Improved eukaryotic detection compatible with large-scale automated analysis of metagenomes

Wojtek Bazant[1], Ann S. Blevins[2], Kathryn Crouch[1]*#, Daniel P. Beiting[2]*#

1. Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, United Kingdom

2. Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA

* To whom correspondence should be addressed. E-mail: kathryn.crouch@glasgow.ac.uk, beiting@upenn.edu

# Indicates co-senior authors

Keywords: metagenome, shotgun metagenomics, eukaryotes, bioinformatics, fungi, mycobiome

Abstract

Background

Eukaryotes such as fungi and protists frequently accompany bacteria and archaea in microbial communities. Unfortunately, their presence is difficult to study with 'shotgun' metagenomic sequencing since prokaryotic signals dominate in most environments. Recent methods for eukaryotic detection use eukaryote-specific marker genes, but they do not allow for quantification of eukaryote signal and do not incorporate strategies to handle the presence of eukaryotes that are not represented in the reference marker gene set.

Results

Here we present CORRAL (for Clustering Of Related Reference ALignments), a tool for identification of eukaryotes in shotgun metagenomic data based on alignments to eukaryote-specific marker genes and Markov clustering. Using a combination of simulated datasets and large publicly available human microbiome studies, we demonstrate that our method is not only sensitive and accurate but is also capable of inferring the presence of eukaryotes not included in the marker gene reference, such as novel species and strains. Finally, we deploy CORRAL on our MicrobiomeDB.org resource, producing an atlas of eukaryotes present in various environments of the human body and linking their presence to study covariates.

Conclusion

CORRAL allows eukaryotic detection to be automated and carried out at scale. Since our approach is independent of the reference used, it may be applicable to other contexts where shotgun metagenomic reads are matched against redundant but non-exhaustive databases, such as identification of novel bacterial strains or taxonomic classification of viral reads.

Background

Eukaryotic microbes are a large and phylogenetically diverse group of organisms that includes both pathogens and commensals, the latter of which are emerging as important modulators of health and disease. Protists include many important pathogens of humans and other animals, such as *Cryptosporidium*, *Toxoplasma*, *Eimeria*, *Trypanosoma*, and *Plasmodium*. Many fungi are also well-studied pathogens affecting a diverse range of hosts. For example, *Aspergillus fumigatus* is an important cause of respiratory disease in humans (1); *Magnaporthe oryzae* is the most important fungal

disease of rice globally (2); while *Pseudogymnoascus destructans* is the cause of White-Nose Syndrome, one of the most devastating diseases of bats (3). However, recent data also suggest that non-pathogenic commensal fungi are critical modulators of the human antibody repertoire (4–6), intestinal barrier integrity (7), and colonization resistance (8). The diverse array of host-microbe interactions and host phenotypes influenced by eukaryotic microbes underscores the importance of studying this class of organisms in their natural habitats. Unfortunately, the ability to carry out culture-independent analysis of eukaryotic microbes is severely hindered by their low abundance relative to bacteria, which makes accurate detection a challenge and consequently eukaryotes are commonly overlooked in metagenomic studies (9). For example, an analysis of stool metagenomes in healthy adults participating in the Human Microbiome Project reports only 0.01% reads aligning to fungal genomes (10).

Several methods have been developed to improve the detection of eukaryotes in complex samples. Targeted sequencing of internal transcribed spacer regions (ITS) is a common approach but prevents simultaneous profiling of other members of the microbiome (11). Alternatively, collections of curated fungal genomes have been successfully used for strain-level identification of *Blastocystis* from stool (12). However, pitfalls associated with non-specific or erroneous parts of reference genomes (13) combined with computational challenges associated with carrying out alignments to very large collections of reference genomes (14) limit applicability of these approaches to the discovery of eukaryotes from the vast amount of metagenomic data already available in the public domain. One attractive solution to this challenge was recently proposed in important work by Lind and Pollard (15), who base their method for sensitive and specific identification of eukaryotes in metagenomic studies, EukDetect, on alignments to a collection of over 500,000 universal, single-copy eukaryotic marker genes.

We recently sought to add the EukDetect reference database and software to our web-based resource, MicrobiomeDB.org (16), to allow for automated detection of eukaryotes across a range of human metagenomic studies currently available on the site. Since the EukDetect pipeline does not allow for adjustment of filtering thresholds and is not packaged for containerized deployments, we decided to implement our own tool built with a more flexible software architecture. Our approach retains the EukDetect reference database, as well as the use of Bowtie2 (17) since it has been shown to be a sensitive aligner (18). To better understand the filtering process used by EukDetect, we carried out a simulation-based evaluation. We observed that filtering of read alignments based on mapping quality (MAPQ) scores (19) – though necessary for EukDetect's high specificity – removes correct alignments for which Bowtie2 has inferior but closely scored alternatives.

Considering that the difficulty of detecting a taxon may be affected by similarity of its marker gene sequences to its most similar neighbor led us to develop CORRAL (for Clustering Of Related Reference ALignments), an approach for processing marker gene alignments based on exploiting information in shared alignments to reference genes through Markov clustering. This allows for sensitive and accurate detection which also extends to species not present in the reference but which are similar to one or more known taxa present in the reference.

Results

Species-specific impact of MAPQ filtering

To evaluate how read mapping and filtering parameters influence eukaryotic detection, we carried out a series of simulations using the EukDetect database of eukaryotic marker genes as both a source of reads with known identity and a reference to which to align these reads. When metagenomic reads are simulated from this reference and then simply mapped back, thus exactly matching the reference, they are accurately mapped to the correct taxon with a recall (fraction of correctly mapped reads among all reads) and precision (fraction of correctly mapped reads among all reads that mapped) of 95.1% for

