Accurate and sensitive detection of microbial eukaryotes from metagenomic shotgun sequencing data

Abigail L. Lind[1] and Katherine S. Pollard[1,2,3,*]

1. Gladstone Institute of Data Science and Biotechnology, San Francisco, CA.
2. Institute for Human Genetics, Department of Epidemiology and Biostatistics, and Institute for Computational Health Sciences, University of California, San Francisco, CA.
3. Chan-Zuckerberg BioHub, San Francisco, CA.

* Corresponding author: katherine.pollard@gladstone.ucsf.edu

Abstract

Microbial eukaryotes are found alongside bacteria and archaea in natural microbial systems, including host-associated microbiomes. While microbial eukaryotes are critical to these communities, they are often not included in metagenomic analyses. Here we present EukDetect, a bioinformatics approach based on universal eukaryotic marker genes, to identify eukaryotes in shotgun metagenomic sequencing data. EukDetect is accurate, sensitive, has a broad taxonomic coverage of microbial eukaryotes, and is resilient against bacterial contamination in eukaryotic genomes. This enables us to identify and make observations about eukaryotes in public human and plant microbiome datasets. Using EukDetect, we describe the spatial distribution of eukaryotes in the human gut, finding that fungi and protists are present in the lumen and mucosa throughout the large intestine. We find that there is a succession of eukaryotes that colonize the human gut during the first years of life, similar to patterns of succession observed in bacteria in the developing gut. By comparing DNA and RNA sequencing of

paired samples from the human gut, we find that many eukaryotes continue active transcription after passage through the gut, while others do not, suggesting they are dormant or nonviable. Finally, we identify eukaryotes in *Arabidopsis* leaf samples, many of which are not identifiable from public protein databases. EukDetect is an accurate and sensitive pipeline to routinely characterize eukaryotes in shotgun sequencing datasets from diverse microbiomes.

Introduction

Eukaryotic microbes are ubiquitous in microbial systems, where they function as decomposers, predators, parasites, and producers (Bik et al. 2012). Eukaryotic microbes inhabit virtually every environment on earth, from extremophiles found in geothermal vents, to endophytic fungi growing within the leaves of plants, to parasites inside the animal gut. Microbial eukaryotes have complex interactions with their hosts in both plant and animal associated microbiomes. In plants, microbial eukaryotes assist with nutrient uptake and can protect against herbivory (Rodriguez et al. 2009). In animals, microbial eukaryotes in the gastrointestinal tract can metabolize plant compounds (Akin and Borneman 1990). Microbial eukaryotes can also cause disease in both plants and animals, as in the case of oomycete pathogens in plants or cryptosporidiosis in humans (Kamoun et al. 2015; Haque 2007). Outside of host associated environments, microbial eukaryotes are integral to the ecology of marine and soil ecosystems, where they are primary producers, partners in symbioses, decomposers, and predators (Caron et al. 2017; Brussaard et al. 2007).

Eukaryotic species are often not captured in studies of microbiomes (Hernández-Santos and Klein 2017). As eukaryotic species do not have a 16S ribosomal RNA gene, 16S amplicon sequencing cannot detect them. In studies specifically targeting eukaryotes, amplicon sequencing of eukaryotic-specific (18S) or fungal-specific (ITS) genes are effective, but can have limited taxonomic resolution and often cannot disambiguate closely related species or strains (Schoch et al. 2012; Lücking et al. 2020). 18S or ITS sequencing is also performed less frequently than global shotgun methods of microbiome sequencing. Whole genome shotgun sequencing captures fragments of DNA from the entire pool of species present from a microbiome, including eukaryotes. However, eukaryotic species are estimated to be present in many environments at lower abundances than bacteria, and comprise a much smaller fraction of the reads of a shotgun sequencing library than do bacteria (Qin et al. 2010; Bik et al. 2012). As interest in detecting rare taxa and rare variants in microbiomes has increased alongside falling sequencing costs, more and more microbiome deep sequencing datasets are generated where eukaryotes could be detected and analyzed in a meaningful way.

Apart from sequencing depth, a second major barrier to detecting eukaryotes from shotgun sequencing is the availability of methodological tools. Research questions about eukaryotes in microbiomes have primarily been addressed by using eukaryotic reference genomes, genes, or proteins for either *k*-mer matching or read mapping approaches (Nash et al. 2017; Chehoud et al. 2015). However, databases of eukaryotic genomes and proteins have widespread contamination from bacterial sequences, and these methods therefore frequently misattribute bacterial reads to eukaryotic species

(Lu and Salzberg 2018; Steinegger and Salzberg 2020). Some gene-based taxonomic profilers, such as Metaphlan2, can detect eukaryotic species, but these target a small subset of microbial eukaryotes (122 eukaryotes detectable at the species level as of the mpa_v30 release) (Truong et al. 2015).

To enable detection of eukaryotic species from any environment sequenced with whole genome shotgun metagenomics, we have developed EukDetect, an alignment-based approach that uses universally conserved marker genes found in all fungi, protists, roundworms, and flatworms with publicly available sequenced genomes in NCBI GenBank. This gene-based strategy avoids the pitfall of spurious bacterial contamination in eukaryotic genomes, which has confounded other approaches. The EukDetect pipeline will enable different fields using shotgun metagenomics to address emerging questions about microbial eukaryotes and their roles in human, other host-associated, and environmental microbiomes.

We apply the EukDetect pipeline to public human and plant associated microbiome datasets and make inferences about the roles of eukaryotes in these environments. We show that our marker gene approach greatly expands the number of detectable disease-relevant eukaryotic species in both human and plant associated microbiomes.

Results

Bacterial sequences are ubiquitous in eukaryotic genomes

Eukaryotic reference genomes and proteomes have many sequences that are derived from bacteria, which have entered these genomes either spuriously through

contamination during sequencing and assembly, or represent true biology of horizontal transfer from bacteria to eukaryotes (Lu and Salzberg 2018; Steinegger and Salzberg 2020). In either case, these bacterial-derived sequences overwhelm the ability of either *k*-mer matching or read mapping-based approaches to detect eukaryotes from microbiome sequencing, as bacteria represent the majority of the sequencing library from most microbiomes. We demonstrated this problem by simulating shotgun sequencing reads from the genomes of 971 common human gut microbiome bacteria representing all major bacterial phyla in the human gut, including Bacteroidetes, Actinobacteria, Firmicutes, Proteobacteria, and Fusobacteria (Zou et al. 2019). We aligned these reads against a database of 2,449 genomes of fungi, protists, roundworms, and flatworms from NCBI GenBank. We simulated 2 million 126-bp paired-end reads per bacterial genome. Even with stringent read filtering (mapping quality > 30, aligned portion > 100 base pairs), a large number of simulated reads aligned to this database, from almost every individual bacterial genome (Figure 1A). In total, 112 bacteria had more than 1% of simulated reads aligning to the eukaryotic genome database, indicating that the potential for spurious alignment of reads to eukaryotes from human microbiome bacteria is not limited to a few bacterial taxa. A further complication is that bacterial reads aligned to many different eukaryotic genomes in species from different taxonomic groups (Figure 1B), indicating that bacterial contamination of eukaryotic genomes is widespread, and that simply removing a small number of contaminated genomes from an analysis will not be sufficient. In total, 18% of animal genomes, 10% of protist genomes, 11% of unassigned genomes, and 2% of fungal genomes had more than 1,000 reads from a single bacterial genome align with

high confidence. As the majority of the sequencing library of a given human microbiome sample is expected to be mostly comprised of bacteria (Qin et al. 2010), this spurious signal drowns out any true eukaryotic signal.

In practice, progress has been made in identifying eukaryotes by using alignment based approaches coupled with extensive masking, filtering, and manual curation (Nash et al. 2017; Chehoud et al. 2015). However, these approaches still typically require extensive manual examination because of the incomplete nature of the databases used to mask. These manual processes are not scalable for analyzing the large amounts of sequencing data that are generated by microbiome sequencing studies and the rapidly growing number of eukaryotic genomes.

Using universal marker genes to identify eukaryotes

To address the problem of identifying eukaryotic species in microbiomes from shotgun sequencing data, we created a database of marker genes that are uniquely eukaryotic in origin and uncontaminated with bacterial sequences. We chose genes that are universally conserved to achieve the greatest identification power. We focused on universal eukaryotic genes rather than clade-, species-, or strain-specific genes because many microbial eukaryotes have variable size pangenomes and universal genes are less likely to be lost in a given lineage (Gentekaki et al. 2017; McCarthy and Fitzpatrick 2019). The chosen genes contain sufficient sequence variation to be informative about which eukaryotic species are present in sequencing data.

More than half of eukaryotic microbial species with sequenced genomes in NCBI Genbank do not have annotated genes or proteins. Therefore, we used the

Benchmarking Universal Single Copy Orthologs (BUSCO) pipeline, which integrates sequence similarity searches and *de novo* gene prediction to determine the location of putative conserved eukaryotic genes in each genome (Simão et al. 2015). Other databases of conserved eukaryotic genes exist, including PhyEco and the recently available EukCC and EukProt (Saary et al. 2020; Richter et al. 2020; Wu et al. 2013). We found the BUSCO pipeline was advantageous, as BUSCO and the OrthoDB database have been benchmarked in multiple applications including gene predictor training, metagenomics, and phylogenetics (Waterhouse et al. 2018; Kriventseva et al. 2019).

