bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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 sequencing techniques, since prokaryotic signals dominate in most environments. Recent methods for eukaryotic detection use eukaryote-specific marker genes, but they do not yet incorporate strategies to handle presence of unknown eukaryotes.

Results

Here we present CORALE (for Clustering of 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. We finally deploy CORALE on our MicrobiomeDB.org resource, demonstrating adequate reliability and throughput.

Conclusion

CORALE allows eukaryotic detection to be automated and carried out at scale. Since our approach is independent of the reference used, it is 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 ("Aspergillus Fumigatus and Aspergillosis" 1999); *Magnaporthe oryzae* is the most important fungal disease of rice globally (Wilson and Talbot 2009); while *Pseudogymnoascus destructans* is the cause of White-Nose Syndrome, one of the most devastating diseases of bats (Wibbelt et al. 2010). However, recent data also suggests that non-pathogenic commensal fungi are significant as modulators of the human antibody repertoire (Doron, Leonardi, et al. 2021)(Doron, Mesko, et al. 2021)(Ost et al. 2021), intestinal barrier integrity (Leonardi et al. 2022), and colonization resistance (Jiang et al. 2017). This 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 (Laforest-Lapointe and Arrieta 2018). For example, an analysis of stool metagenomes in healthy adults participating in the Human Microbiome Project (HMP 2012) reports only 0.01% reads aligning to fungal genomes (Nash et al. 2017).

Several methods have been developed to improve the detection of eukaryotes in complex samples. Targeted sequencing of internal transcribed spacer regions (ITS) (Schoch et al. 2012) is a common approach but prevents simultaneous profiling of other members of the microbiome. Alternatively, collections of curated fungal genomes have been successfully used for strain-level identification of *Blastocystis* from stool (Beghini et al. 2017). However, pitfalls associated with non-specific or erroneous parts of reference genomes (R Marcelino, Holmes, and Sorrell 2020) combined with computational challenges associated with carrying out alignments to very large collections of reference genomes (Breitwieser, Baker, and Salzberg 2018) 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 (Lind and Pollard 2021), who base their method for sensitive and specific identification of eukaryotes in metagenomic studies, EukDetect, on alignments to over 500,000 universal, single-copy eukaryotic marker genes.

We recently sought to add EukDetect results to our web-based resource, MicrobiomeDB.org (Oliveira et al. 2018), in order to allow for querying samples based on presence 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 it is not packaged for containerised deployments, we decided to implement our own tool built with a more flexible software architecture. We retained EukDetect's reference of marker genes with the aim of producing directly comparable results, and kept `bowtie2` (Langmead and Salzberg 2012) since it has been shown to be a sensitive aligner (Thankaswamy-Kosalai, Sen, and Nookaew 2017). To better understand the filtering process used by EukDetect, conducted a simulation-based evaluation. We noticed that a filter based on mapping quality (MAPQ) (Li et al. 2009) scores, though necessary for EukDetect's high specificity, removes correct alignments for which `bowtie2` has inferior but closely scored alternatives.

Considering how the difficulty of detecting a taxon may be affected by similarity of its sequence to nearest references led us to develop CORALE (for Clustering of Reference Alignments), an approach to processing marker gene alignments based on exploiting information in shared alignments to reference genes through clustering. This allows for sensitive and accurate detection while also enabling inference of novel species not present in the reference.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Species-specific impact of MAPQ filtering

We base our study of alignments to marker genes on the software for sampling reads `wgsim` (Li 2011), the aligner `bowtie2` used with EukDetect's settings for aligning the reads we sample, and a custom Python program for capturing properties of alignments like the MAPQ score and aggregating results.

Not surprisingly, when reads in a metagenomic sample are simulated from the reference and then aligned back, thus exactly matching the reference, they are accurately mapped to the correct taxon with a precision and recall of 95.1%. Applying a MAPQ $\geq$ 30 filter increases precision to 99.7% and decreases recall to 91.7%. This translates to 8% of reads mapping with MAPQ $<$ 30, and just under half of those being incorrectly mapped. In contrast, of reads with MAPQ $\geq$ 30 only 0.3% are incorrectly mapped.

Stratifying these values by the source taxon of the reads reveals a structural component to the difficulty of mapping the reads, as well as the efficacy of the MAPQ $\geq$ 30 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 point), and after applying the MAPQ $\geq$ 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, dotted 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 decreases precision for five taxa: fungi *Fusarium cf. fujikuroi NRRL 66890* and *Escovopsis sp. Ae733*, and protists *Favella ehrenbergii*, *Leishmania peruviana*, and *Mesodinium rubrum*.