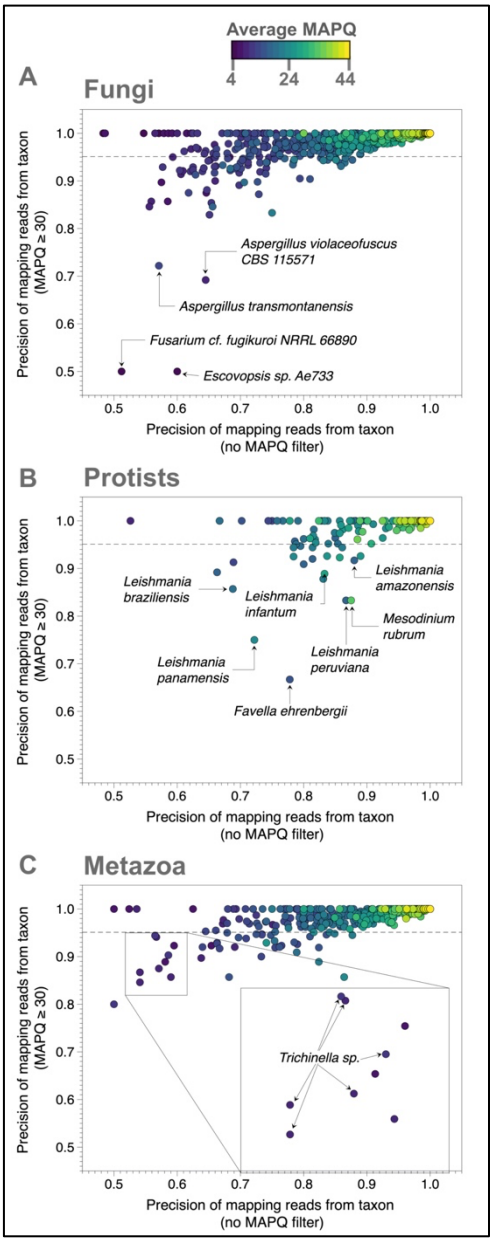

Figure 1: Species-specific impact of MAPQ filtering. Precision of read mapping comparing MAPQ ≥ 30 (Y-axis) versus no MAPQ filter (X-axis) for A) fungi, B) protists, and C) metazoa. Points are colored by average MAPQ scores. Horizontal dashed line indicates prefilter precision and recall of 95.1%. Select taxa for which the MAPQ filter either only marginally improved or impaired precision are labeled.

each. Applying a MAPQ ≥ 30 filter increases precision to 99.7% and decreases recall to 91.7%. This translates to 92% of the simulated reads mapping with MAPQ ≥ 30, with only 0.3% of these mapping incorrectly, and out of the remaining 8%, almost half mapping incorrectly.

Examining these data at the level of individual taxa from which the reads were sourced reveals a structural component to the difficulty of mapping the reads, as well as the efficacy of the MAPQ filter (**Figure 1**). For example, out of 3977 taxa whose reads map back to the reference, reads from 1908 taxa map with 100% precision (**Figure 1, upper rightmost points**), and after applying the MAPQ ≥ 30 filter, 1105 more taxa map with 100% precision. Despite this clear improvement after filtering, 146 taxa still map with precision lower than the pre-filter overall total of 95.1% (**Figure 1, dashed line**). This set of taxa includes numerous species of *Aspergillus* (**Figure 1A**), *Leishmania* (**Figure 1B**), and *Trichinella* (**Figure 1C**), all of which are important pathogens of humans and other mammals. Furthermore, filtering based on MAPQ decreases precision for five taxa, including the fungi *Fusarium cf. fujikuroi NRRL 66890* and *Escovopsis sp. Ae733* (**Figure 1A**), and the protists *Favella ehrenbergii*, *Leishmania peruviana*, and *Mesodinium rubrum* (**Figure 1B**). Taken together, these results suggest that relying on MAPQ filter alone may not allow for robust detection of multiple eukaryotes of public health importance.

Since the diversity of eukaryotic microbial life extends far beyond the currently discovered species, let alone species present in the EukDetect reference (20), we next modified this simulation above to evaluate the possibility of detecting 'novel' species. To do this, species-level markers in the EukDetect reference were split into a holdout set of 371 taxa from which we simulated reads that were then mapped back to the remaining 3343 taxa in the EukDetect reference, thus mimicking a scenario in which a metagenomic sample contains reads from eukaryotes not represented in the reference. In this circumstance, the MAPQ ≥ 30 filter is not on average an improvement. Same-genus precision and recall are 82% and 30%, respectively, without the filter. Applying the MAPQ filter results in a similar precision (83.6%) but a much-diminished recall of 7%. Source taxon is a structural component here as well – applying the MAPQ ≥ 30 filter increases the number of taxa which only map to the correct genus from 48 to 152 but increases the number of taxa that fail to map from 49 to 175.

There is extensive strain variation in complex microbial communities, so we next set out to evaluate the ability to identify eukaryotes when a sample contains a novel strain of a species present in the reference database **(Figure 2)**. We carried out a third simulation in which sampled reads were mutated before mapping back to the reference. As mutation rate increases, recall declines from 95.1% to less than

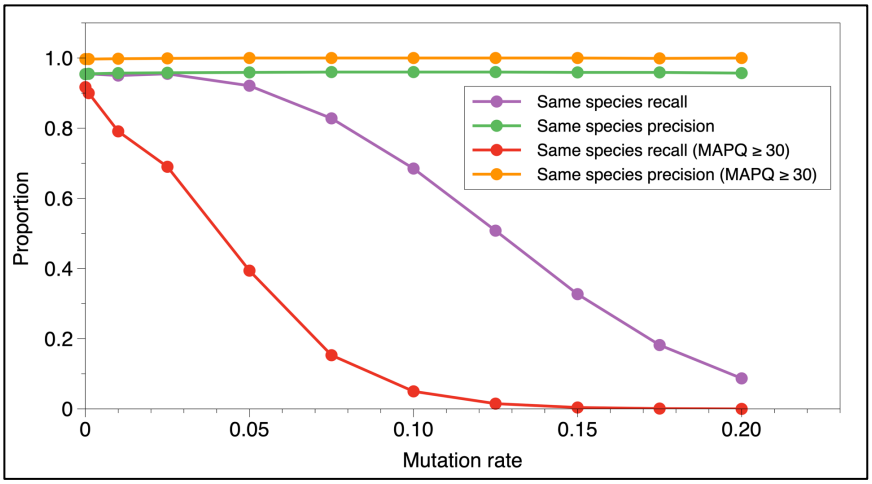

Figure 2: Mutation rate influences MAPQ filter performance. Proportion of taxa where recall or precision are as described (legend), as mutation rate is increased from 0 to 0.2.

10% when mutation rate is 0.2. In this range, precision stays between 95-96% for all reads and ≥ 99% for reads with MAPQ ≥ 30 – an observation consistent with previous reports of bowtie2 preserving precision over recall [22]. Applying the MAPQ ≥ 30 filter results in a rapid decline in recall. For example, when mutation rate is 0.1, recall is 68.3% overall but drops to 5.0% when a MAPQ ≥ 30 filter is applied. These results indicate there may be many taxa which match the reference sufficiently closely to allow for sensitive detection, but only if one does not apply the MAPQ ≥ 30 filter.