We ran the BUSCO pipeline using the OrthoDB Eukaryote lineage (Kriventseva et al. 2019) on all 2,571 microbial eukaryotic genomes (2,050 fungi, 362 protists and non-assigned organisms, and 159 roundworms and flatworms) downloaded from NCBI GenBank in January 2020. Marker gene sequences were then extensively quality filtered (see Methods). Genes with greater than 99% sequence identity were combined and represent the most recent common ancestor of the species with collapsed genes.

A final set of 389,724 marker genes from 214 BUSCO orthologous gene sets was selected, of which 367,811 genes correspond to individual species and the remainder to internal nodes (i.e., genera, families) in the taxonomy tree. The full EukDetect database of universal marker genes in eukaryotes has species-level markers for 2,012 fungi, 346 protists and unassigned taxa, and 158 roundworms and flatworms (Figure 2, Table S3). This represents 98% of all eukaryotic microbial taxonomic groups with sequenced genomes in GenBank, of which more than half do not have annotated genes. Confirming that reads from bacteria would not spuriously align to this eukaryotic

marker database, zero reads from our simulated bacterial shotgun data (Figure 1) align to this database.

Despite the BUSCO universal single-copy nomenclature, the EukDetect database genes are present at variable levels in different species (Figure 2). We uncovered a number of patterns that explain why certain taxonomic groups have differing numbers of marker genes. First, many of the species present in our dataset, including microsporidia and many of the pathogenic protists, function as parasites. As parasites rely on their host for many essential functions, they frequently experience gene loss and genome reduction (McCutcheon and Moran 2012), decreasing the representation of BUSCO genes. It is also possible that the BUSCO pipeline does not identify all marker genes in some clades with few representative genomes, and therefore few inputs into the hidden Markov model used for locating putative gene sequences. Conversely, some clades have high rates of duplication for these typically single-copy genes. These primarily occur in fungi; studies of fungal genomes have demonstrated that hybridization and whole genome duplication occurs frequently across the fungal tree of life (Stukenbrock 2016; Marcet-Houben and Gabaldón 2015).

As the representation of each marker gene differs across taxonomic groups, we identified a subset of 50 marker genes that are more frequently found in all taxonomic groups (Table S2). This core set of genes has the potential to be used for strain identification, abundance estimation, and phylogenetics, as has been demonstrated for universal genes in bacteria (Wu et al. 2013).

EukDetect: a pipeline to accurately identify eukaryotes using universal marker genes

We incorporated the complete database into EukDetect, a bioinformatics pipeline (https://github.com/allind/EukDetect) that first aligns metagenomic sequencing reads to the marker gene database, and filters each read based on alignment quality, sequence complexity, and alignment length. Sequencing reads can be derived from DNA or from RNA sequencing. EukDetect then removes duplicate sequences, calculates the coverage and percent identity of detected marker genes, and reports these results in a taxonomy-informed way (see Methods for more detail).

EukDetect is sensitive and accurate even at low sequence coverage

To determine the performance of EukDetect, we simulated reads from representative species found in human microbiomes, including fungi from the *Candida* genus and the *Malassezia* genus, protists from the *Blastocystis* genus and *Trichomonas vaginalis*, and the flatworm *Schistosoma mansoni*. These species represent different groups of the eukaryotic tree of life, and are variable in their genome size and representation in the EukDetect marker gene database. The number of reads aligning to the database and the number of observed marker genes vary based on the total number of marker genes per organism and the overall size of its genome (Figure 3). We simulated metagenomes with sequencing coverage across a range of values for each species. We found that EukDetect aligns at least one read to 80% or more of all marker genes per species at low simulated genome coverages between 0.25x or 0.5x (Figure 3A).

Using a detection limit cutoff requiring at least 4 unique reads aligning to 2 or more marker genes, *Candida albicans* and *Malassezia restricta* are detectable at

coverages as low as 0.0025x, which constitutes only 141 and 75 reads simulated from each genome, respectively (Figure 3B). *Schistosoma mansoni* and *Trichomonas vaginalis* are detectable at 0.005x coverage (7,233 and 3,500 simulated reads). Due to the low number of marker genes in the database for *Blastocystis hominis*, this species is detectable at a genome coverage of 0.05x or greater (373 simulated reads).

In the case of species where multiple close relatives have sequenced genomes, off-target alignment can occur where reads from one species erroneously align to another. This occurs due to errors in sequencing masking or due to high local similarities between gene sequences. We simulated reads from *Candida albicans, Candida dubliniensis, Malassezia dermatis, Malassezia globosa,* and *Schistosoma mansoni*, as these species have close relatives with a sequenced genome in the EukDetect marker database, and determined the impact of off-target read alignments. We observed that the absolute number of off-target reads is lower than the number of on-target reads across different simulated genome coverages (Figure 3C). One distinguishing feature of these off-target reads is that the overall percent identity is lower than the percent identity for correctly determined reads (Figure 3D). Therefore, this difference in percent identity and in read counts can be used to disambiguate cases where off-target alignment occurs.

In cases where high numbers of eukaryotic reads are expected, the majority of sequences will to align to markers from species most closely related to the eukaryote present in the sample. However, due to factors such as sequencing errors, reads will align to other closely related species as well, causing a similar phenomenon as seen with amplicon sequencing where abundant taxa inflate rare taxa counts (Rosen et al.

2012). A combination of factors including the number of reads aligning to marker genes, the amount of observed marker genes per species, and the percent identity of aligned reads should all be considered in cases where two closely related species are identified by EukDetect. The most conservative option is to consider the most recent common ancestor of the closely related species, which usually resolves at the genus level. This information is provided by EukDetect (Figure S1).

Vignettes of microbial eukaryotes in microbiomes

To demonstrate how EukDetect can be used to understand microbiome eukaryotes, we investigated several biological questions about eukaryotes in microbiomes using publicly available datasets from human and plant associated microbiomes.

The spatial distribution of eukaryotic species in the gut

While microbes are found throughout the human digestive tract, studies of the gut microbiome often examine microbes in stool samples, which cannot provide information about the spatial distribution of microbes in the gut (Tropini et al. 2017). Analyses of human and mouse gut microbiota have shown differences in the bacteria that colonize the lumen and the mucosa of the GI tract, as well as major differences between the upper GI tract, the small intestine, and the large intestine, related to the ecology of each of these environments. Understanding the spatial organization of microbes in the gut is critical to dissecting how these microbes interact with the host and contribute to host phenotypes.

To determine the spatial distribution of eukaryotic species in the human gut, we analyzed data from two studies that examined probiotic strain engraftment in the human gut. These studies used healthy adult humans recruited in Israel, and collected stool samples along with biopsies performed by endoscopy from 18 different body sites along the upper GI tract, small intestine, and large intestine (Suez et al. 2018; Zmora et al. 2018). A total of 1,613 samples from 88 individuals were sequenced with whole-genome shotgun sequencing.

In total, eukaryotes were detected in 324 of 1,613 samples with the EukDetect pipeline (Table S4). The most commonly observed eukaryotic species were subtypes of the protist *Blastocystis* (235 samples), the yeast *Saccharomyces cerevisiae* (66 samples), the protist *Entamoeba dispar* (17 samples), the yeast *Candida albicans* (9 samples), and the yeast *Cyberlindnera jadinii* (6 samples). Additional fungal species were observed in fewer samples, and included the yeast *Malassezia restricta*, a number of yeasts in the *Saccharomycete* class including *Candida tropicalis* and *Debaromyces hansenii*, and the saprophytic fungus *Penicillium roqueforti*.

The prevalence and types of eukaryotes detected varied along the GI tract. In the large intestine and the terminal ileum, microbial eukaryotes were detected at all sites, both mucosal and lumen-derived. However, they were mostly not detected in the small intestine or upper GI tract (Figure 4). One exception to this is a sample from the gastric antrum mucosa which contained sequences assigned to the fungus *Malassezia restricta*. *Blastocystis* species were present in all large intestine and terminal ileum samples, while fungi were present in all large intestine samples and the lumen of the terminal ileum. The protist *Entamoeba dispar* was detected almost exclusively in stool

samples; only one biopsy-derived sample from the descending colon lumen contained sequences assigned to an *Entamoeba* species. The saprophytic fungus *Penicillium roqueforti*, which is used in food production and is likely to be allochthonous in the gut due to its likely inability to grow in that environmnt (Suhr and Hallen-Adams 2015), was only detected in stool samples.

Our detection of microbial eukaryotes in mucosal sites is intriguing. Bacteria colonizing the mucosa have greater opportunities to interact with host factors. We detected both protists and fungi (*Blastocystis, Malassezia,* and *Saccharomyces cerevisiae)* in mucosal as well as lumen samples, suggesting that they may be closely associated with host cells and not just transiently passing through the GI tract. These findings are consistent with previous studies of *Blastocystis*-infected mice that found *Blastocystis* in both the lumen and the mucosa of the large intestine (Moe et al. 1997). Taken together, these findings suggest that certain eukaryotic species can directly interact with the host.

A succession of eukaryotic microbes in the infant gut

The gut bacterial microbiome undergoes dramatic changes during the first years of life (Bäckhed et al. 2015). We sought to determine whether eukaryotic members of the gut microbiome also change over the first years of life. To do so, we examined longitudinal shotgun sequencing data from the three-country DIABIMMUNE cohort, where stool samples were taken from infants in Russia, Estonia, and Latvia during the first 1200 days of life (Vatanen et al. 2016). In total, we analyzed 791 samples from 213 individuals.

