**Multi-cohort fecal metagenomic analysis reveals the altered fungal signatures in colorectal cancer and the carcinogenic potential of *Aspergillus rambellii***

**Yufeng LIN**, study design, dry lab analysis, wet lab validation, and written.

**Wet Lab:** Yali Liu, Jason Kang, Thomas Kwong.

**Dry Lab:** Yiwei Wang (SBS, CUHK), Han Jing (Med-X, XJTU), Changan Liu, Weixin Liu, SK, Yanqiang Ding.

**Manuscript revision:** Nick Ding, Thomas Kwong, Harry Lau.

Bisi, Sunny Wong, Jun Yu

Introduction

Colorectal cancer (CRC) is the third most common cancer and has the second-highest mortality rate after lung cancer globally1,2. It is estimated that CRC incidence will increase by approximately 80% to over two million cases in the next two decades3. Sporadic CRC, which arises without known contribution from germline causes or significant family history, accounted for about 75% of CRC, implying the importance of environmental factors in CRC pathogenesis4.

The gut microbiome provides numerous essential metabolic and physiological functions for our bodies, including digestion, vitamins synthesis, immune system development and more2. Recent studies have linked gut microbiota alteration to CRC occurence5. Dysbiosis such as reduced gut microbial diversity and enrichment of oncogenic microorganisms have been associated with CRC carcinogenesis. For instance, the notorious *Fusobacterium nucleatum* promoted glycolysis and oncogenesis of CRC by targeting IncRNA ENO1-IT15. Meta-analysis with approximately 1,000 individuals from five cohorts has revealed the microbial signatures of genes specific for CRC6 and the association between the gut microbiome and choline degradation7.

Even though>90% of the gut microbiome are composed of bacteria, a perturbed gut fungal composition had also been described to be associated with inflammatory bowel disease8 and liver cirrhosis9. Fungi could influence the immunological responses of the host by dampening or promoting local inflammatory reactions 8–12. For instance, the mammalian intestinal fungal community interacts with the immune system through the innate immune receptor Dectin-111. The commensal fungi were shown to protect from colitis-associated colon cancer by prompting inflammasome activation and IL-18 maturation in murine model13. Therefore, it is apparent that fungus plays a more significant role in CRC development than previously anticipated. However, except for our previous study14 aiming to discover potential fungal biomarkers for CRC detection, the exact role of fungi in CRC pathogenesis remains unexplored mainly due to their relatively low abundance and lack of well-characterized reference fungal genomes.

In this study, we performed a meta-analysis of eight available datasets. After rigorous and consistent data processing, a total of 1,329 samples from four continents, including 525 healthy individuals, 350 adenoma patients, and 454 CRC patients, were included in this analysis. Specific fungal diversity and features associated with different stages of CRC were identified. We also explored the intra-fungi and fungi-bacteria co-occurrence patterns in CRC and compared CRC correlations, adenoma, and healthy control.

Methodology

Study inclusion and data attainment

Faecal shotgun metagenomic data on CRC-related research with a minimum of 2 subject categories (CRC patients and healthy controls) were retrieved from the NCBI database and from 2014 to 2020. Eight published fecal shotgun metagenomics cohorts and our recently completed cohort (unpublished) were included in this meta-analysis. We downloaded seven public faecal shotgun CRC datasets from European Nucleotide Archive (ENA) using the following ENA identifiers: ERP005534 for Zeller et al.15, ERP008729 for Feng et al.16, PRJEB10878 for Yu et al.17, PRJEB12449 for Vogtmann et al.18, PRJNA389927 for Hanningan et al.19, PRJEB27928 for Wirbel et al.6, and SRP136711 for Thomas et al.7. And the eighth cohort was downloaded from the DNA Data Bank of Japan (DDBJ) with the Accession numbers: DRA006684 and DRA008156 for Yachida et al.20. All published datasets contained at least two groups, CRC patients and healthy individuals; five published encompassed the adenoma patients7,15,16,19,20 (table 1 and supplementary table 1). Our cohort was generated with the new fecal metagenomic data from samples collected in Hong Kong from 2009 to 2012. A subset of samples from this patient collective was published previously14. These nine studies were organized from eight countries and various sampling procedures, sample storage, and DNA extraction protocols.