CORRAL leverages Markov clustering for reference-based eukaryote detection

To address the challenges described above and to fully leverage the valuable eukaryotic marker gene reference database created by Lind and Pollard (15), we developed CORRAL (Clustering Of Related Reference Alignments) as a Nextflow workflow wrapping a Python module. CORRAL retrieves sequence files, aligns reads to the EukDetect reference of markers, and produces a taxonomic profile through a multi-step process (**Figure 3**). First, we run Bowtie2 and keep all alignments that are at least 60 nucleotides in length (**Figure 3, step 1**), ensuring that sequence matches contain enough information to be marker-specific. We then run Markov Clustering (MCL) on a graph composed of marker genes as nodes and counts of shared alignments as edge weights to obtain marker clusters (**Figure 3, step 2**). Next, percent match identities of alignments are calculated and aggregated by marker to obtain an identity average for each marker gene, as well as per cluster to obtain a cluster average (**Figure 3, step 3**). Each marker whose identity average is lower than the cluster average is considered an inferior representation of signal in the sample, and taxa with ≥ 50% of such markers are rejected (**Figure 3, step 4**). Remaining taxa are then gathered into taxonomic clusters using MCL on counts of multiply aligned reads (**Figure 3, step 5**), which allows us to incorporate ambiguity of identification into any taxa reported. Unambiguous matches (defined as having average alignment identity of at least 97%, and two different reads aligned to at least two markers) are reported (**Figure 3, step 6**), while other taxa in clusters where there are any unambiguous matches reported are rejected. Finally, for each remaining taxon cluster, we report it as one hit if it is a strong ambiguous match (defined as having at least four markers and eight reads) by joining names of taxa in the cluster and prepending with a "?" (**Figure 3, step 7**).

This approach represents a set of default parameters – based on our observations in simulated and human microbiome data – that can be altered when configuring CORRAL. Additionally, CORRAL has rich reporting capabilities, including the ability to quantify abundance of eukaryotes using a 'copies per million (CPM)' metric (see Methods).

CORRAL infers the presence of novel species

To demonstrate CORRAL's ability to handle reads from a taxon that is not in the provided reference, we returned to a holdout set simulation like the one described above. We simulated a metagenomic dataset consisting of 338 samples, each containing a single 'novel' eukaryotic species (from the holdout set) at 0.1x genome coverage. Using this data set, we compared CORRAL, EukDetect with default

settings, EukDetect with a relaxed MAPQ filter of ≥ 5, and a simple "4 reads + 2 markers (MAPQ ≥ 30)" scenario in which taxa are reported when at least four reads align with MAPQ ≥ 30 to at least two

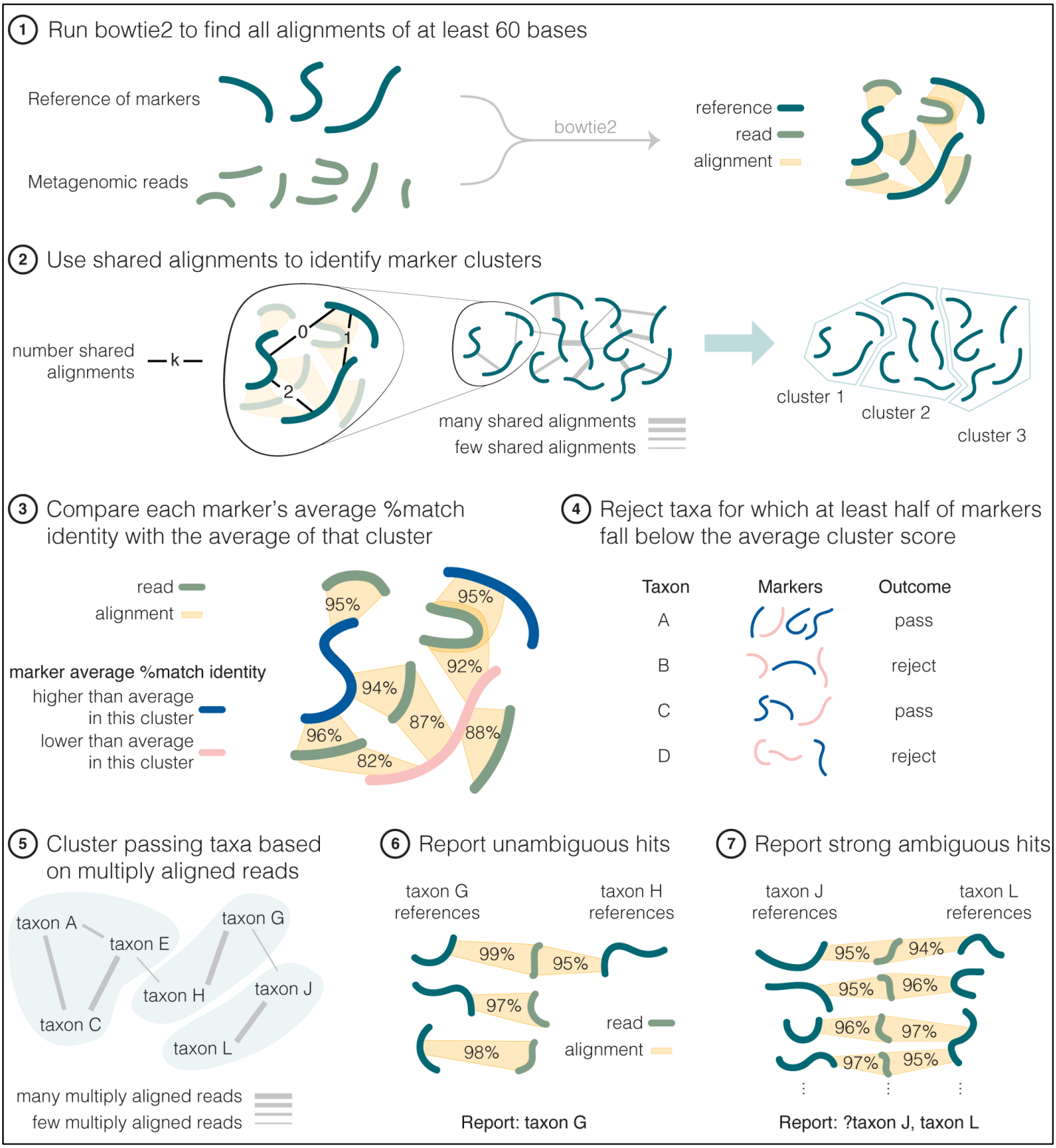

Figure 3: The CORRAL workflow. Schematic showing all seven steps of the CORRAL workflow.

markers. We evaluated accuracy based on whether a single taxon was reported, rather than no hit or more than one hit (since each simulated sample was created to only contain one novel species). If there was a hit, we also evaluated taxonomic proximity – whether the hit was of the same genus as the novel species.