Since the diversity of eukaryotic taxa extends far beyond the currently discovered species, let alone species present in the EukDetect reference (Mora et al. 2011), we then modify this experiment to study the possibility of detecting 'novel' species; we split the species level markers in the EukDetect reference into a holdout set of 371 taxa and a remaining reference of 3343 taxa, sample from the holdout set, and align to an index built from the remaining reference. In these circumstances, the MAPQ $\geq$ 30 filter is not on average an improvement: same-genus precision and recall are 82% and 30% without the filter, compared to 83.6% precision and a much-diminished recall of 7% with the filter. Source taxon is a structural component here, too: while applying the MAPQ $\geq$ 30 filter increases the number of taxa which only map to the correct genus from 48 to 152, it also increases the number of taxa which don't map at all from 49 to 175.

The `wgsim` tool lets us sample reads with mutations, and thus study detection of 'semi-novel' taxa like a non-reference strain of a known species. We modify the setup of aligning reads back to the reference as in the first experiment to run `wgsim` with a mutation rate parameter μ, and increase it in increments. Over μ in range of [0, 0.200], recall declines from 95.1% to below 10% while precision stays between 95-96% for all reads and $\geq$ 99% for reads with MAPQ $\geq$ 30 - an observation of bowtie2 preserving precision over recall consistent with e.g. [16]. We then observe that applying the MAPQ $\geq$ 30 filter causes recall to drop much faster: for example, when $\mu = 0.1$, recall is 68.3% for all reads and 5.0% with the MAPQ $\geq$ 30 filter.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

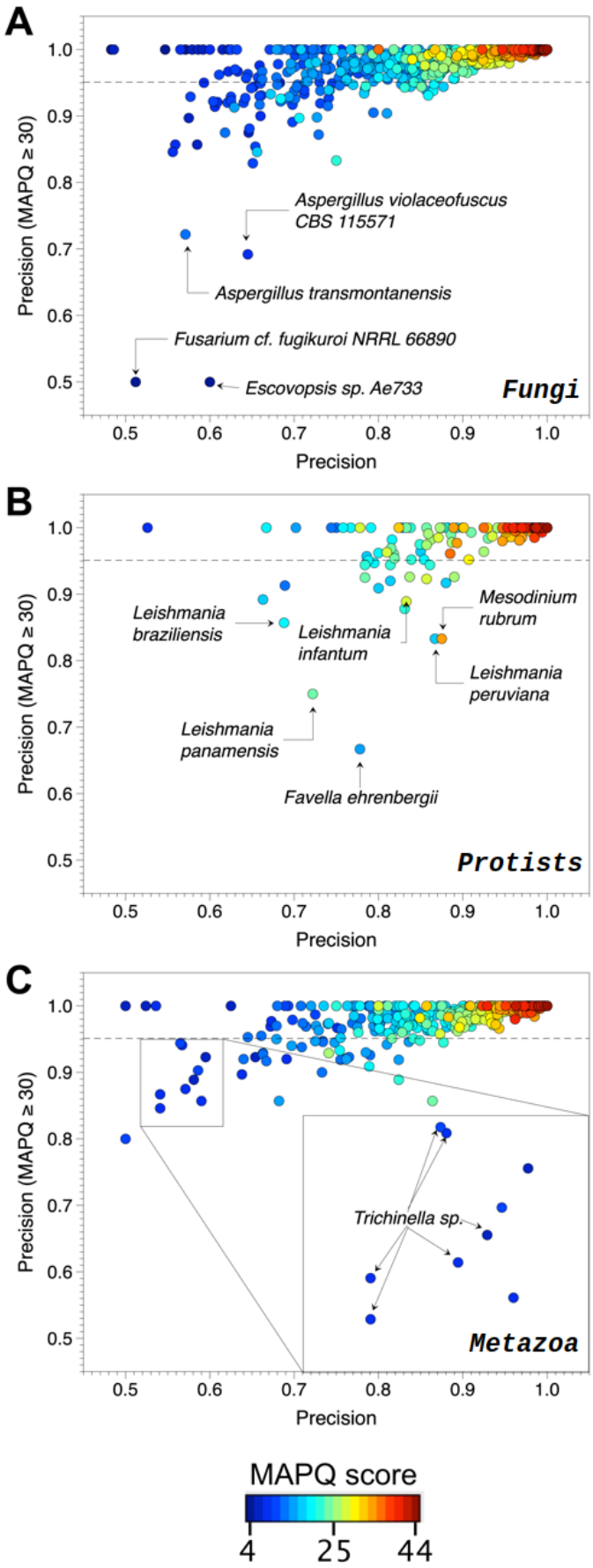

Figure 1: **The improvement from the MAPQ ≥ 30 filter varies by species.**

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CORALE leverages Markov clustering for reference-based eukaryote detection