Hong Kong study recruitment and sequencing

Recruitment criteria included presentations of digestive symptoms to the outpatient gastroenterology clinics and asymptomatic individuals 50 years or older receiving colonoscopy screening from the Chinese University of Hong Kong Jockey Club Bowel Cancer Education Centre. Stool samples were collected by participants and stored at –20°C within 4 hours. For long-term storage, all samples were stored at –80°C within 24 hours of stool collection. Total DNA was extracted from stool samples by using the QIAamp DNA Stool Mini Kit according to the manufacturer's instructions (Qiagen, Germany). All subjects had intact colonic at the time of stool collection. An independent Chinese cohort of 112 control subjects, 111 patients with CRC, and 197 patients with colorectal adenomas was recruited. Part of the samples had been published in the previous research21.

Sample filter criteria

To ensure consistent and high-quality data, samples were subjected to filtering before analysis. Abnormal conditions, history surgery patients, and ambiguous stage patients were discarded. We only included the PCR-free cohort because the PCR-free kits could reduce bias and cell spike-in controls for more accurate quantification22. Last, we excluded the low-alignment reads samples (less than 1,000,000), which might cause by the low sequencing depth, the host reads contamination, etc. In the second part, we tended to remove the outlier or suspected contaminated cases, such as the high-fungi (the fungi contained more than 1% in all intestinal microbiome), low-Fungi (the fungi covered less than 0.01% in the gut microbiota), and bacterial or fungal contamination (one species more than 50% of the intestinal microbiome) samples. For the last division, the low-fungal sequence depth sample (fungi aligned read counts less than 10,000) would be dropped, which was consistent with a previous study exploring that at least 30% of individuals couldn't be detected fungi in all gastrointestinal segments23.

Sequence pre-processing and taxonomic and functional profiling

We applied the KneadData's default parameters to quality control all the metagenomic samples, which aims to perform principled in silico separation of bacterial reads from these "contaminant" reads, be they from the host or other user-defined sources. In the second step, taxonomic profiles were generated with the Kraken2 v2.0.9-beta across the custom database. Our custom library contained 9,543 bacterial and 909 fungi references from NCBI (https://www.ncbi.nlm.nih.gov/), FungiDB (https://fungidb.org/fungidb/), Ensemble (http://fungi.ensembl.org/index.html), and Broad Institute (https://www.broadinstitute.org/); and was established with the Jellyfish program by counting distinct 31-mer. We discarded all reads of quality less than 20 and shorter than 50 nucleotides, and the other parameters were default. Each query was classified to a taxon with the highest total hits of k-mer matched by pruning the general taxonomic trees affiliated with mapped genomes. The final metagenomic read counts were normalized by multiple methods, rarefied abundance, relative abundance (see supplementary table 2 and supplementary table 8), and median normalized(see supplementary table 3 and supplementary table 9) with the script (https://github.com/ifanlyn95/multi-CRC-fungi). To prevent the denominator from being zero, all zero values will be replaced by the normal distribution with a mean value of one-tenth of the non-zero minimum value and one-hundredth of the non-zero minimum value of the variance. The median normalized means dividing the median of the control group of each feature in various cohorts as :

: means the relative abundance of fungi or bacteria in sample , which belongs to cohort . In contrast, cohort has exactly sample to sample .

Feature selections criteria

We had three criteria to select the potential candidates, whether it is bacteria or fungi. First, we excluded the rarefied candidates with an average abundance of less than 0.1% in all fungi. We selected the same trend features (SSTF), required more than 3/4 cohorts (not less than six cohorts) to perform the same trends. And the log2 of Multiple Median Fold Change (log2MultMedFC) was the evaluation index of SSTF. We define the log2MultMedFC as:

means the counts of CRC/CTRL samples in an individual cohort.

means the fungal names.

means the relative abundance of species in sample .

The second measure was based on the Wilcoxon test. The significance of differential abundance was tested on a per species basis using a Wilcoxon test and adjust the p-value with the conservative Bonferroni correction. And the last strict criterion was Fold Change. We only focused on the absolute value of log2 of features' Fold Change more significant than 0.5. In addition, we ignored the unclassified strain of bacteria because we could not explain it. The scripts were also shared on Github (https://github.com/ifanlyn95/multi-CRC-fungi).