Out of the four methods outlined above, CORRAL performs best at reporting a single novel species as a single result in the correct genus (**Figure 4**). EukDetect's proportion of 'No result' is higher than for "EukDetect (MAPQ ≥ 5)" and exactly the same as "4 reads + 2 markers (MAPQ ≥ 30)", which indicates that use of a MAPQ ≥ 30 filter makes inferring novel species more difficult. Furthermore, although relaxing the MAPQ filter from ≥ 30 to ≥ 5 decreases the proportion of 'No results' **(Figure 4,** leftmost group of bars**)**, it substantially increases the proportion of results that yielded multiple hits that were not in the same genus **(Figure 4,** rightmost group of bars**).** This indicates that simply modifying EukDetect to use a less stringent filter (MAPQ ≥ 5) would not be a desirable adjustment, because while it improves sensitivity to detect eukaryotic signal in a sample, it compromises the ability to recognize that this signal consists of only a single species. In contrast, CORRAL yielded the lowest proportion of 'No result' and the highest proportion of single hits to the same genus as the novel species. We also noticed that CORRAL showed the highest proportion of single hits that were not in the same genus. To better understand the source of these potential 'off-target' hits, we separated CORRAL results into ambiguous and non-ambiguous, which is easily achieved since the report produced by CORRAL flags ambiguous hits with a '?'. This closer examination of the CORRAL results showed that most of these hits to other genera were indeed flagged as ambiguous and therefore could easily be excluded by the user. Taken together, these data show that CORRAL demonstrates higher sensitivity and specificity than all other methods tested.

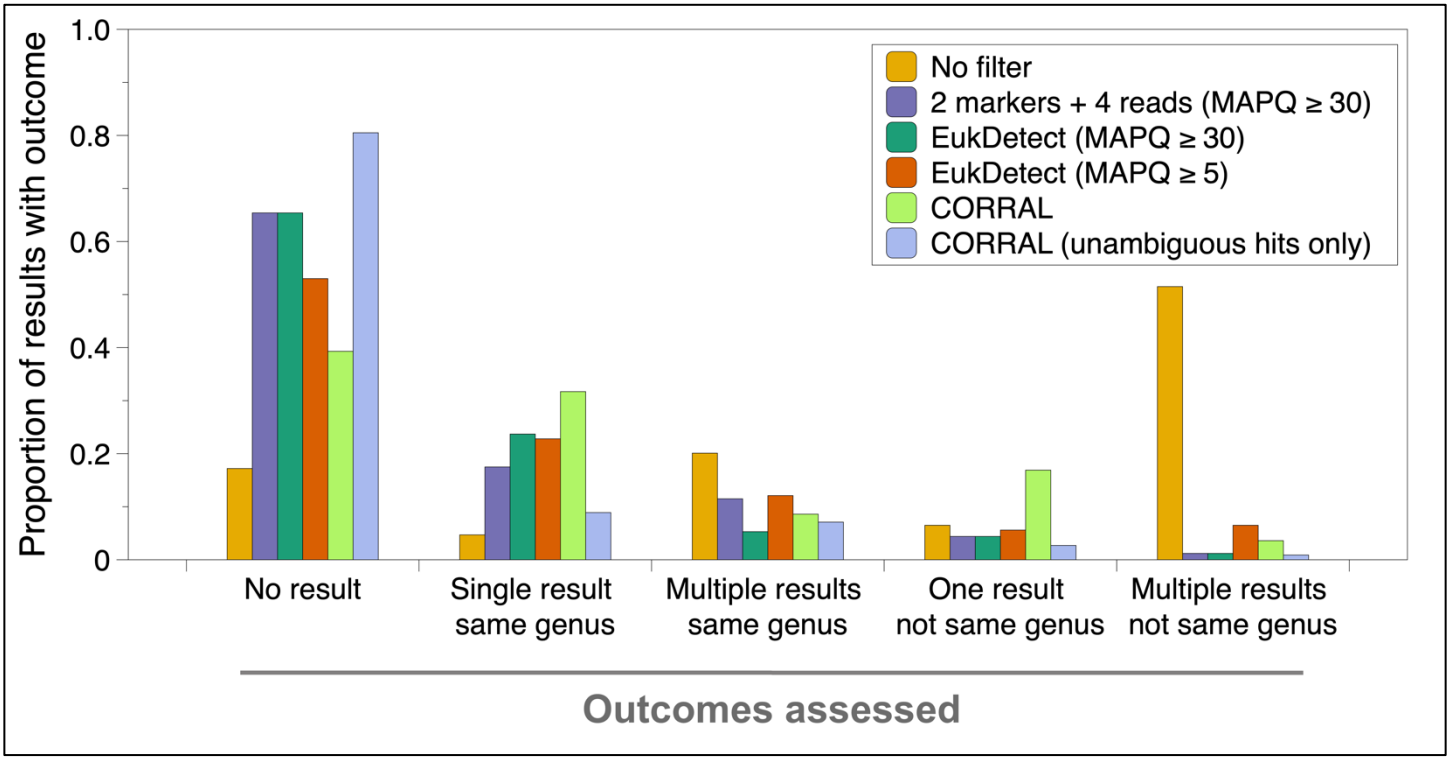

Figure 4: CORRAL yields high sensitivity and specificity when predicting the presence of eukaryotes in metagenomic data. Proportion of results (Y-axis) corresponding to each of 5 possible outcomes (X-axis). CORRAL (light green bar) balances high sensitivity (low proportion of 'No result'; leftmost group of bars) with high specificity (highest proportion of single results in the same genus as the 'novel' taxon from the holdout set). CORRAL allows users to understand which hits may be ambiguous. Considering only unambiguous hits shows that CORRAL has the lowest proportion of single or multiple results occurring outside of genus of the 'novel' taxon (lavender bar).

Evaluating CORRAL on human microbiome data

To move beyond the simulations described above we next tested CORRAL on data from real

microbiome studies where some expectations exist about which eukaryotes might be present. We first evaluated the DIABIMMUNE study (21), for which 136 data points about 30 different eukaryotes were reported across 1154 samples in the original EukDetect publication (15). Processing these same 1154 samples, CORRAL is in exact concordance with EukDetect on 122/136 data points and adds an additional 97 data points. CORRAL reports common taxa at a higher frequency. For example, *S. cerevisiae* is detected by CORRAL 67 times, while EukDetect only identifies this organism 31 times. The other additional hits detected by CORRAL, but not EukDetect, consist primarily of yeast and other fungi that have been previously reported in the human gut, and thus seem plausible. In summary, these results are evidence that, when applied to real metagenomic data, CORRAL improves sensitivity for eukaryote detection without compromising specificity.