We wrote our software, CORALE - for Clustering of Reference Alignments - as a Nextflow workflow wrapping a Python module. It retrieves sequence files, aligns them to the reference of markers, and produces a taxonomic profile through a seven-step procedure (Figure 2). First, we run `bowtie2` and keep all alignments that are at least 60 nucleotides in length (Figure 2, step 1), which ensures that the matches contain enough information to be marker-specific. We then run Markov Clustering (MCL) on a graph of marker genes as nodes and counts of shared alignments as edge weights (Figure 2, step 2) to obtain marker clusters. Next, we calculate % match identities of alignments (Figure 2, step 3) and aggregate them by marker to obtain an identity average for each marker gene, as well as per cluster to obtain a cluster average. Each marker whose identity average is lower than the cluster average is an inferior representation of signal in the sample, so we next reject each taxon with $\geq 50\%$ of such markers (Figure 2, step 4). We then group remaining taxa into taxon clusters using MCL with counts of multiply aligned reads (Figure 2, step 5), which allows us to reflect ambiguity of identification in reporting the hits. We then report unambiguous matches (defined as having average alignment identity of at least 97%, and two different reads aligned to at least two markers) as is (Figure 2, step 6), while rejecting other taxa in taxon clusters where there were any unambiguous matches reported. Finally, for each remaining taxon cluster, we report it as one hit if 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 2, step 7).

A CORALE user can alter this procedure and adjust any thresholds. We set above parameter values as default based on our observations in simulated and human microbiome data, but the software also supports other filters developed in the course of our experimentation. One example is a filter based on the fraction of primary to secondary alignments for each taxon, which is a simpler alternative to the filter based on marker cluster averages but seems to be slightly inferior. Additionally, CORALE has rich reporting capabilities - for example, we produce 'copies per million (CPMs)', a quantitative estimate of abundance calculated as number of reads normalized by marker length and sequencing depth.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