Microbial eukaryotes are fairly commonly found in stool in the first few years of life. We detected reads assigned to a eukaryotic species from 107 samples taken from 68 individuals (Table S5). The most frequently observed species were *Candida parapsilosis* (30 samples), *Saccharomyces cerevisiae* (27 samples), various subtypes of the protist *Blastocystis* (17 samples), *Malassezia* species (13 samples), and *Candida albicans* (9 species). These results are consistent with previous reports that *Candida* species are prevalent in the neonatal gut (Olm et al. 2019; Fujimura et al. 2016). Less frequently observed species were primarily yeasts in the *Saccharomycete* class, and the coccidian parasite *Cryptosporidium meleagridis* was detected in one sample.

These species do not occur uniformly across time. When we examined covariates associated with different eukaryotic species, we found a strong association with age. The median age at collection of the samples analyzed with no eukaryotic species was 450 days, but *Blastocystis* protists and *Saccharomycetaceae* fungi were primarily observed among older infants (median 690 days and 565 days, respectively) (Figure 5A). Fungi in the *Debaromycetaceae* family were observed among younger infants (median observation 312 days). Samples containing fungi in the *Malasseziaceae* family were older than *Debaromycetaceae* samples (median observation 368 days), though younger than the non-eukaryotic samples. To determine whether these trends were statistically significant, we compared the mean ages of two filtered groups: individuals with no observed eukaryotes and individuals where only one eukaryotic family was observed (Figure 5B). We found that *Saccharomycetaceae* fungi and *Blastocystis* protists were detected in significantly older children (Wilcoxon rank-sum p=0.0012 and p=2.6e-06, respectively). In contrast, *Debaromycetaceae* fungi were

found in significantly younger infants (p=0.035). As only three samples containing *Malasseziaceae* came from individuals where no other eukaryotic families were detected, we did not analyze this family.

These findings suggest that, as observed with bacteria, the eukaryotic species that colonize the gastrointestinal tract of children change during early life (Bäckhed et al. 2015). Our results support a model of eukaryotic succession in the infant gut, where *Debaromycetaceae* fungi, notably *Candida parapsilosis* in these data, dominate the eukaryotic fraction of the infant gut during the first year of life, and *Blastocystis* and *Saccharomyces* fungi, which are commonly observed in the gut microbiomes of adults, rise to higher prevalence in the gut in the second year of life and later (Figure 5C). Altogether these results expand the picture of eukaryotic diversity in the human early life GI tract.

Differences in RNA and DNA detection of eukaryotic species suggests differential transcriptional activity in the gut

While most sequencing-based analyses of microbiomes use DNA, microbiome transcriptomics can reveal the gene expression of microbes in a microbiome, which has the potential to shed light on function. RNA sequencing of microbiomes can also be used to distinguish dormant or dead cells from active and growing populations. We sought to determine whether eukaryotic species are detectable from microbiome transcriptomics, and how these results compare to DNA shotgun sequencing.

We leveraged data from the Inflammatory Bowel Disease (IBD) Multi'omics Data generated by the Integrative Human Microbiome Project (IHMP), which generated RNA-

Seq, whole-genome sequencing, viral metagenomic sequencing, and 16S amplicon sequencing from stool samples of individuals with Crohn's disease, ulcerative colitis, and no disease (IBDMDB; http://ibdmdb.org) (Lloyd-Price et al. 2019). We analyzed samples with paired whole genome DNA shotgun sequencing and RNA sequencing. In the IHMP-IBD dataset, there were 742 sets of paired DNA and RNA sequencing from a sample from 104 individuals, 50 of whom were diagnosed with Crohn's disease, 28 of whom were diagnosed with ulcerative colitis, and 19 with no with IBD. Samples were collected longitudinally, and the median number of paired samples per individual was seven.

Microbial eukaryotes were prevalent in this dataset. EukDetect identified eukaryotic species in 407 / 742 RNA-sequenced samples and 398 / 742 DNA-sequenced samples. The most commonly detected eukaryotic species were *Saccharomyces cerevisiae*, *Malassezia* species, *Blastocystis* species, and the yeast *Cyberlindnera jadinii* (Table S6). Eukaryotic species that were detected more rarely were primarily yeasts from the *Saccharomycete* class. We examined whether eukaryotic species were present predominantly in DNA sequencing, RNA sequencing, or both, and found different patterns across different families of species (Figure 6). Of the 585 samples where we detected fungi in the *Saccharomycetaceae* family, 314 of those samples were detected from the RNA-sequencing alone and not from the DNA. A further 178 samples had detectable *Saccharomycetaceae* fungi in both the DNA and the RNA sequencing, while 17 samples only had detectable *Saccharomycetaceae* fungi in the DNA sequencing. Fungi in the *Malasseziaceae* family were only detected in DNA sequencing (115 samples total), and the bulk of *Cyberlindnera jadinii* and

Debaryomycetaceae fungi (*Candida albicans* and *Candida tropicalis*) were detected in DNA sequencing alone. In contrast, *Blastocystis* protists were mostly detected both from RNA and from DNA, and *Pichiaceae* fungi (*Pichia* and *Brettanomyces* yeasts) were detected in both RNA and DNA sequencing, DNA sequencing alone, and in a small number of samples in RNA-seq alone.

From these findings, we can infer information about the abundance and possible functions of these microbial eukaryotes. *Blastocystis* is the most frequently observed gut protist in the human GI tract in industrialized nations (Scanlan et al. 2014), and its high relative abundance is reflected in the fact that it is detected most frequently from both RNA and DNA sequencing. In contrast, the much greater detection of *Saccharomycetaceae* fungi from RNA than from DNA suggests that these cells are transcriptionally active, and that while the absolute cell counts of these fungi may not be detectable from DNA sequencing, they are actively transcribing genes and therefore may impact the ecology of the gut microbiome. In contrast, fungi in the *Malasseziaceae* family are found at high relative abundances on human skin and while they have been suggested to play functional roles in the gut (Limon et al. 2019; Sokol et al. 2017; Aykut et al. 2019), the data from this cohort suggest that the *Malasseziaceae* fungi are rarely transcriptionally active by the time they are passed in stool. Fungi in the *Phaffomycetaceae, Pichiaceae*, and *Debaromycetaeae* families likely represent a middle ground, where some cells are transcriptionally active and detected from RNA-sequencing, but others are not active in stool. These results suggest that yeasts in the *Saccharomycete* clade survive passage through the GI tract and may contribute functionally to the gut microbiome.

Eukaryotes in the plant leaf microbiome

Because the EukDetect database uses all fungal, protist, flatworm, and roundworm genomes available for download from NCBI, this classification tool is not limited to human microbiome analyses. We analyzed 1,394 samples taken from 275 wild *Arabidopsis thaliana* leaves (Regalado et al. 2020). From these samples, we detect 37 different eukaryotic species and 25 different eukaryotic families (Table S7). We found pathogenic *Peronospora* oomycetes in 374 samples and gall forming *Protomyces* fungi in 312 samples. Other frequently observed eukaryotes in these samples are *Dothideomycete* fungi in 101 samples, *Sordariomycete* fungi in 29 samples, *Agaricomycete* fungi in 26 samples, and *Tremellomycete* fungi in 10 samples. *Malassezia* fungi are also detected in 14 samples, though we hypothesize that these are contaminants from human skin. Many of the most commonly detected microbial eukaryotic species in these samples, including the *Arabidopsis*-isolated yeast *Protomyces* sp. C29 in 311 samples and the epiphytic yeast *Dioszegia crocea* in 10 samples, do not have annotated genes associated with their genomes and have few to no gene or protein sequences in public databases. These taxa and others would not be identifiable by aligning reads to existing gene or protein databases. These results demonstrate that EukDetect can be used to detect microbial eukaryotes in non-human associated microbiomes and reveal more information than aligning reads to gene or protein databases.

Discussion

Here we present EukDetect, an analysis pipeline that leverages a database of conserved eukaryotic genes to identify eukaryotes present in metagenomic sequencing. By using conserved eukaryotic genes, this pipeline avoids inaccurately classifying sequences based on bacterial contamination of eukaryotic genomes, which is widespread (Figure 1) (Lu and Salzberg 2018; Steinegger and Salzberg 2020). The EukDetect pipeline is sensitive and can detect fungi, protists, and worms present at low sequencing coverage in metagenomic sequencing data.

We apply the EukDetect pipeline to public datasets from the human gut microbiome and the plant leaf microbiome, detecting eukaryotes uniquely present in each environment. We find that fungi and protists are present at all sites within the lower GI tract of adults, and that the eukaryotic composition of the gut microbiome changes during the first years of life. Using paired DNA and RNA sequencing from the iHMP IBDMBD project, we demonstrate the utility of EukDetect on RNA data and show that some eukaryotes are differentially detectable in DNA and RNA sequencing, suggesting that some eukaryotic cells are dormant or dead in the GI tract, while others are actively transcribing genes. Finally, we find oomycetes and fungi in the *Arabidopsis thaliana* leaf microbiome, many of which would not have been detectable using existing protein databases.