Importantly, CORRAL differs from EukDetect in how it treats reads that might originate from a novel species. For example, in sample G78909 from DIABIMMUNE, EukDetect reports *Penicillium nordicum*, while our method reports a novel *Penicillium*. In sample G80329, our method agrees with EukDetect regarding detection of *Candida parapsilosis*, and also identifies the sample as positive for *C. albicans*. Finally, in sample G78500 EukDetect reports *Saccharomyces cerevisiae* and *Kazachstania unispora*, which our method reports to be reads from a single taxon: a strain of *S. cerevisiae* that differs from the reference strain.

Automating eukaryote detection with CORRAL

Top 15 eukaryotic taxa detected
across 8 metagenomic studies on MicrobiomeDB

Taxon	Sample count	Eukaryote type
Malassezia restricta	364	Fungi
Candida albicans	255	Fungi
Saccharomyces cerevisiae	190	Fungi
Purpureocillium lilacinum	181	Fungi
Malassezia globosa	129	Fungi
Candida parapsilosis	114	Fungi
Blastocystis sp. subtype 3	88	Protist
Clavispora lusitaniae	59	Fungi
Blastocystis	47	Protist
Cyberlindnera jadinii	43	Fungi
Blastocystis sp. subtype 1	42	Protist
Candida tropicalis	39	Fungi
Malassezia	39	Fungi
Malassezia sympodialis	19	Fungi
Candida	16	Fungi

Table 1: CORRAL expands eukaryote identification when deployed at scale on MicrobiomeDB.org. Top 15 eukaryotes (by prevalence) detected across eight metagenomic studies encompassing 6337 samples.

In addition to making our software simple to install through pip and easily parametrized, we integrated CORRAL into the automated data loading workflow for our open-science platform, MicrobiomeDB.org. As of Release 27 (17 May 2022), the site contains 6337 samples from 8 published metagenomic studies (21–28). Automated analysis of these samples by CORRAL occurs at the time a study is loaded for public release onto the database website. In the case of these 6337 samples, this results in the identification of eukaryotes in 1453/6337 (23%) of the samples, yielding 2084 data points for 190 different eukaryotic taxa. A large majority, 1851/2084 or 89% of these data points, are fungal taxa. Of the 233 data points for non-fungal eukaryotes detected in these samples, 200 (86%) are species belonging to the genus *Blastocystis*, one of the most common protozoan parasites found in the human GI tract (29). A summary of the top 15 most frequently observed eukaryotes (**Table 1**) reveals that *Malassezia restricta*, a common commensal and opportunistic pathogen; and *Candida albicans*, a prevalent component of gut flora, are the top two most common fungal taxa identified on MicrobiomeDB using CORRAL, detected in 364 and 255 samples, respectively.

Integration of CORRAL in MicrobiomeDB enables exploration of associations between eukaryotic microbes and host phenotypes

Although CORRAL can be run as stand-alone software, one advantage of integrating this software into MicrobiomeDB is that the results can viewed across many different studies, sample types and in many

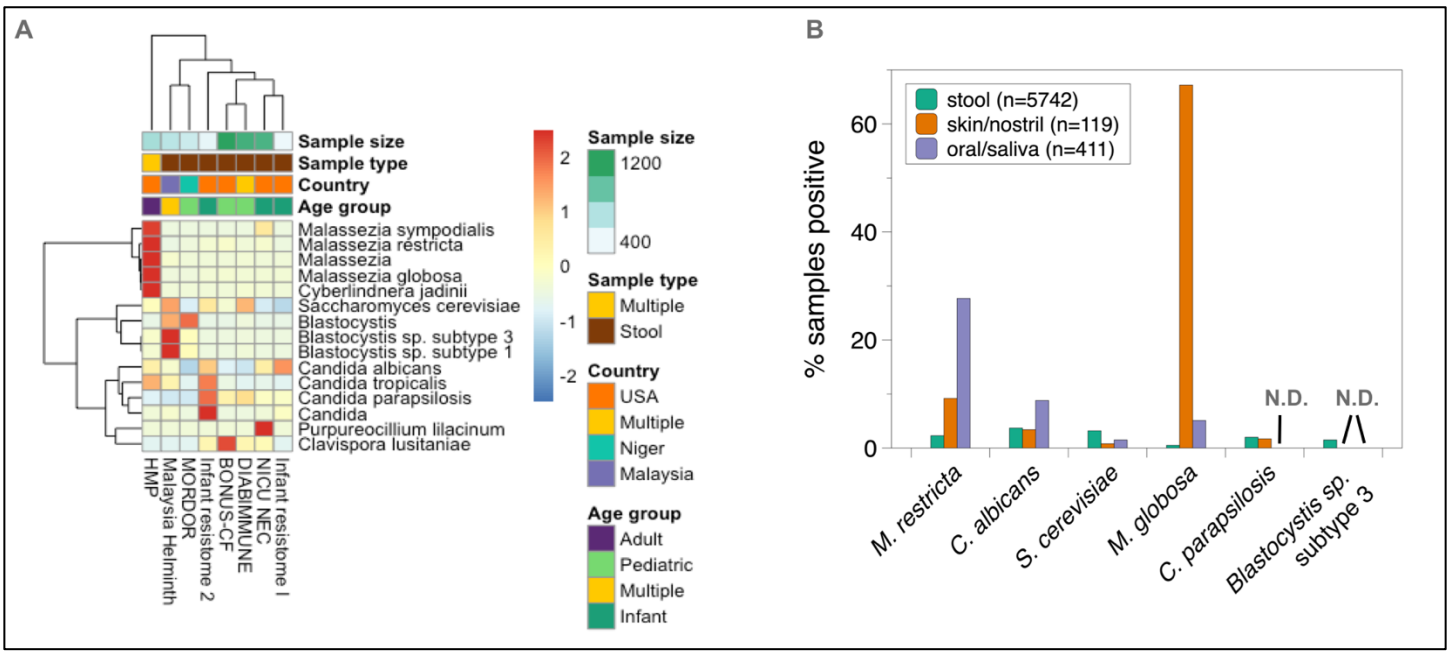

Figure 5: Integration of CORRAL results with study metadata on MicrobiomeDB. A) Heatmap showing row Z scores for the top 15 eukaryotes (by prevalence) across all eight metagenomic datasets currently publicly available on MicrobiomeDB.org. Study name and metadata are shown below and above the heatmap, respectively. B) % of all stool, skin swab or nostril swab (skin/nostril), or oral swab or saliva (oral/saliva) metagenomic samples on MicrobiomeDB that were positive for six selected eukaryotes (X-axis) by analysis by CORRAL.