The random forest based machine learning approach

Our machine learning analyses exploited the taxonomic species-level median normalized relative abundance by Kraken2 and its plugin; Bracken was an import. To get test error estimates with lower bias, the LOSO (leave one set out) was used to do nested cross-validation. The features selection and model training was calculated with randomForest package in R. To choose the best model, and we utilized the max average AUC and best AUC in multi-features and single-feature as the selected criteria, respectively. Only species appearing in the top three ranking features in at least one cohort were included in multi-features model characters selection. The code generating the analyses and the figures is available at https://github.com/ifanlyn95/multi-CRC-fungi.

Association calculation and comparison

Co-occurrence and co-exclusion relationships within fungi and between fungi and bacteria were estimated using the DGCA algorithm24, the methodology for systematically assessing the difference in feature-feature regulatory relationships under different conditions. P-values less than 0.05 were considered significant. The inclusion criterion for network plot features had a correlation index less than -0.2 or more effective than 0.5.

In the comparison of different stages, DGCA leverages the permutation samples to calculate empirical p-values. Another important index is the z-score, which represents the relative strength of differential correlation. We considered the empirical p-values less than 0.05, and the absolute values of the z-score larger than 5 were a significantly different correlation between different stages. The positive z-score refers to the correlation in CRC is weaker relative strength than the healthy control, while the negative z-score means the correlation is more positive in CRC.

Based upon a threshold for correlation significance (p-value less than 0.05) and the sign of correlation in each condition (i.e., healthy control or CRCs), species-species correlations in each condition could be categorized into three classes, i.e. significant positive correlation, no significant correlation, and significant negative correlation. Therefore, there were nine classes for differential correlation between two stages, namely '+/+', '+/0', '+/-', '0/+', '0/0', '0/-', '-/+', '-/0', and '-/-'. It revealed the altered trends between the two stages.

Additional validation experiments on cancer cell line

TBA

Results

Filtering and pre-processing of a large population from various regions for the fungal meta-analysis

We collected shotgun metagenomic sequencing data from eight cohorts: xxx. All the raw sequencing data were reprocessed using the KneadData, Kraken225, and Bracken26 for taxonomic profiling. Each sample has about 107.19 (median) high-quality paired reads that match the bacterial database, and 104.31 (median) paired sequences were aligned to the fungal genome (figure 1a). The median ratio of fungi to bacteria was 10-2.80 (figure 1a), consistent with a previous study27 which reported that fungi make up about 0.1% of the total enteric microbes. The rarefaction curve (figure 1b) showed that all cohort samples reached a plateau at 10,000 sequencing reads. Hence, the minimum rarefied fungal counts of each individual were defined as 10,000 in the downstream analysis. We applied strict sample filtering criteria to ensure rigorous outcomes and reduce the outlier effect (figure 1c). Collectively, we discarded 216 samples with unsatisfied sequencing quality, 211 outlier or contaminated samples. Additionally, 296 samples with low-fungi sequencing depth were filtered. It is consistent with a previous study23 that fungi could be detected in approximately 70% of individuals in the gut.

Alterations of enteric fungal and bacterial composition in CRC

Considering the overall fungal composition, *Ascomycota* was the most abundant fungal phylum among all the cohorts, while other dominating fungal phyla showed significant inter-cohort variations. For instance, unlike all other cohorts, the second-most abundant phylum in Yachida's Japanese cohort was *Mucoromycota* but not *Basidiomycota*. Other examples include *Microsporidia* taking up a smaller proportion in Asians than non-Asians (figure 2b).

For the altered microbial composition in CRC, we observed that bacterial phyla *Bacteroidetes* and *Fusobacteria* were enriched in the CRC group compared with the healthy control group. Conversely, *Firmicutes* and *Actinobacteria* were reduced in the CRC group (supplementary figure 1a). When we investigated the individual cohort, 6 of the total cohorts showed significant enrichment of *Fusobacteria* (p-value < 0.05) in the CRC group (supplementary figure 1b). Although no fungi were identified to show a stronger relationship with CRC than Fusobacteria, differentially abundant fungi were identified, which will be discussed in later sessions (figure 2c and supplementary figure 1).

