H, Pei Z, Martinez KA, Rivera-Vinas JI, Mendez K, Cavallin H, et al. The first microbial environment of infants born by C-section: the operating room microbes. Microbiome. 2015;3 Available from: http://www. microbiomejournal.com/content/3/1/59.

Shivaprasad A, Ravi GC, Shivapriya, Rama. A rare case of nasal septal perforation due to Purpureocillium lilacinum: case report and review. Indian J Otolaryngol Head Neck Surg. Springer. 2013;65:184–8.

Silva, S., Negri, M., Henriques, M., Oliveria, R., Williams, D. W., Azeredo, J. (2012). Candida glabrata. Candida parapsilosis, and Candida tropicalis: biology, epidemiology, pathogenicity, and antifungal resistance. FEMS Microbiol Rev 36(2): 288-305 https://doi.org/10.1111/j.1574-6976.2011.00278.x

Simão FA, Waterhouse RM, Loannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. (2017). PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Research 45(W1): W49-W54

Slutsky, B., Staebell M., Anderson, J., Risen, L., Pfaller, M., Soll, D. R. (1987). "White-opaque transition": a second high -frequency switching system in Candida albicans. J Bacteriol 169(1): 189-97 https://doi.org/10.1128/jb.169.1.189-197.1987

Soanes, D. M., Alam, I., Cornell, M., Wong, H. M., Hedeler, C., Norman, W., … Talbot, N. J. (2008). Comparative Genome Analysis of Filamentous Fungi Reveals Gene Family Expansions Associated with Fungal Pathogenesis, 3(6). https://doi.org/10.1371/journal.pone.0002300

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis andpost-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7: 62.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34:W435–9.

Stewart CJ, Marrs ECL, Magorrian S, Nelson A, Lanyon C, Perry JD, et al. The preterm gut microbiota: changes associated with necrotizing enterocolitis and infection. Acta Paediatr. 2012;101:1121–7.

Stewart CJ, Nelson A, Scribbins D, Marrs ECL, Lanyon C, Perry JD, et al. Bacterial and fungal viability in the preterm gut: NEC and sepsis. Arch Dis Child Fetal Neonatal Ed. 2013;98:F298–303.

Surawicz CM, Elmer GW, Speelman P, McFarland LV, Chinn J, van Belle G. Prevention of antibiotic-associated diarrhea by Saccharomyces boulardii: a prospective study. Gastroenterology. gastrojournal.org. 1989;96:981–8.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23: 1282–1288.
Tamburini S, Shen N, Wu HC, Clemente JC. The microbiome in early life: implications for health outcomes. Nat Med. 2016;22:713–22.
Tavanti, A., Gow, N. A. R., Maiden, M. C. J., Odds, F. C., & Shaw, D. J. (2004). Genetic evidence for recombination in Candida albicans based on haplotype analysis. Fungal Genetics and Biology, 41(5), 553–562. https://doi.org/10.1016/j.fgb.2003.12.008
Taylor MW, Radax R, Steger D, Wagner M. (2007). Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. Microbiol Mol Biol Rev 71(2): 295-347
Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 18: 1979–1990.
Thacker RW, Freeman CJ. (2012). Sponge-microbe symbioses: recent advances and new directions. Adv Mar Biol 62: 57-111
The UniProt Consortium. 2017. Uniprot: the universal protein knowledgebase. Nucleic Acids Res 45: D158–D169.
Tóth, R., Cabral, V., Thuer, E., Bohner, F., Németh, T., Papp, C., … Gácser, A. (2018). Investigation of Candida parapsilosis virulence regulatory factors during host-pathogen interaction. Scientific Reports, 8(1), 1–14. https://doi.org/10.1038/s41598-018-19453-4
Trofa D, Gácser A, Nosanchuk JD. 2008. Candida parapsilosis, an emerging fungal pathogen. Clin Microbiol Rev 21: 606–625.
Tsai Y-C, Conlan S, Deming C, NISC Comparative Sequencing Program, Segre JA, Kong HH, et al. Resolving the complexity of human skin metagenomes using single-molecule sequencing. MBio. 2016;7:e01948–15.
Tuch, B. B., Mitrovich, Q. M., Homann, O. R., Hernday, A. D., Monighetti, C. K., de La Vega, F. M., & Johnson, A. D. (2010). The transcriptomes of two heritable cell types illuminate the circuit governing their differentiation. PLoS Genetics, 6(8). https://doi.org/10.1371/journal.pgen.1001070
Vaulot D, Lepère C, Toulza E, De la Iglesia R, Poulain J, Gaboyer F, Moreau H, Vandepoele K, Ulloa O, Gavory F, et al. 2012. Metagenomes of the picoalga Bathycoccus from the Chile coastal upwelling. PLoS One 7: e39648.
Vazquez JA, Sanchez V, Dmuchowski C, Dembry LM, Sobel JD, Zervos MJ. Nosocomial acquisition of Candida albicans: an epidemiologic study. J Infect Dis. academic.oup.com. 1993;168:195–201.
Vervier K, Mahe P, Tournoud M, Veyrieras J, Vert J. 2016. Large-scale machine learning for metagenomics sequence classification. Bioinformatics 32: 1023–1032.
West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. Genomereconstruction for eukaryotes from complex natural microbial communities. Genome Res. genome.cshlp.org. 2018;28:569–80.
West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., Banfield, J. F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. Genome Res 28(4) 569-580 https://doi.org/10.1101/gr.228429.117
Whaley, S. G., Berkow, E. L., Rybak, J. M., Nishimoto, A. T., Barker, K. S., Rogers, P. D. (2017). Azole Antifungal Resistance in Candida albicans and Emerging Non-albicans Candida Species. Frontiers in Microbiology, 7, 2173. https://doi.org/10.3389/fmicb.2016.02173