different experimental contexts, thus allowing researchers to identify associations between eukaryotes and study metadata, potentially leading to novel hypotheses (**Figure 5**). The metagenomic data currently available on MicrobiomeDB were generated from distinct geographic regions and from participants that vary in age from infant to adult. When we viewed the top 15 most prevalent eukaryotic taxa across all 8 datasets on MicrobiomeDB, in the context of this study metadata, interesting trends emerged. For example, species of *Malassezia* were primarily found in the Human Microbiome Project study (HMP) **(Figure 5A),** likely because this study included sample types other than stool. A closer look at *Malassezia* species prevalence by sample type across all 8 studies showed that over 60% of the 119 skin and nostril swab samples were positive for *M. globosa*, while *M. restricta* was more restricted to the oral cavity and saliva **(Figure 5B)**. *Blastocystis* sp. were primarily observed in samples from studies carried out in Niger and Malaysia (MORDOR and Malaysia Helminth studies) **(Figure 5A),** suggesting that these protists may be more prevalent in lower- and middle-income countries. Similarly, *Candida* species were most prevalent in infant samples. The fungi *Clavispora lusitaniae* and *Purpureocillium lilacinum* were each primarily observed in the BONUS-CF and NICU NEC studies, respectively. Interestingly, careful analysis of *P. lilacinum* by the authors of the NICU NEC study identified this organism as a reagent contaminant (30). Taken together, these results suggest that implementing CORRAL at database-scale can accelerate the discovery of species-specific niches, improve identification of taxa that arise from spurious results or contamination, and help researchers link eukaryotic taxa to environmental covariates within and across studies.

CORRAL enables quantification of eukaryotes in metagenomic data

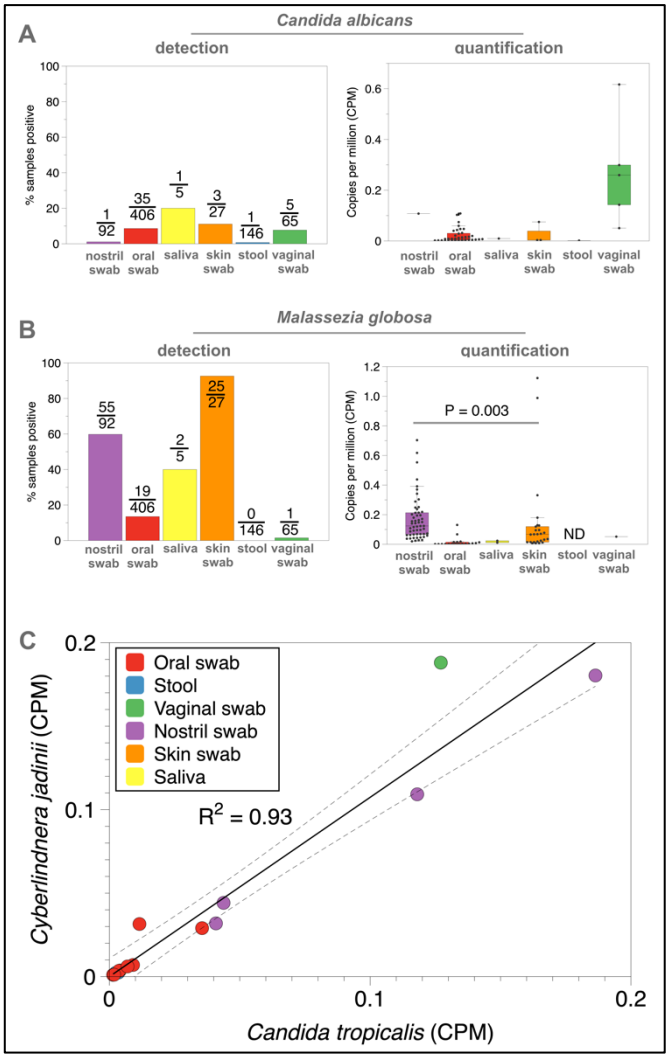

Figure 6: Quantification of eukaryotes by CORRAL. Comparison of detection (presence/absence) and quantification (copies per million; CPM) by CORRAL for A) *Candida albicans* and B) *Malassezia globosa* in the Human Microbiome Project (HMP) study. For detection, number of samples testing positive out of total samples assayed is shown on each bar. P value from Wilcoxon rank-sum test comparing levels of *M. globosa* between nostril swabs and skin swabs. C) Correlation of CPM for *Candida tropicalis* and *Cyberlindnera jadinii* in HMP.

In addition to the presence/absence detection of eukaryotes, CORRAL also reports the relative abundance of the eukaryotes it detects, thus opening the door to using many of the same visualization and analytics already familiar to the microbiome community for interpreting bacterial census data. To demonstrate this, we focused on the Human Microbiome Project (HMP) study, since it is the only metagenomic study on our MicrobiomeDB resource that contains multiple sample types. We compared CORRAL's detection data with relative abundance data for two of the most prevalent fungal taxa detected across all studies on our site, *Candida albicans* and *Malassezia globosa* **(Figure 6)**. Although CORRAL detected *Candida albicans* in less than 10% of vaginal swabs, these positive samples had the highest levels of this organism compared to all other sample types examined **(Figure 6A).** Although the HMP participants were healthy adults, these data may point to individuals that either had or were at risk of developing vaginal yeast infections. Similarly, *Malassezia globosa* was detected in nearly every skin swab examined **(Figure 6B, left)**, consistent with numerous reports of this fungus as a skin-dwelling microbe, yet the abundance of *M. globosa* is significantly higher in the nasal cavity, compared to skin swabs **(Figure 6B, right).** These data underscore how quantitative data can impact our understanding of host-microbe interactions. Although this analysis focused on associations between fungal taxa and sample type, a similar analysis could be carried out using any available experimental metadata loaded into MicrobiomeDB (*e.g.* fungal taxa by clinical status).