In agreement with previous research showing distorted microbiome diversity in the diseased group28, we observed that alpha diversity was reduced in CRC patients compared to healthy individuals when considering all the cohorts together (figure 2d). When considering individual cohorts, the majority showed reduced alpha diversity (chao1 index). Despite the less apparent alteration of fungal composition compared to the bacterial composition in CRC and the heterogeneity in different cohorts, we could still observe remarkable differences in both fungi phyla composition and alpha diversity in the CRC group, which are not negligible.

Identification of fungal species associated with CRC by univariate meta-analysis

We next searched for the potential enteric fungal shifts in CRC patients as compared to healthy individuals. After filtering low abundant (< 0.1%) fungi from the 592 aligned species, 296 species were obtained for further analysis (figure 3a and supplementary table 2, 3). Using the Wilcoxon rank-sum test to compare data from all the cohorts together, 74 differentially abundant fungi were identified, which was named as the main set (FDR < 0.1). Among the 74 identified species, we further shortlisted 33 species that demonstrated significant alterations (FDR < 0.01) as the core set (figure 3a and supplementary table 4).

We then evaluated if these 74 fungi (main set) were consistently altered across all the eight cohorts using SSTF and non-parametric tests (see Methods). We observed that the enrichment and depletion status of the 74 species were consistent in most of the cohorts except the 2019\_Thomas and 2019\_Yachida cohorts. Interestingly, most of the 74 species in the 2019\_Thomas cohorts either showed significant enrichment in CRC patients or no significant difference between CRC versus healthy individuals. Still, very few showed deletion in CRC patients. Whereas in the 2019\_Yachida group, most of the identified 74 fungi showed weak variance in CRC patients versus healthy individuals, unlike other cohorts. We also discovered that 3 of the 74 species showed consistent changes across all the cohorts with *Aspergillus* *rambellii* and *Erysiphe* *pulchra* being enriched while *Trichophyton* *mentagrophytes* being depleted in CRC (figure 3b and supplementary table 6). We further identified 15 species that were consistently altered in 7 out of the eight cohorts. Ten of them were enriched in CRC patients, while the remaining five were depleted (see supplementary table 6). Notably, only *Aspergillus* *rambellii* showed a significant difference (p-value < 0.05) in all the cohorts, except the 2019\_ThomasA cohort (figure 3d and supplementary table 7).

For the 33 species in the core set, ten were enriched in CRC patients, and the remaining 23 were depleted (figure 3c). Notably, we identified that *Aspergillus* *rambellii* showed the most remarkable difference between the CRC patients and the healthy control groups (-log10FDR = 17.29). We observed that alterations of these 33 species in CRC patients versus healthy individuals were relatively consistent in most cohorts except 2019\_Thomas and 2019\_Yachida. Moreover, most of the species showed consistent alterations in at least three cohorts.

In the meantime, we also compared the fungal community between CRC and adenoma patients (supplementary table 5 and supplementary figure 2). Seven of the identified species were significantly differed (FDR < 0.01) in both CRC patients versus adenoma patients and CRC patients versus healthy individuals. These species include *Aspergillus rambellii*, *Moniliophthora perniciosa*, *Erysiphe pulchra*, *Sphaerulina musiva*, *Phytophthora capsici*, *Aspergillus kawachii*, and *Cordyceps sp. RAO-2017* . These species belong to the *Ascomycota* phylum except *Moniliophthora* *perniciosa* and *Phytophthora capsici* (supplementary table 12).

In summary, we identified 74 differentially abundant fungi between CRC patients and healthy individuals, of which 33 significant species were further selected as core-set for the downstream analysis. Despite cohort heterogeneity, we could still identify fungal features that were consistently altered in most of the cohorts.