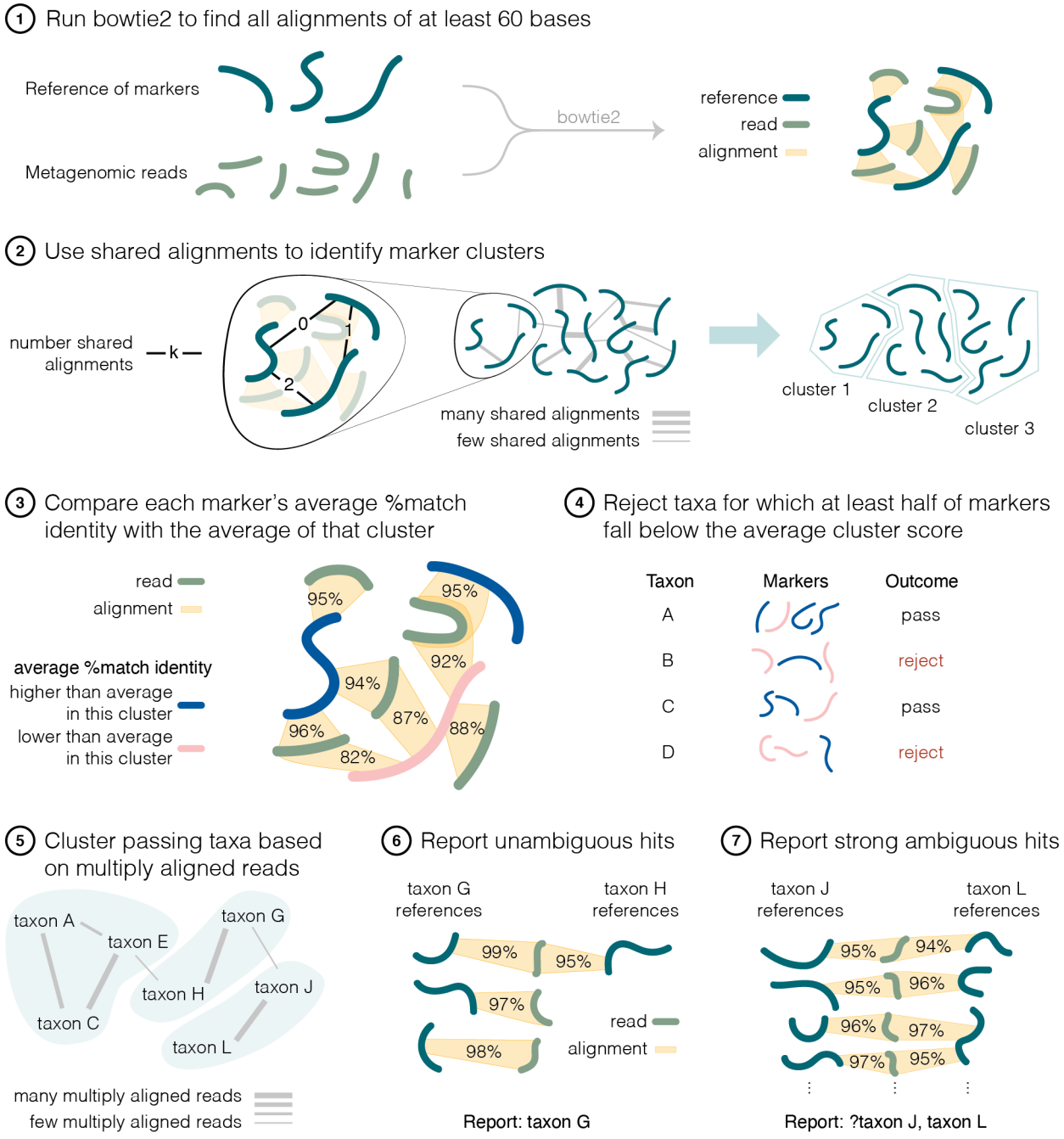

Figure 2: **CORALE - schematic**

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CORALE is capable of inferring presence of novel species

To demonstrate CORALE's ability to sensibly interpret a signal from a taxon that is not in the provided reference, we return to the holdout set and remaining reference described before, and prepare an input of 338 samples each with a single 'novel' eukaryotic species at 0.1 genome coverage. We compare CORALE, EukDetect, and two candidate tools whose outputs provide additional context. The first one is "EukDetect (MAPQ $\geq$ 5)": a version of EukDetect modified to filter MAPQ $\geq$ 5 instead of MAPQ $\geq$ 30, and the second one is "4 reads + 2 markers (MAPQ $\geq$ 30)": a run of `bowtie2` like in EukDetect followed by reporting taxa where at least four reads align with MAPQ $\geq$ 30 to at least two markers. We judge correctness measured by taxonomic proximity - whether results are the same genus as the 'novel' source of reads - as well as correctness measured by correct plurality of results.

Out of the four tools, CORALE does best at reporting a single unknown species as a single result in the correct genus (Figure 3). EukDetect's proportion of No Hits is higher than for "EukDetect (MAPQ $\geq$ 5)" and exactly the same as "4 reads + 2 markers (MAPQ $\geq$ 30)", which indicates that the MAPQ $\geq$ 30 filter makes inferring novel species more difficult.