Quantification data produced by CORRAL also allow conventional statistical analyses to be readily applied, either manually by downloading data from MicrobiomeDB, or directly within the website using data visualization applications ('apps') built using the R/Shiny (16,31). For example, we used the 'Correlation App' on MicrobiomeDB to search for co-associated fungal taxa. This analysis identified a strong positive correlation between the abundance of the fungi *Candida tropicalis* and *Cyberlindnera jadinii* in the HMP dataset **(Figure 6C;** R^2 = 0.93**)**. Interestingly, this correlation was evident even in sample types where the relative abundance of these organisms was low or high **(Figure 6C;** oral swabs vs. nostril swabs, respectively**).** Importantly, due to

the relatively low prevalence of eukaryotes in metagenomic samples, observing this type of correlation may only be possible when eukaryotic data can be mined at scale, using large collections of studies. Whether or how these two fungi interact is beyond the scope of this study, nevertheless these data underscore the ability to use CORRAL in conjunction with MicrobiomeDB to generate hypotheses about fungal community interactions which can then be experimentally tested.

Discussion

CORRAL (Cluster of Related Reference ALignments) is open-source software that uses multiple alignments and Markov clustering to achieve high sensitivity and specificity for identification of eukaryotes in metagenomic data, while also enabling inferences about the presence of eukaryotes not represented in the reference. We highlight the utility of this software using simulated metagenomic samples containing 'novel' species and strains. We also deploy CORRAL on our open-science platform, MicrobiomeDB.org, which allowed automated processing of thousands of samples currently on the site, thus generating the first cross-study atlas of eukaryotes from metagenomic data. With CORRAL now part of our standard data loading workflow for metagenomic data on MicrobiomeDB, this atlas will continue to grow as new studies are loaded. This demonstrates the value of combining robust software with web-based tools for conducting large-scale screens of metagenomic data, thereby creating a resource that will allow investigators to access eukaryotic data from a vast range of sample types and studies, irrespective of whether the original study investigators intended to examine eukaryotes in their data.

The high cost of metagenomic sequencing, the relative low abundance of most eukaryotes in the microbiome, and the inherent limitation of reference-based methods for identification of taxa remain major challenges to identification of eukaryotes. CORRAL helps to address some of these issues by being able to work with minimal information required to plausibly report the presence and abundance of eukaryotes, even when the source reads do not perfectly match the marker gene reference. Future improvements in genome assembly will provide more complete information on eukaryote-specific genomic sequences which could be used to create a larger reference with more taxa and more sequences per taxon, improving both specificity and sensitivity of hits reported by CORRAL.

Our strategy of clustering of related read alignments could be further improved by making use of information about taxonomic similarity between reference sequences. Not relying on external data about similarity of different proteins has the benefit of flexibility but lacks the capacity to act on implied 'improbability' of reported taxa. For example, it is relatively unlikely that a sequenced sample containing reads which map to multiple closely related *Leishmania* species does in fact contain different species of *Leishmania*, because the reference sequences are highly similar, and the species readily hybridize [31]. Conversely, reads sharing alignments to markers across a large taxonomic distance are more likely to come from a single source because of relative implausibility of the sample containing multiple eukaryotes of unknown genera – for example, they might all be contamination from a metazoan host. Incorporating such speculations about 'likely' and 'unlikely' results into a detection method is an ambitious undertaking, because it involves making and modeling assumptions about vast numbers of eukaryotic taxa, most of which have not been sequenced and not yet well studied. It could, however, yield methods with a more natural choice of threshold parameters, and further gains in sensitivity and specificity. Since the computational approach used by CORRAL is independent of the reference sequences used, our software could potentially be applied to processing alignments to any reference that is anticipated to be redundant and incomplete, and where reads are expected to map with varying identity. This includes identification of bacteria to the strain-level resolution required in genomic epidemiology, as well as taxonomic classification of viral reads to reference sequences (32), identification of antibiotic resistance genes (33), or bacterial virulence genes (34).

Methods

Simulations

We used wgsim (35) to sample 100 basepair reads with base error rate of 0 from the EukDetect reference (the 1/23/2021 version, latest at time of writing, consisting of BUSCOs from OrthoDB (36)). Bowtie2 (17) was used to align reads to references with identical settings to those used in EukDetect: the end to end (default) mode and the `--no-discordant` flag.

To check correctness of simulated alignments, we retrieved the rank of the nearest taxon containing source and match by using the ETE toolkit and the NCBI database version dated 2020/1/14 packaged with EukDetect. Alignments were deemed correct if the source and match were of the same species, or genus in case of hold-out analysis where the species was missing from the reference by design.

Our formulas to calculate precision and recall are as used in the OPAL method of assessing taxonomic metagenome profilers (37): precision is a fraction of correctly mapped reads among all reads that are mapped, and recall is a fraction of correctly mapped reads among all reads.

When simulating whole samples, we obtained 338 simulated samples from a holdout set of 371 taxa, because we skipped 33 cases in which `wgsim` considers the sequences too fragmented to source reads at a set coverage, and errors out. The number of reads to source per marker to obtain 0.1 coverage was calculated as previously described (38).

To run EukDetect, we edited the default config file such that it lists the simulated samples. To run “EukDetect (MAPQ >= 5)”, we additionally modified the source code of our local installation. To run “4 reads + 2 markers (MAPQ >= 30)”, we ran CORRAL configured to use these three filters instead of the default procedure described in this publication.

CORRAL quantifies abundance for each found taxon with ‘copies per million’ (CPMs) as the number of reads assigned to the taxon normalized by marker length and sequencing depth, in line with the quantity being calculated in the integrated metagenomic profiling tool, HUMAnN (39).

Deploying CORRAL on MicrobiomeDB.org

CORRAL is integrated into the standard MicrobiomeDB workflow for metagenomic datasets (see *https://github.com/VEuPathDB/MicrobiomeWorkflow*) along with bioBakery tools for bacterial abundance estimation. CORRAL output is loaded as both binary (presence/absence) and quantitative Copies Per Million (CPM) values for each sample and can be used along with other sample details related to the collection, processing and analysis of data for filtering and stratification of bacterial abundance data as well as directly for exploring correlations between eukaryotic abundance and other sample data.

Acknowledgements

We thank the authors of the metagenomic studies cited in this work (21–28) for their assistance with loading and representing their data on MicrobiomeDB. This work was partially supported by a grant from the Bill and Melinda Gates Foundation (D.P.B.) and Astarte Medical (D.P.B.). The funders had no role in data collection and analysis, decision to publish, or preparation of this manuscript.

Data availability

All our software is publicly available under the MIT license: CORRAL (*github.com/wbazant/CORRAL*), its main Python module, (*github.com/wbazant/marker_alignments*), and a mix of Python, Make, and Bash scripts to produce simulations, comparisons, and figures for this publication (*github.com/wbazant/markerAlignmentsPaper*).