One important limitation of our approach is that only eukaryotic species with sequenced genomes or that have close relatives with a sequenced genome can be detected by the EukDetect pipeline. Due to taxonomic gaps in sequenced genomes, the EukDetect database does not cover the full diversity of the eukaryotic tree of life. We focused our applications on environments that have been studied relatively well. But

nonetheless some eukaryotes that are known to live in animal GI tracts, such as *Dientamoeba fragilis* (Osman et al. 2016), have not been sequenced and would have been missed if they were present in the samples we analyzed. Improvements in single-cell sequencing (Ahrendt et al. 2018) and in metagenomic assembly of eukaryotes (West et al. 2018) will increase the representation of uncultured eukaryotic microbes in genome databases. Because EukDetect uses universal genes, it will be straightforward to expand its database as more genomes are sequenced.

Taken together, the work reported here demonstrates that eukaryotes can be effectively detected from shotgun sequencing data with a database of conserved eukaryotic genes. As more metagenomic sequencing data becomes available from host associated and environmental microbiomes, tools like EukDetect will reveal the contributions of microbial eukaryotes to diverse environments.

Methods

Identifying universal eukaryotic marker genes in microbial eukaryotes

Microbial eukaryotic genomes were downloaded from NCBI GenBank on January 14, 2020 for all species designated as “Fungi”, “Protists”, “Roundworms”, “Flatworms”, and “Other”. One genome was downloaded for each species; priority was given to genomes designated as ‘reference genomes’ or ‘representative genomes’. If a species with multiple genomes did not have a designated representative or reference genome, the genome assembly that appeared most contiguous was selected.

To identify marker genes in eukaryotic genomes, we ran the Benchmarking Universal Single-Copy Orthologs (BUSCO) version 4 pipeline on all eukaryotic genomes

with the Eukaryota OrthoDB version 10 gene models (255 universal eukaryote marker genes) using the augustus optimization parameter (command --long) (Simão et al. 2015; Kriventseva et al. 2019). To ensure that no bacterial sequences were erroneously annotated with this pipeline, we also ran BUSCO with the same parameters on a set of 971 common human gut bacteria from the Culturable Genome Reference (Zou et al. 2019) and found that this pipeline annotated 41 of the 255 universal marker genes in one or more bacteria. We discarded these predicted markers from each set of BUSCO predicted genes in eukaryotes. The 214 marker genes never predicted in a bacterial genome are listed in Table S1.

Constructing the EukDetect database

After running the BUSCO pipeline on each eukaryotic genome and discarding gene models that were annotated in bacterial genomes, we extracted full length nucleotide and protein sequences from each complete BUSCO gene. Protein sequences were used for filtering only. We examined the length distribution of the predicted proteins of each potential marker gene, and genes whose proteins were in the top or bottom 5% of length distributions were discarded. We masked simple repetitive elements with RepeatMasker version open-4.0.7 and discarded genes where 10% or more of the sequence was masked (Smit et al. 2013).

After filtering, we used CD-HIT version 4.7 to collapse sequences that were greater than 99% identical (Fu et al. 2012). For a small number of clusters of genomes, all or most of the annotated BUSCO genes were greater than 99% identical. This arose from errors in taxonomy (in some cases, from fungi with a genome sequence deposited

both under the anamorph and teleopmorph name) and from genomes submitted to GenBank with a genus designation only. In these cases, one genome was retained from the collapsed cluster. Some collapsed genes were gene duplicates from the same genome which are designated with the flag “SSCollapse” in the EukDetect database, where “SS” designates “same species”. Genes that were greater than 99% identical between species in the same genus were collapsed, re-annotated as representing NCBI Taxonomy ID associated with the last common ancestor of all species in the collapsed cluster, and annotated with the flag “SGCollapse”. Genes that were collapsed where the species in the collapsed cluster came from multiple NCBI Taxonomy genera were collapsed, annotated as representing the NCBI Taxonomy ID associated with the last common ancestor of all species in the collapsed cluster, and annotated with the flag “MGCollapse”. The EukDetect database is available from Figshare (https://doi.org/10.6084/m9.figshare.12670856.v4). A smaller database of 50 conserved BUSCO genes that could potentially be used for phylogenetics is available from Figshare (https://doi.org/10.6084/m9.figshare.12693164.v1). These 50 marker genes are listed in Table S2.

The EukDetect pipeline

The EukDetect pipeline uses the Snakemake workflow engine (Köster and Rahmann 2012). Sequencing reads are aligned to the EukDetect marker database with Bowtie2 (Langmead and Salzberg 2012, 2). Reads are filtered for mapping quality greater than 30 and alignments that are less than 80% of the read length of the gene are discarded. Aligned reads are then filtered for sequence complexity with komplexity

(https://github.com/eclarke/komplexity) and are discarded below a complexity threshold of 0.5. Duplicate reads are removed with samtools (Li et al. 2009). Reads aligning to marker genes are counted and the percent sequence identity of reads is calculated. The marker genes in the EukDetect database are each linked to NCBI taxonomy IDs. Further filtering is performed where only taxonomy IDs that have 4 or more aligned reads mapping to 2 or more marker genes are reported. Results are reported both as a count, coverage, and percent sequence identity table for each taxonomy IDs that passes filtering, and as a taxonomy of all reported reads constructed with ete3 (Huerta-Cepas et al. 2016). A schematic of the EukDetect pipeline is depicted in Figure S1.

Simulated reads

Paired-end Illumina reads were simulated from *Candida albicans, Candida dubliniensis, Malassezia restricta, Malassezia dermatis, Malassezia globose, Trichomonas vaginalis, Blastocystis hominis*, and *Schistosoma mansoni* with InSilicoSeq (Gourlé et al. 2019). Each species was simulated at 13 coverage depths: 0.0001x, 0.001x, 0.01x, 0.05x, 0.1x, 0.25x, 0.5x, 0.75x, 1x, 2x, 3x, 4x, and 5x. Simulated reads were processed with the EukDetect pipeline.

Analysis of public datasets

All sequencing data and associated metadata were taken from public databases and published studies. Sequencing data for determining eukaryotic GI tract distribution was downloaded from the European Nucleotide Archive under accession PRJEB28097 (Zmora et al. 2018; Suez et al. 2018). Sequencing data and metadata for the

DIABIMMUNE three country cohort was downloaded directly from the DIABIMMUNE project website (https://diabimmune.broadinstitute.org/diabimmune/three-country-cohort) (Vatanen et al. 2016). DNA and RNA sequencing from the IHMP IBDMDB project was downloaded from the NCBI SRA under Bioproject PRJNA398089 (Lloyd-Price et al. 2019). Sequencing data from the *Arabidopsis thaliana* leaf microbiome was downloaded from the European Nucleotide Archive under accession PRJEB31530 (Regalado et al. 2020).

Data availability

The EukDetect pipeline is available on github (https://github.com/allind/EukDetect). The EukDetect database can be downloaded from Figshare (https://doi.org/10.6084/m9.figshare.12670856.v4)

Figure legends.

Figure 1. Human gut microbiome bacterial shotgun sequence reads are misattributed to eukaryotes. (A) Metagenomic sequencing reads were simulated from 971 species total from all major phyla in human stool (2 million reads per species) and aligned to all microbial eukaryotic genomes used to develop EukDetect. Even after stringent filtering (Methods), many species have thousands of reads aligning to eukaryotic genomes, which would lead to false detection of eukaryotes in samples with only bacteria. An additional 111 species with > 20,000 reads aligned to eukaryotic genomes are not

shown. Red dashed line indicates species with greater than 1,000 reads aligning to eukaryotic genomes. (B) Taxonomic distribution of eukaryotic genomes with >1,000 aligned bacterial reads. All major taxonomic groups of microbial eukaryotes are affected.

Figure 2. EukDetect database marker genes are represented in 98% of currently sequenced eukaryotes. (A) Total number of marker genes per species across taxonomic groups, including genes present in multiple copies. (B) Total number of marker genes identified per species by taxonomic group, excluding duplicates of genes.

Figure 3. EukDetect pipeline performance is sensitive for yeasts, protists, and worms at low sequence coverage. (A) Number of marker genes with at least one aligned read per species up to 1x genome coverage. Horizontal red line indicates the total number of marker genes per species (best possible performance). (B) Number of marker genes with at least one aligned read per species up to 0.01x genome coverage. Vertical red line indicates a detection cutoff where 4 or more reads align to 2 or more markers. (C) Counts of reads aligned to on- and off-target marker genes (i.e, correctly classified vs. incorrectly classified) across simulated genome coverages for species with close relatives in the EukDetect database. On-target alignments dominate at all coverages. (D) Percent sequence identity of all on-target and all off-target reads across simulated genome coverages. On-target reads have consistently higher percent identity.

Figure 4. Distribution of eukaryotic species in the gastrointestinal tract taken from biopsies. Eukaryotes were detected at all sites in the large intestine and in the terminal ileum, in both lumen and mucosal samples. One biopsy of gastric antrum mucosa in the stomach contained a *Malassezia* yeast. Slashes indicate no eukaryotes detected in any samples from that site. See Figures S2 and S3 for locations of *Blastocystis* subtypes and locations of fungi.

Figure 5. Changes in eukaryotic gut microbes during the first years of life. (A) Age at collection in the DIABIMMUNE three-country cohort for samples with no eukaryote or with any of the four most frequently observed eukaryotic families. (B) The mean age at collection of samples from individuals with no observed eukaryotes compared to the mean age at collection of individuals where one of three eukaryotic families. Individuals where more than one eukaryotic family was detected are excluded. *Malasseziaceae* is excluded due to low sample size. Group comparisons were performed with an unpaired Wilcoxon rank-sum test. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. (C) Model of eukaryotic succession in the first years of life. *Debaryomycetaceae* species predominate during the first two years of life but are detected later. *Blastocystidae* species and *Saccharomycetaceae* species predominate after the first two years of life, but are detected during the second year. *Malasseziaceae* species are constant.