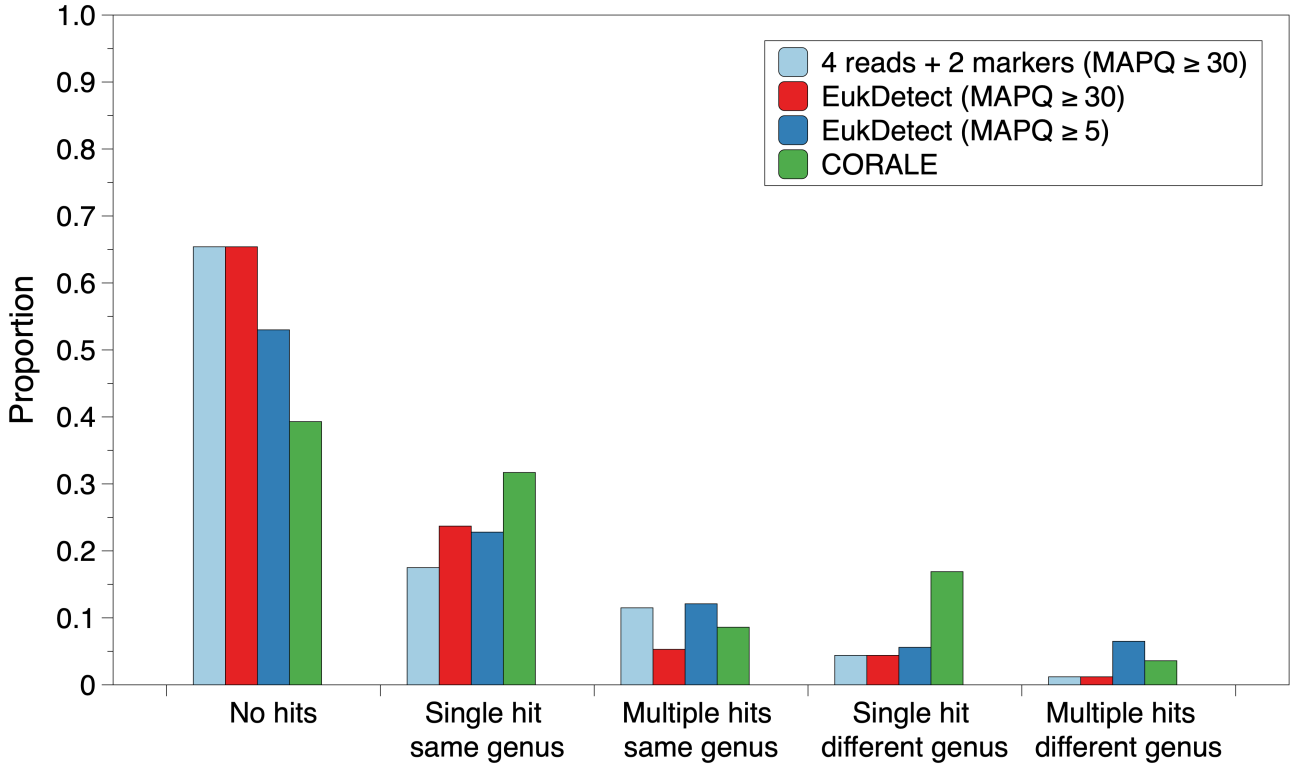

Figure 3: **Simulated unknown taxa - 0.1 coverage**

In the results above, fractions of samples where signal is detected for CORALE, EukDetect, "EukDetect (MAPQ $\geq$ 5)", and "4 reads + 2 markers (MAPQ $\geq$ 30)" are respectively 0.607, 0.346, 0.47, and 0.346, and fractions of detected signal reported as one species are respectively 0.8, 0.812, 0.604, and 0.632. This indicates modifying EukDetect to filter on MAPQ $\geq$ 5 would not be a desirable adjustment, because while it improves the tool's ability to recognise eukaryotic signal in the sample, it compromises the tool's ability to recognise that this signal consists of only a single species. Our method maintains the second ability while improving the first one.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Evaluating CORALE on human microbiome data

To test our method on data where expectations exist about which eukaryotes might be present, we turn to the the DIABIMMUNE study (Vatanen et al. 2016), for which 136 data points about 30 different eukaryotes were reported across 1154 samples in the original EukDetect publication (Lind and Pollard 2021). Processing these same 1154 samples, CORALE is in exact concordance with EukDetect on 122/136 data points, and adds additional 97 data points. CORALE reports common taxa at a higher frequency, for example, *S. cerevisiae* appears 67 times, while EukDetect only identifies this organism 31 times. This suggests CORALE is more sensitive than EukDetect. Furthermore, since the other additional hits are not obviously implausible - they consist primarily of yeast and other fungi that have been previously reported in the human gut - there is evidence that CORALE does not compromise specificity.

Importantly, CORALE 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 an additional *C. albicans*. Finally, in sample G78500 EukDetect reports *Saccharomyces cerevisiae* and *Kazachstania unispora*, which our method reports to be reads from one taxon slightly different from the reference *Saccharomyces cerevisiae*.

Automating eukaryote detection with CORALE

In addition to making our software freely available, we integrate CORALE into the automated data loading workflow for our open-science platform, MicrobiomeDB.org. As of Release 25 (2 Dec 2021), the site contains 5113 samples from 6 published metagenomic studies (Hayden et al. 2020) (Vatanen et al. 2016) (Kostic et al. 2015) (Olm et al. 2019) (Gibson et al. 2016) (Gasparrini et al. 2019). Most of the eukaryotes detected by CORALE are fungal - CORALE identifies 97 distinct fungal species across 1661/5113 (32%) metagenomic samples. A summary of the top 10 most frequently observed fungi (Figure 4A) reveals that *Malassezia restricta*, a common commensal and opportunistic pathogen, and *Candida albicans*, a prevalent component of gut flora, are most commonly detected, with presence in 363 and 229 samples, respectively. Since these results are integrated with all other sample annotations on MicrobiomeDB, users can easily identify associations between eukaryotes and metadata (Figure 4B and 4C). For example, *Malassezia globosa* (identified in 130 samples) is primarily found on skin and nostrils, while *C. albicans* (Figure 4B and 4C), *C. parapsilosis*, *Clavispora lusitaniae*, and *S. cerevisiae* are all primarily or exclusively found in stool.