All results are publicly viewable and downloadable on MicrobiomeDB. In addition, the following files are available as supplemental material:

LINK: Simulated whole samples - results for different methods

LINK: Simulated reads - per-species breakdown and aggregate stats

LINK: Comparison of CORRAL and EukDetect on DIABIMMUNE study

LINK: Summary of MicrobiomeDB results

Bibliography

1. Latgé JP. Aspergillus fumigatus and aspergillosis. Clin Microbiol Rev. 1999 Apr;12(2):310–50.

2. Wilson RA, Talbot NJ. Under pressure: investigating the biology of plant infection by Magnaporthe oryzae. Nat Rev Microbiol. 2009 Mar;7(3):185–95.

3. Wibbelt G, Kurth A, Hellmann D, Weishaar M, Barlow A, Veith M, et al. White-nose syndrome fungus (Geomyces destructans) in bats, Europe. Emerging Infect Dis. 2010 Aug;16(8):1237–43.

4. Doron I, Leonardi I, Li XV, Fiers WD, Semon A, Bialt-DeCelie M, et al. Human gut mycobiota tune immunity via CARD9-dependent induction of anti-fungal IgG antibodies. Cell. 2021 Feb 18;184(4):1017-1031.e14.

5. Doron I, Mesko M, Li XV, Kusakabe T, Leonardi I, Shaw DG, et al. Mycobiota-induced IgA antibodies regulate fungal commensalism in the gut and are dysregulated in Crohn's disease. Nat Microbiol. 2021 Dec;6(12):1493–504.

6. Ost KS, O'Meara TR, Stephens WZ, Chiaro T, Zhou H, Penman J, et al. Adaptive immunity induces mutualism between commensal eukaryotes. Nature. 2021 Aug;596(7870):114–8.

7. Leonardi I, Gao IH, Lin W-Y, Allen M, Li XV, Fiers WD, et al. Mucosal fungi promote gut barrier function and social behavior via Type 17 immunity. Cell. 2022 Mar 3;185(5):831-846.e14.

8. Jiang TT, Shao T-Y, Ang WXG, Kinder JM, Turner LH, Pham G, et al. Commensal fungi recapitulate the protective benefits of intestinal bacteria. Cell Host Microbe. 2017 Dec 13;22(6):809-816.e4.

9. Laforest-Lapointe I, Arrieta M-C. Microbial eukaryotes: a missing link in gut microbiome studies. mSystems. 2018 Apr;3(2).

10. Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, et al. The gut mycobiome of the Human Microbiome Project healthy cohort. Microbiome. 2017 Nov 25;5(1):153.

11. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci USA. 2012 Apr 17;109(16):6241–6.

12. Beghini F, Pasolli E, Truong TD, Putignani L, Cacciò SM, Segata N. Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. ISME J. 2017 Dec;11(12):2848–63.

13. R Marcelino V, Holmes EC, Sorrell TC. The use of taxon-specific reference databases compromises metagenomic classification. BMC Genomics. 2020 Feb 27;21(1):184.

14. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol. 2018 Nov 16;19(1):198.

15. Lind AL, Pollard KS. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. Microbiome. 2021 Mar 3;9(1):58.

16. Oliveira FS, Brestelli J, Cade S, Zheng J, Iodice J, Fischer S, et al. MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments. Nucleic Acids Res. 2018 Jan 4;46(D1):D684–91.

17. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.

18. Thankaswamy-Kosalai S, Sen P, Nookaew I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics. 2017 Jul;109(3–4):186–91.

19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

20. Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. How many species are there on Earth and in the ocean? PLoS Biol. 2011 Aug 23;9(8):e1001127.

21. Vatanen T, Kostic AD, d'Hennezel E, Siljander H, Franzosa EA, Yassour M, et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. Cell. 2016 May 5;165(4):842–53.

22. Hayden HS, Eng A, Pope CE, Brittnacher MJ, Vo AT, Weiss EJ, et al. Fecal dysbiosis in infants with cystic fibrosis is associated with early linear growth failure. Nat Med. 2020 Feb;26(2):215–21.

23. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe. 2015 Feb 11;17(2):260–73.

24. Olm MR, Bhattacharya N, Crits-Christoph A, Firek BA, Baker R, Song YS, et al. Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria. Sci Adv. 2019 Dec 11;5(12):eaax5727.

25. Gibson MK, Wang B, Ahmadi S, Burnham C-AD, Tarr PI, Warner BB, et al. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. Nat Microbiol. 2016 Mar 7;1:16024.

26. Gasparrini AJ, Wang B, Sun X, Kennedy EA, Hernandez-Leyva A, Ndao IM, et al. Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. Nat Microbiol. 2019 Dec;4(12):2285–97.

27. Doan T, Hinterwirth A, Worden L, Arzika AM, Maliki R, Abdou A, et al. Gut microbiome alteration in MORDOR I: a community-randomized trial of mass azithromycin distribution. Nat Med. 2019 Aug 12;25(9):1370–6.

28. Tee MZ, Er YX, Easton AV, Yap NJ, Lee IL, Devlin J, et al. Gut microbiome of helminth infected indigenous malaysians is context dependent. BioRxiv. 2022 Jan 24;

29. Silberman JD, Sogin ML, Leipe DD, Clark CG. Human parasite finds taxonomic home. Nature. 1996 Apr 4;380(6573):398.

30. Olm MR, West PT, Brooks B, Firek BA, Baker R, Morowitz MJ, et al. Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. Microbiome. 2019 Feb 15;7(1):26.

31. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. shiny: Web Application Framework for R [Internet]. R Package; 2019 [cited 2020 Jan 8]. Available from: https://CRAN.R-project.org/package=shiny

32. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. mSphere. 2018 Apr;3(2).

33. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020 Jan 8;48(D1):D517–25.

34. Liu B, Zheng D, Zhou S, Chen L, Yang J. VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res. 2022 Jan 7;50(D1):D912–7.

35. lh3/wgsim: Reads simulator [Internet]. [cited 2022 Mar 1]. Available from: https://github.com/lh3/wgsim

36. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019 Jan 8;47(D1):D807–11.

37. Meyer F, Bremges A, Belmann P, Janssen S, McHardy AC, Koslicki D. Assessing taxonomic metagenome profilers with OPAL. Genome Biol. 2019 Mar 4;20(1):51.

38. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014 Feb;15(2):121–32.

39. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife. 2021 May 4;10. sciwheel/placeholder/bibliography