Figure 6. Detection of eukaryotes from paired DNA and RNA sequenced samples from the IHMP IBD cohort. Plots depict the most commonly detected eukaryotic families, and whether a given family was detected in the DNA sequencing alone, the RNA

sequencing alone, or from both the RNA and the DNA sequencing from a sample. Some samples shown here come from the same individual sampled at different time points.

Supplemental files.

Figure S1. Schematic of the EukDetect pipeline.

Figure S2. Distribution of Blastocystis in the gastrointestinal tract taken from biopsies. Fungi were detected at all sites in the large intestine and terminal ileum, in both lumen and mucosal samples. Slashes indicate no Blastocystis detected in any samples from that site.

Figure S3. Distribution of fungal species in the gastrointestinal tract taken from biopsies. Fungi were detected at all sites in the large intestine in both lumen and mucosal samples, and in the lumen of the terminal ileum. One biopsy of gastric antrum mucosa in the stomach contained a Malassezia yeast. Slashes indicate no fungi detected in any samples from that site.

Table S1. OrthoDB v10 Eukaryota marker genes that were not predicted in any of 971 bacterial genomes. (CSV)

Table S2. Genomes with species-level marker genes in the EukDetect database. (CSV)

Table S3. Subset of 50 conserved marker genes found broadly across microbial eukaryotes. (CSV)

Table S4. Eukaryotes in gut biopsy and stool samples. (CSV)

Table S5. Euakaryotes in DIABIMMUNE three-country cohort samples. (CSV)

Table S6. Eukaryotes in IHMP IBD samples. (CSV)

Table S7. Eukaryotes from *Arabidopsis thaliana* leaf samples. (CSV)

References

Ahrendt SR, Quandt CA, Ciobanu D, Clum A, Salamov A, Andreopoulos B, Cheng J-F, Woyke T, Pelin A, Henrissat B, et al. 2018. Leveraging single-cell genomics to expand the fungal tree of life. *Nat Microbiol* **3**: 1417–1428.

Akin DE, Borneman WS. 1990. Role of Rumen Fungi in Fiber Degradation. *J Dairy Sci* **73**: 3023–3032.

Aykut B, Pushalkar S, Chen R, Li Q, Abengozar R, Kim JI, Shadaloey SA, Wu D, Preiss P, Verma N, et al. 2019. The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL. *Nature* **574**: 264–267.

Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H, et al. 2015. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**: 690–703.

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK. 2012. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol Evol* **27**: 233–243.

Brussaard L, de Ruiter PC, Brown GG. 2007. Soil biodiversity for agricultural sustainability. *Agric Ecosyst Environ* **121**: 233–244.

Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, Bell CJ, Bharti A, Dyhrman ST, Guida SM, et al. 2017. Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nat Rev Microbiol* **15**: 6–20.

Chehoud C, Albenberg LG, Judge C, Hoffmann C, Grunberg S, Bittinger K, Baldassano RN, Lewis JD, Bushman FD, Wu GD. 2015. A Fungal Signature in the Gut Microbiota of Pediatric Patients with Inflammatory Bowel Disease. *Inflamm Bowel Dis* **21**: 1948–1956.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma Oxf Engl* **28**: 3150–3152.

Fujimura KE, Sitarik AR, Havstad S, Lin DL, Levan S, Fadrosh D, Panzer AR, LaMere B, Rackaityte E, Lukacs NW, et al. 2016. Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat Med* **22**: 1187–1191.

Gentekaki E, Curtis BA, Stairs CW, Klimeš V, Eliáš M, Salas-Leiva DE, Herman EK, Eme L, Arias MC, Henrissat B, et al. 2017. Extreme genome diversity in the hyper-prevalent parasitic eukaryote Blastocystis. *PLOS Biol* **15**: e2003769.

Gourlé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. 2019. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinforma Oxf Engl* **35**: 521–522.

Haque R. 2007. Human Intestinal Parasites. *J Health Popul Nutr* **25**: 387–391.

Hernández-Santos N, Klein BS. 2017. Through the Scope Darkly: The Gut Mycobiome Comes into Focus. *Cell Host Microbe* **22**: 728–729.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33**: 1635–1638.

Kamoun S, Furzer O, Jones JDG, Judelson HS, Ali GS, Dalio RJD, Roy SG, Schena L, Zambounis A, Panabières F, et al. 2015. The Top 10 oomycete pathogens in molecular plant pathology. *Mol Plant Pathol* **16**: 413–434.

Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**: 2520–2522.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* **47**: D807–D811.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–9.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinforma Oxf Engl* **25**: 2078–9.

Limon JJ, Tang J, Li D, Wolf AJ, Michelsen KS, Funari V, Gargus M, Nguyen C, Sharma P, Maymi VI, et al. 2019. Malassezia Is Associated with Crohn's Disease and Exacerbates Colitis in Mouse Models. *Cell Host Microbe* **25**: 377-388.e6.

Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**: 655–662.

Lu J, Salzberg SL. 2018. Removing contaminants from databases of draft genomes. *PLOS Comput Biol* **14**: e1006277.

Lücking R, Aime MC, Robbertse B, Miller AN, Ariyawansa HA, Aoki T, Cardinali G, Crous PW, Druzhinina IS, Geiser DM, et al. 2020. Unambiguous identification of fungi: where do we stand and how accurate and precise is fungal DNA barcoding? *IMA Fungus* **11**: 14.

Marcet-Houben M, Gabaldón T. 2015. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLOS Biol* **13**: e1002220.

McCarthy CGP, Fitzpatrick DA. 2019. Pan-genome analyses of model fungal species. *Microb Genomics* **5**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6421352/ (Accessed July 6, 2020).

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10**: 13–26.

Moe KT, Singh M, Howe J, Ho LC, Tan SW, Chen XQ, Ng GC, Yap EH. 1997. Experimental Blastocystis hominis infection in laboratory mice. *Parasitol Res* **83**: 319–325.

Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, Stewart CJ, Metcalf GA, Muzny DM, Gibbs RA, et al. 2017. The gut mycobiome of the Human Microbiome Project healthy cohort. *Microbiome* **5**: 153.

Olm MR, West PT, Brooks B, Firek BA, Baker R, Morowitz MJ, Banfield JF. 2019. Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* **7**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6377789/ (Accessed July 6, 2020).

Osman M, El Safadi D, Cian A, Benamrouz S, Nourrisson C, Poirier P, Pereira B, Razakandrainibe R, Pinon A, Lambert C, et al. 2016. Prevalence and Risk Factors for Intestinal Protozoan Infections with Cryptosporidium, Giardia, Blastocystis and Dientamoeba among Schoolchildren in Tripoli, Lebanon. *PLoS Negl Trop Dis* **10**: e0004496.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59–65.

Regalado J, Lundberg DS, Deusch O, Kersten S, Karasov T, Poersch K, Shirsekar G, Weigel D. 2020. Combining whole-genome shotgun sequencing and rRNA gene amplicon analyses to improve detection of microbe–microbe interaction networks in plant leaves. *ISME J* 1–15.

Richter DJ, Berney C, Strassert JFH, Burki F, Vargas C de. 2020. EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotic life. *bioRxiv* 2020.06.30.180687.

Rodriguez RJ, Jr JFW, Arnold AE, Redman RS. 2009. Fungal endophytes: diversity and functional roles. *New Phytol* **182**: 314–330.

Rosen MJ, Callahan BJ, Fisher DS, Holmes SP. 2012. Denoising PCR-amplified metagenome data. *BMC Bioinformatics* **13**: 283.

Saary P, Mitchell AL, Finn RD. 2020. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis. *bioRxiv* 2019.12.19.882753.

Scanlan PD, Stensvold CR, Rajilić-Stojanović M, Heilig HGHJ, De Vos WM, O'Toole PW, Cotter PD. 2014. The microbial eukaryote Blastocystis is a prevalent and diverse member of the healthy human gut microbiota. *FEMS Microbiol Ecol* **90**: 326–330.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Consortium FB. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci* **109**: 6241–6246.

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

Smit A, Hubley R, Green P. 2013. Repeatmasker Open-4.0. http://www.repeatmasker.org (Accessed January 10, 2015).

Sokol H, Leducq V, Aschard H, Pham H-P, Jegou S, Landman C, Cohen D, Liguori G, Bourrier A, Nion-Larmurier I, et al. 2017. Fungal microbiota dysbiosis in IBD. *Gut* **66**: 1039–1048.

Steinegger M, Salzberg SL. 2020. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol* **21**: 115.

Stukenbrock EH. 2016. The Role of Hybridization in the Evolution and Emergence of New Fungal Plant Pathogens. *Phytopathology* **106**: 104–112.

Suez J, Zmora N, Zilberman-Schapira G, Mor U, Dori-Bachash M, Bashiardes S, Zur M, Regev-Lehavi D, Ben-Zeev Brik R, Federici S, et al. 2018. Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT. *Cell* **174**: 1406-1423.e16.

Suhr MJ, Hallen-Adams HE. 2015. The human gut mycobiome: pitfalls and potentials--a mycologists perspective. *Mycologia* **107**: 1057–1073.

Tropini C, Earle KA, Huang KC, Sonnenburg JL. 2017. The gut microbiome: Connecting spatial organization to function. *Cell Host Microbe* **21**: 433–442.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**: 902–903.