A

Top 10 fungal species detected

across all metagenomic studies on MicrobiomeDB

species	sample count
Malassezia restricta[1]	363
Candida albicans[2]	229
Purpureocillium lilacinum[3]	181
Saccharomyces cerevisiae[2]	145
Malassezia globosa[1]	130
Candida parapsilosis[2]	113
Clavispora lusitaniae[2]	60
Cyberlindnera jadinii	43
Candida tropicalis	32
Malassezia sympodialis[1]	17

[1] Primarily on skin and nostrils
[2] Primarily or exclusively in stool
[3] Exclusively in infant gut

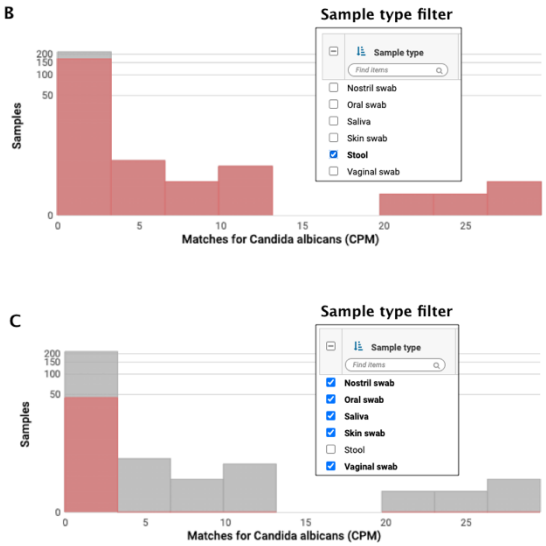

Figure 4: **MicrobiomeDB - top fungal species**

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Use cases and limitations of CORALE

CORALE is a freely available tool that aims to report on eukaryotes in metagenomes without overwhelming the results by false positives. Thanks to our novel approach, we achieve better sensitivity, similar specificity, and additional capabilities, and our tool is reliable enough to let us integrate eukaryotic detection into our open-science platform, MicrobiomeDB.org, and process thousands of samples. CORALE empowers the microbiome research community with a means of setting up broad screens of metagenomic data to identify samples where eukaryotes are present.

Importantly, using shotgun metagenomics followed by processing with CORALE is not without limitations. A primary hurdle for this approach is the high cost of WGS sequencing to the depth required to detect most eukaryotes. CORALE addresses this setback gracefully, being able to work with minimal information required to plausibly report different kinds of hits. Additionally, future improvements in genome assembly would 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 CORALE.

Future work

Our approach 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.

Additionally, the efficacy of using many alignments per read in combination with clustering demonstrated by CORALE shows that to develop new metagenomics tools, it can be worthwhile to view protein sequences as a similarity-based network: naturally occuring proteins form isolated clusters of similar sequence (Smith 1970). With additional theoretical work, this view could become a basis for predictions about presence of eukaryotes, potentially providing probabilistic estimates of certainty on reported results.

Conclusion

CORALE (for Clustering of Reference Alignments) is a tool for identification of eukaryotes in shotgun metagenomic studies in which the results are not overwhelmed by false positives. While CORALE is based on the same marker gene reference as EukDetect, it does not use EukDetect's approach to filtering, most notably not including the MAPQ $\geq$ 30 filter which we show to have species-specific impact on results. The CORALE approach, based on multiple alignments and Markov clustering, results in sensitive and accurate detection, and is capable of inferring presence of eukaryotes not included in the reference. We highlight this feature using simulated samples with 'novel' species, as well as data from DIABIMMUNE, a large infant gut metagenome study. We also successfully deploy CORALE on our MicrobiomeDB.org resource, demonstrating the appropriateness of our method for large-scale screens of metagenomic data for the purpose of detecting eukaryotes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

To simulate reads, we use `wgsim` (Li 2011) to sample from EukDetect's reference (the 1/23/2021 version, latest at time of writing, consisting of BUSCOs from OrthoDB (Kriventseva et al. 2019)). We use `bowtie2` (Langmead and Salzberg 2012) to align reads to references, in end to end (default) mode and the `--no-discordant` flag as in EukDetect. When using `wgsim` we set read length to 100, and base error rate to 0.

To check correctness of simulated alignments, we retrieve the rank of the nearest taxon containing source and match by using the ETE toolkit (Huerta-Cepas, Serra, and Bork 2016) and the NCBI database version dated 2020/1/14 packaged with EukDetect. We deem the alignment correct if the source and match are of the same species, and in case of hold-out analysis where the species is missing from the reference by construction, same genus.

When simulating whole samples, we obtain 338 simulated samples from a holdout set of 371 taxa, because we skip 33 cases the `wgsim` considers the sequences too fragmented to source reads at a set coverage, and errors out. We calculate the number of reads to source per marker to obtain 0.1 coverage as in (Sims et al. 2014).