Whaley, S. G., Tsao, S., Weber, S., Zhang, Q., Barker, K. S., Raymond, M., Rogers, P. D. (2016). The RTA3 Gene, Encoding a Putative Lipid Translocase, Influences the Susceptibility of Candida albicans to Fluconazole. Antimicrobial Agents and Chemotherapy 60(10): 6060-6066 https://doi.org/10.1128/AAC.00732-16

Wildschutte H, Wolfe DM, Tamewitz A, Lawrence JG. Protozoan predation, diversifying selection, and the evolution of antigenic diversity in Salmonella. Proc Natl Acad Sci U S A. 2004;101:10644–9.

Williams TA. (2014). Evolution: Rooting the Eukaryotic Tree of Life. Current Biology. 24(4): R151-R152

Wilson MR, O'Donovan BD, Gelfand JM, Sample HA, Chow FC, Betjemann JP, et al. Chronic meningitis investigated via metagenomic next-generation sequencing. JAMA Neurol. 2018;75:947–55.

Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, Lapidus A, Wu D, McCutcheon JP, McDonals BR,et al. 2010. One bacterial cell, one complete genome. PLoS One 5: e10314.

Woznica, A., Gerdt, J. P., Hulett, R. E., Clardy, J., King, N. (2017). Mating in the Closest Living Relatives of Animals Is Induced by a Bacterial Chondroitinase. Cell 170: 1175-1183 https://doi.org/10.1016/j.cell.2017.08.005

Xu, J. (2006). Fundamentals of Fungal Molecular Population Genetic Analyses. Current Issues in Molecular Biology, (July), 75–90. Retrieved from http://www.horizonpress.com/cimb/v/v8/06.pdf

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40: W445–W451.

Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S, Bhattacharya D. 2011. Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332: 714–717.

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Backstrom D, Juzokaite L, … Ettema TJ. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541(7637): 353-358

Zhang N, O'Donnell K, Sutton DA, Nalim FA, Summerbell RC, Padhye AA, et al. Members of the Fusarium solani species complex that cause infections in both humans and plants are common in the environment. J Clin Microbiol. 2006;44:2186–90.

Zhou, X., Liao, W. J., Liao, J. M., Liao, P., Lu, H. (2015). Ribosomal proteins: functions beyond the ribosome. J Mol Cell Biol, 7(2): 92-104 https://doi.org/10.1093/jmcb/mjv014

Zhu, Y. O., Siegal, M. L., Hall, D. W., & Petrov, D. A. (2014). Precise estimates of mutation rate and spectrum in yeast. Proceedings of the National Academy of Sciences, 111(22), E2310–E2318. https://doi.org/10.1073/pnas.1323011111