Vatanen T, Kostic AD, d'Hennezel E, Siljander H, Franzosa EA, Yassour M, Kolde R, Vlamakis H, Arthur TD, Hämäläinen A-M, et al. 2016. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **165**: 842–853.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* **35**: 543–548.

West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. 2018. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res* **28**: 569–580.

Wu D, Jospin G, Eisen JA. 2013. Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLOS ONE* **8**: e77033.

Zmora N, Zilberman-Schapira G, Suez J, Mor U, Dori-Bachash M, Bashiardes S, Kotler E, Zur M, Regev-Lehavi D, Brik RB-Z, et al. 2018. Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell* **174**: 1388-1405.e21.

Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, Sun H, Xia Y, Liang S, Dai Y, et al. 2019. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* **37**: 179–185.

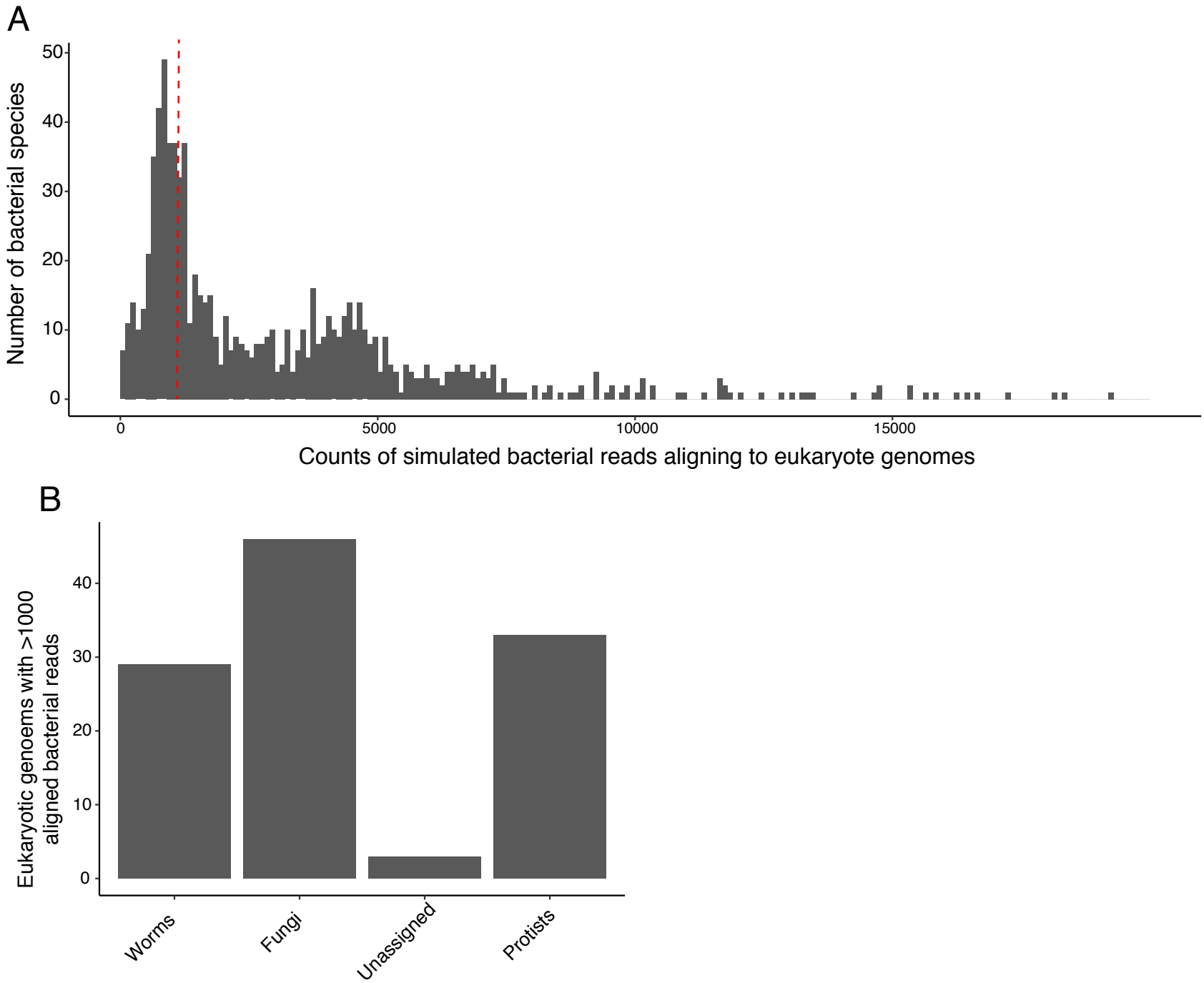


Figure 1. Human gut microbiome bacterial shotgun sequence reads are misattributed to eukaryotes. (A) Metagenomic sequencing reads were simulated from 971 species total from all major phyla in human stool (2 million reads per species) and aligned to all microbial eukaryotic genomes used to develop EukDetect. Even after stringent filtering (Methods), many species have thousands of reads aligning to eukaryotic genomes, which would lead to false detection of eukaryotes in samples with only bacteria. 111 species with > 20,000 reads aligned to eukaryotic genomes are not shown. Red dashed line indicates species with greater than 1,000 reads aligning to eukaryotic genomes. (B) Taxonomic distribution of eukaryotic genomes with >1,000 aligned bacterial reads. All major taxonomic groups of microbial eukaryotes are affected.

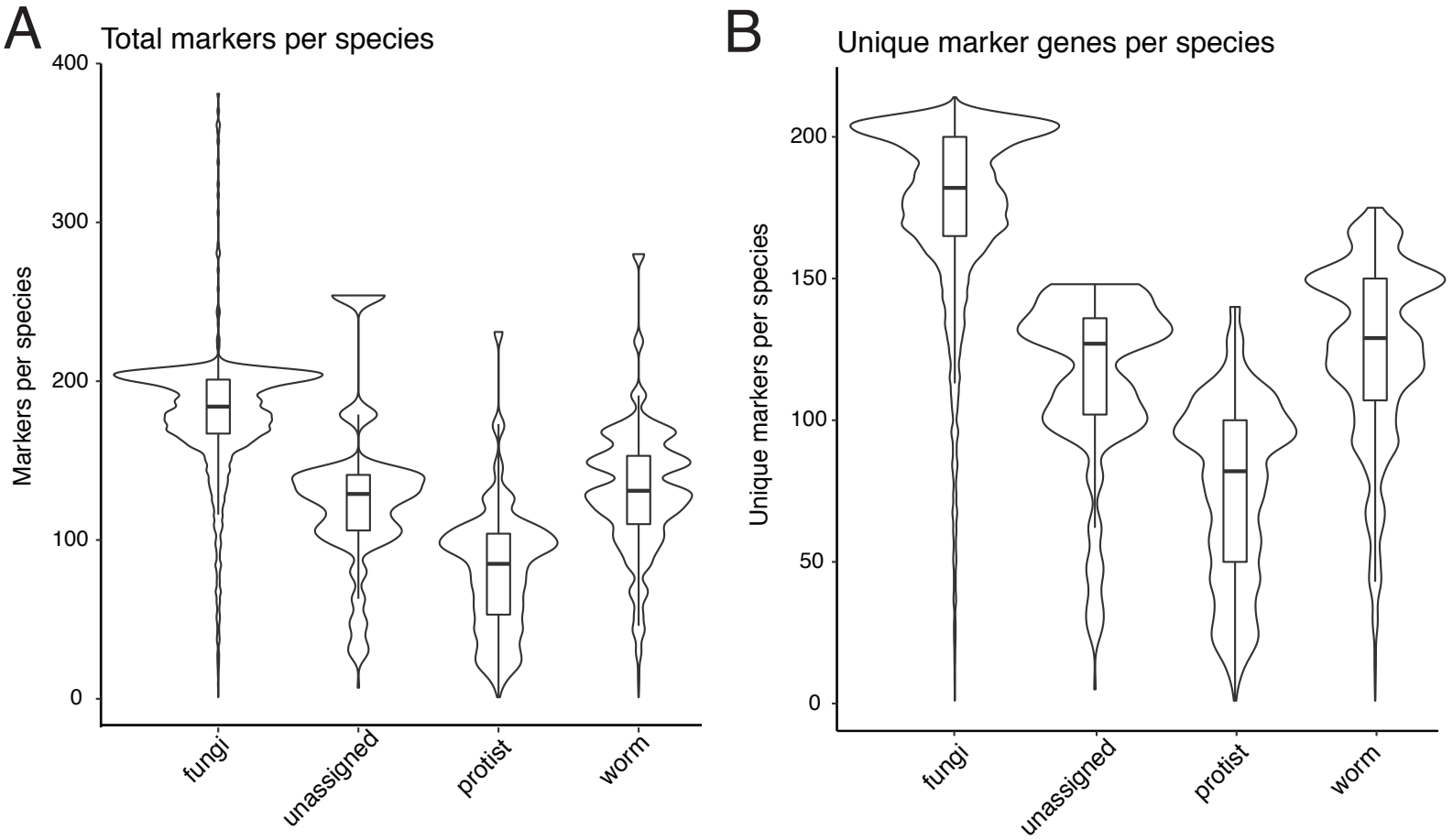


Figure 2. EukDetect database marker genes are represented in 98% of currently sequenced eukaryotes. (A) Total number of marker genes per species across taxonomic groups, including genes present in multiple copies. (B) Total number of marker genes identified per species by taxonomic group, excluding duplicates of genes.