To run EukDetect, we edit the default config file such that it lists the simulated samples. To run "EukDetect (MAPQ $\geq$ 5)", we additionally modify the source code of our local installation. To run "4 reads + 2 markers (MAPQ $\geq$ 30)", we run CORALE configured to use these three filters instead of the default procedure described in this publication.

Data availability

All our software is publicly available under the MIT license: CORALE (*github.com/wbazant/CORALE*), 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 CORALE and EukDetect on DIABIMMUNE study

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

"Aspergillus Fumigatus and Aspergillosis." 1999. *Clinical Microbiology Reviews* 12 (2): 310–50. https://doi.org/10.1128/cmr.12.2.310.

Beghini, Francesco, Edoardo Pasolli, Tin Duy Truong, Lorenza Putignani, Simone M Cacciò, and Nicola Segata. 2017. "Large-Scale Comparative Metagenomics of Blastocystis, a Common Member of the Human Gut Microbiome." *The ISME Journal* 11 (12): 2848–63.

Breitwieser, Florian P, DN Baker, and Steven L Salzberg. 2018. "KrakenUniq: Confident and Fast Metagenomics Classification Using Unique k-Mer Counts." *Genome Biology* 19 (1): 1–10.

Doron, Itai, Irina Leonardi, Xin V. Li, William D. Fiers, Alexa Semon, Meghan Bialt-DeCelie, Mélanie Migaud, et al. 2021. "Human Gut Mycobiota Tune Immunity via CARD9-Dependent Induction of Anti-Fungal IgG Antibodies." *Cell* 184 (4): 1017–1031.e14. https://doi.org/10.1016/j.cell.2021.01.016.

Doron, Itai, Marissa Mesko, Xin V Li, Takato Kusakabe, Irina Leonardi, Dustin G Shaw, William D Fiers, et al. 2021. "Mycobiota-Induced IgA Antibodies Regulate Fungal Commensalism in the Gut and Are Dysregulated in Crohn's Disease." *Nature Microbiology* 6 (12): 1493–1504.

Gasparrini, Andrew J., Bin Wang, Xiaoqing Sun, Elizabeth A. Kennedy, Ariel Hernandez-Leyva, I. Malick Ndao, Phillip I. Tarr, Barbara B. Warner, and Gautam Dantas. 2019. "Persistent Metagenomic Signatures of Early-Life Hospitalization and Antibiotic Treatment in the Infant Gut Microbiota and Resistome." *Nature Microbiology* 4 (12): 2285–97. https://doi.org/10.1038/s41564-019-0550-2.

Gibson, Molly K., Bin Wang, Sara Ahmadi, Carey-Ann D. Burnham, Phillip I. Tarr, Barbara B. Warner, and Gautam Dantas. 2016. "Developmental Dynamics of the Preterm Infant Gut Microbiota and Antibiotic Resistome." *Nature Microbiology* 1 (4). https://doi.org/10.1038/nmicrobiol.2016.24.

Hayden, Hillary S., Alexander Eng, Christopher E. Pope, Mitchell J. Brittnacher, Anh T. Vo, Eli J. Weiss, Kyle R. Hager, et al. 2020. "Fecal Dysbiosis in Infants with Cystic Fibrosis Is Associated with Early Linear Growth Failure." *Nature Medicine* 26 (2): 215–21. https://doi.org/10.1038/s41591-019-0714-x.

HMP. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486 (7402): 207.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data." *Molecular Biology and Evolution* 33 (6): 1635–38.

Jiang, Tony T., Tzu-Yu Shao, W. X. Gladys Ang, Jeremy M. Kinder, Lucien H. Turner, Giang Pham, Jordan Whitt, Theresa Alenghat, and Sing Sing Way. 2017. "Commensal Fungi Recapitulate the Protective Benefits of Intestinal Bacteria." *Cell Host & Microbe* 22 (6): 809–816.e4. https://doi.org/10.1016/j.chom.2017.10.013.

Kostic, Aleksandar D., Dirk Gevers, Heli Siljander, Tommi Vatanen, Tuulia Hyötyläinen, Anu-Maaria Hämäläinen, Aleksandr Peet, et al. 2015. "The Dynamics of the Human Infant Gut Microbiome in Development and in Progression Toward Type 1 Diabetes." *Cell Host & Microbe* 17 (2): 260–73. https://doi.org/10.1016/j.chom.2015.01.001.

Kriventseva, Evgenia V, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A Simão, and Evgeny M Zdobnov. 2019. "OrthoDB V10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs." *Nucleic Acids Research* 47 (D1): D807–11.