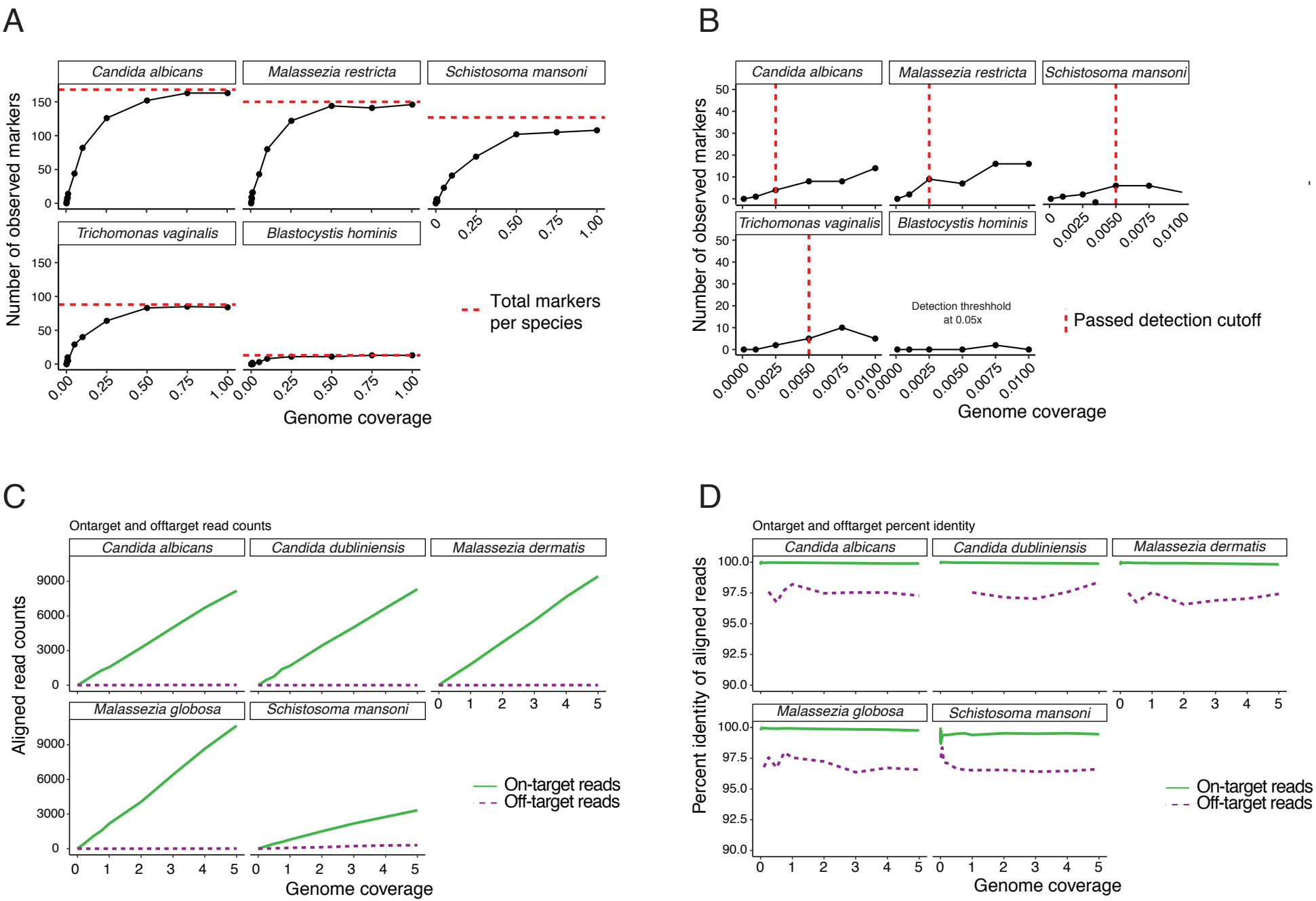


Figure 3. EukDetect pipeline performance is sensitive for yeasts, protists, and worms at low sequence coverage. (A) Number of marker genes with at least one aligned read per species up to 1x genome coverage. Horizontal red line indicates the total number of marker genes per species (best possible performance). (B) Number of marker genes with at least one aligned read per species up to 0.01x genome coverage. Vertical red line indicates a detection cutoff where 4 or more reads align to 2 or more markers. (C) Counts of reads aligned to on- and off-target marker genes (i.e, correctly classified vs. incorrectly classified) across simulated genome coverages for species with close relatives in the EukDetect database. On-target alignments dominate at all coverages. (D) Percent sequence identity of all on-target and all off-target reads across simulated genome coverages. On-target reads have consistently higher percent identity.

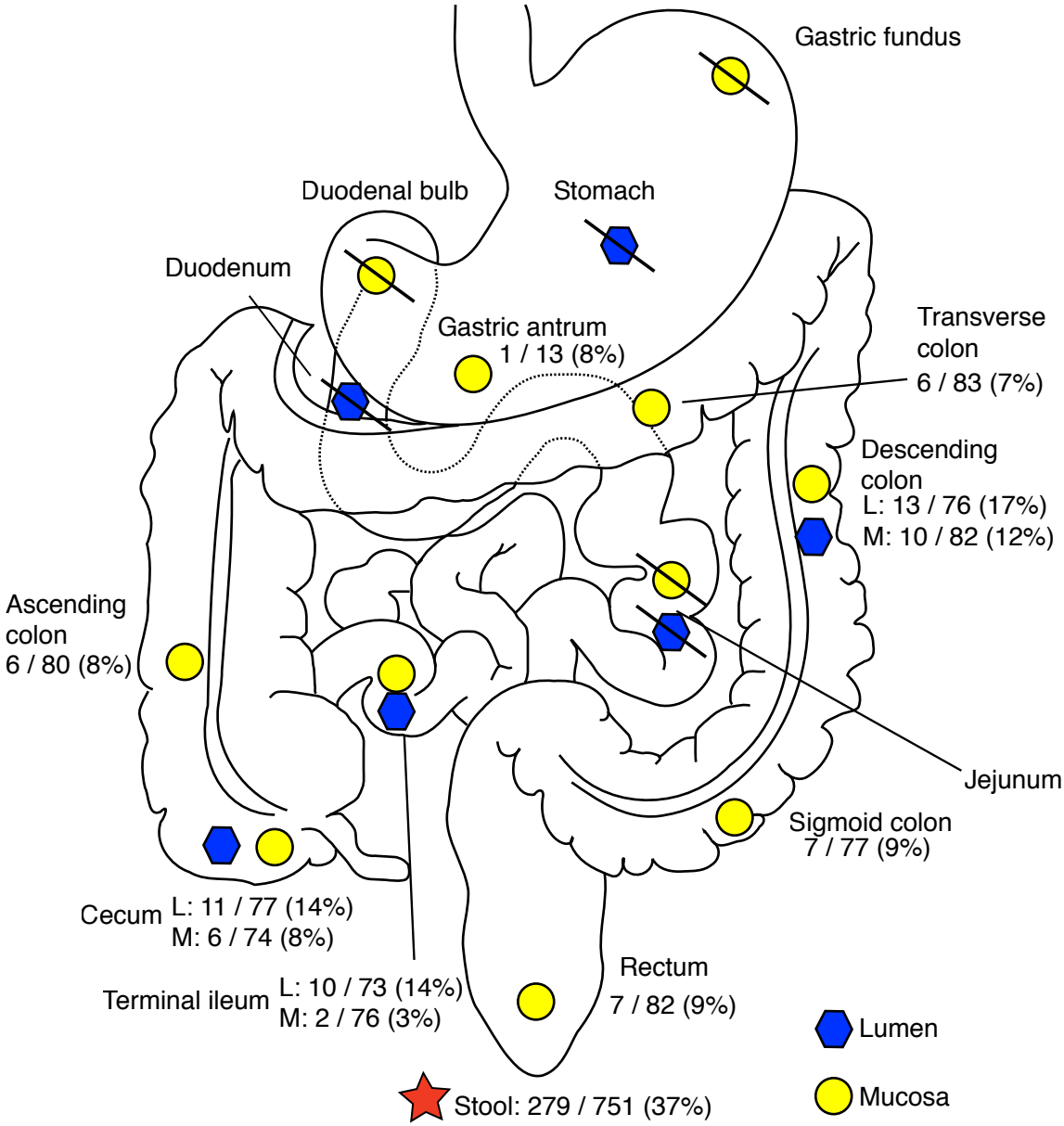
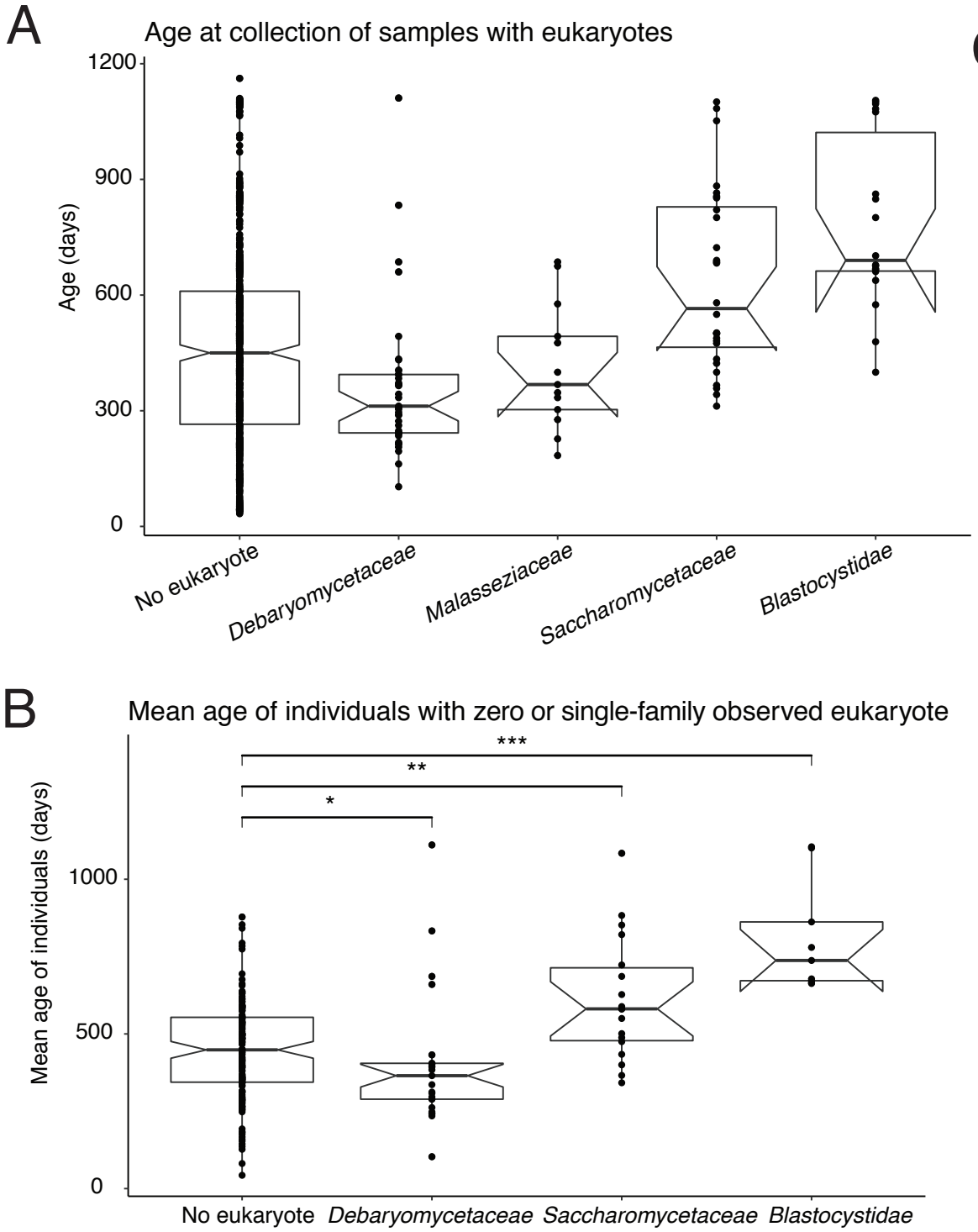


Figure 4. Distribution of eukaryotic species in the gastrointestinal tract taken from biopsies. Eukaryotes were detected at all sites in the large intestine and in the terminal ileum, in both lumen and mucosal samples. One biopsy of gastric antrum mucosa in the stomach contained a Malassezia yeast. Slashes indicate no eukaryotes detected in any samples from that site. See Figures S2 and S3 for locations of Blastocystis subtypes and locations of fungi.



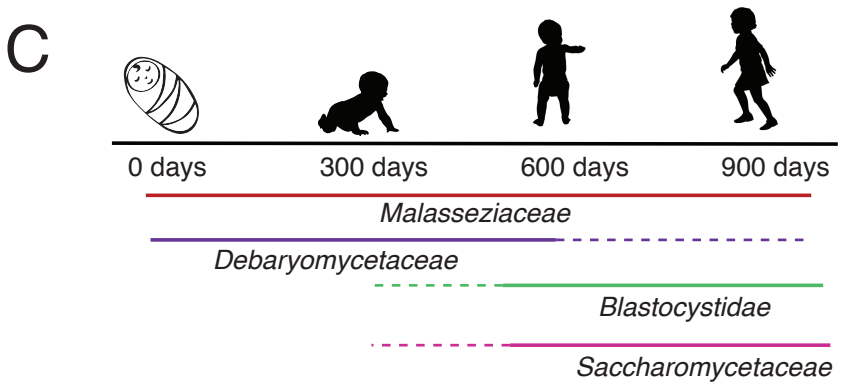


Figure 5. Changes in eukaryotic gut microbes during the first years of life. (A) Age at collection in the DIABIMMUNE three-country cohort for samples with no eukaryote or with any of the four most frequently observed eukaryotic families. (B) The mean age at collection of samples from individuals with no observed eukaryotes compared to the mean age at collection of individuals where one of three eukaryotic families. Individuals where more than one eukaryotic family was detected are excluded. Malasseziaceae is excluded due to low sample size. Group comparisons were performed with an unpaired Wilcoxon rank-sum test. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. (C) Model of eukaryotic succession in the first years of life. Debaryomycetaceae species predominate during the first two years of life but are detected later. Blastocystidae species and Saccharomycetaceae species predominate after the first two years of life, but are detected during the second year. Malasseziaceae species are constant.

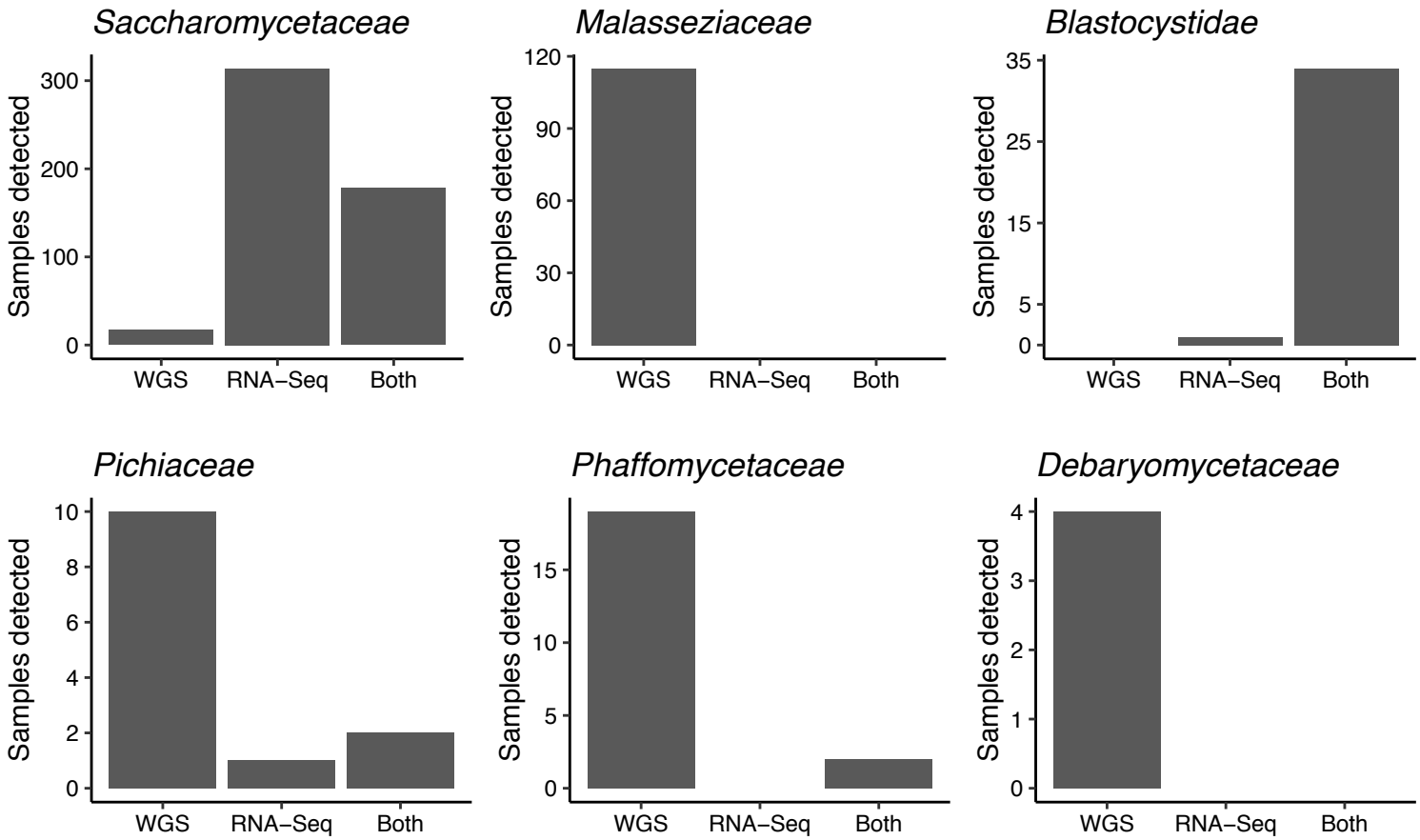


Figure 6. Detection of eukaryotes from paired DNA and RNA sequenced samples from the IHMP IBD cohort. Plots depict the most commonly detected eukaryotic families, and whether a given family was detected in the DNA sequencing alone, the RNA sequencing alone, or from both the RNA and the DNA sequencing from a sample. Some samples shown here come from the same individual sampled at different time points.