Laforest-Lapointe, Isabelle, and Marie-Claire Arrieta. 2018. "Microbial Eukaryotes: A Missing Link in Gut Microbiome Studies." *MSystems* 3 (2): e00201–17.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Leonardi, Irina, Iris H. Gao, Woan-Yu Lin, Megan Allen, Xin V. Li, William D. Fiers, Meghan Bialt De Celie, et al. 2022. "Mucosal Fungi Promote Gut Barrier Function and Social Behavior via Type 17 Immunity." *Cell*. https://doi.org/10.1016/j.cell.2022.01.017.

Li, Heng. 2011. "Wgsim-Read Simulator for Next Generation Sequencing." *Github Repository*.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Lind, Abigail L, and Katherine S Pollard. 2021. "Accurate and Sensitive Detection of Microbial Eukaryotes from Whole Metagenome Shotgun Sequencing." *Microbiome* 9 (1): 1–18.

Mora, Camilo, Derek P. Tittensor, Sina Adl, Alastair G. B. Simpson, and Boris Worm. 2011. "How Many Species Are There on Earth and in the Ocean?" Edited by Georgina M. Mace. *PLoS Biology* 9 (8): e1001127. https://doi.org/10.1371/journal.pbio.1001127.

Nash, Andrea K, Thomas A Auchtung, Matthew C Wong, Daniel P Smith, Jonathan R Gesell, Matthew C Ross, Christopher J Stewart, et al. 2017. "The Gut Mycobiome of the Human Microbiome Project Healthy Cohort." *Microbiome* 5 (1): 1–13.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.09.483664; this version posted March 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Oliveira, Francislon S, John Brestelli, Shon Cade, Jie Zheng, John Iodice, Steve Fischer, Cristina Aurrecoechea, et al. 2018. "MicrobiomeDB: A Systems Biology Platform for Integrating, Mining and Analyzing Microbiome Experiments." *Nucleic Acids Research* 46 (D1): D684–91.

Olm, Matthew R., Nicholas Bhattacharya, Alexander Crits-Christoph, Brian A. Firek, Robyn Baker, Yun S. Song, Michael J. Morowitz, and Jillian F. Banfield. 2019. "Necrotizing Enterocolitis Is Preceded by Increased Gut Bacterial Replication, Klebsiella , and Fimbriae-Encoding Bacteria." *Science Advances* 5 (12). https://doi.org/10.1126/sciadv.aax5727.

Ost, Kyla S., Teresa R. O'Meara, W. Zac Stephens, Tyson Chiaro, Haoyang Zhou, Jourdan Penman, Rickesha Bell, et al. 2021. "Adaptive Immunity Induces Mutualism Between Commensal Eukaryotes." *Nature* 596 (7870): 114–18. https://doi.org/10.1038/s41586-021-03722-w.

R Marcelino, Vanessa, Edward C Holmes, and Tania C Sorrell. 2020. "The Use of Taxon-Specific Reference Databases Compromises Metagenomic Classification." *BMC Genomics* 21 (1): 1–5.

Schoch, Conrad L. et al. 2012. "Nuclear Ribosomal Internal Transcribed Spacer (ITS) Region as a Universal DNA Barcode Marker for fungi." *Proceedings of the National Academy of Sciences* 109 (16): 6241–46. https://doi.org/10.1073/pnas.1117018109.

Sims, David, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. 2014. "Sequencing Depth and Coverage: Key Considerations in Genomic Analyses." *Nature Reviews Genetics* 15 (2): 121–32.

Smith, John Maynard. 1970. "Natural Selection and the Concept of a Protein Space." *Nature* 225 (5232): 563–64.

Thankaswamy-Kosalai, Subazini, Partho Sen, and Intawat Nookaew. 2017. "Evaluation and Assessment of Read-Mapping by Multiple Next-Generation Sequencing Aligners Based on Genome-Wide Characteristics." *Genomics* 109 (3-4): 186–91.

Vatanen, Tommi, Aleksandar D Kostic, Eva d'Hennezel, Heli Siljander, Eric A Franzosa, Moran Yassour, Raivo Kolde, et al. 2016. "Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans." *Cell* 165 (4): 842–53.

Wibbelt, Gudrun, Andreas Kurth, David Hellmann, Manfred Weishaar, Alex Barlow, Michael Veith, Julia Prüger, et al. 2010. "White-Nose Syndrome Fungus (Geomyces Destructans) in Bats, Europe." *Emerging Infectious Diseases* 16 (8): 1237–43. https://doi.org/10.3201/eid1608.100002.

Wilson, Richard A., and Nicholas J. Talbot. 2009. "Under Pressure: Investigating the Biology of Plant Infection by Magnaporthe Oryzae." *Nature Reviews Microbiology* 7 (3): 185–95. https://doi.org/10.1038/nrmicro2032.