*A. rambellii* is the most significant fungus associated with CRC

To identify the most crucial candidates associated with CRC, we utilized the stricter criteria (see methods). We found that *A.* *rambellii* and *A.* *kawachii* were the only two significant fungi (figure 3c). *A. rambellii* was the only candidate with a significant difference among seven cohorts, excluding the 2019\_Thomas cohort (figure 3d). In contrast, *A. kawachii* was significantly different among 2014\_ZellerG, 2016\_VogtmannE, 2017\_JunY, and our unpublished dataset (figure 3d). Although they belong to the same genus, they were shown to have a different association with CRC, with *A. rambellii* being enriched in CRC, while *A. kawachii* being the opposite. Notably, they were also reported to have opposing actions in previous studies. *A. rambellii* has been demonstrated to accumulate aflatoxins (AF) and the aflatoxin precursor sterigmatocystin (ST)29. AF and ST are well known as the most carcinogenic natural products 30. In contrast, *A. kawachii* was reported to enhance anticancer effects of cancer-curing herbs, such as Korean mistletoe31 and fermented silkworm larvae32. All these previous pieces of literature supported our findings. Collectively, our meta-analysis revealed the essential fungi, *A.* *rambellii* and *A. kawachii*, that were significantly associated with CRC across multiple metagenomic studies.

The fungal-bacterial combination model performed best in CRC diagnosis

To identify the significant differentially abundant bacteria between CRC and healthy individuals, we performed Wilcoxon rank-sum test with stringent selection criteria (q-value < 0.01, , unclassified species removed) (supplementary table 10). Thirty-one differentially abundant bacteria were identified in CRC, which was more significant than fungi (supplementary table 10). At least half of the identified bacteria have been reported to be CRC-related, such as Fusobacterium nucleatum, Parvimonas micra, and Gemella morbillorum, or potential probiotics including *Roseburia* *intestinalis*, *Bifidobacterium* *bifidum*33–44, and *Streptococcus* *thermophilus*.45–50. The consistent results in differentially abundant bacteria analysis implied that our methods used in the previous differentially abundant fungi analysis were credible.

In previous researches6,7, the CRC discriminator was trained with the bacteria only. Further, we asked whether the fungal could improve the diagnosis ability of the classifier and identify the potential value of gut fungi in clinical; we trained the model with single or multiple features to diagnose CRC with the fungal and bacterial core set. Among the single-feature models, the average AUC value of six features was greater than 60%. They were composed of four bacteria (*Fusobacterium nucleatum*, *Parvimonas micra*, *Gemella morbillorum*, and *Porphyromonas asaccharolytica*) and two fungi (*Aspergillus rambellii* and *Aspergillus kawachii*) (table 2). Among these, *P. micra* showed the best operation, with an average AUC value of 67.79%, but it played the bad performance in 2016\_VogtmannE (AUC: 56.15%), in which A. rambellii performed the best achievement (AUC: 67.57%). It revealed that the fungi could supplement the bacteria in some situations. And then, we wanted to know whether the classifier model would be improved when utilizing both fungi and bacteria as the candidates, so we trained and compared the multi-features model with pure fungi, bacteria, and the fungal-bacterial combination, with 17, 12, and 14 characters, respectively (figure 4a, supplementary figure k1, and supplementary figure k2). Apart from 2015\_FengQ, the combination classifier improved 1.44% - 10.60% compared with the bacterial classifier (figure 4). Unexpectedly, the fungal classifier played more efficient than bacterial one in 2016\_VogtmannE (fungi: 77.27% vs bacteria: 70.63%) and 2019\_WirbelJ (fungi: 93.23% vs bacterial 89.39%). In general, the fungal-bacterial combination classifier was more sensitive compared with the pure fungal or bacterial model.

Correlation between CRC-associated bacteria and CRC-associated fungi

Due to the complex and multifactorial nature of CRC, we asked whether the interactions among fungi were associated with CRC. We performed the correlation analysis with DGCA24 on the 33 differentially abundant fungi in the core set. We observed that the correlations within the core-set fungi were stronger in CRC than healthy control (figure 4 and supplementary figure 4). There were only four strong positive (correlation index ≥ 0.5) and three negative interactions (correlation index ≤ -0.15) in healthy individuals (figure 4a) as compared to the nine strong positive and four negative associations exhibited in CRC patients (figure 4b). Additionally, there were nine strong positives and one negative interaction in adenoma patients.

*Aspergillus* *rambellii*, *Rhizophagus* *irregularis*, *Rhizophagus* *clarus*, *Phytopythium* *vexans*, and *Edhazardia* *aedis* showed significant correlations in all three conditions (Healthy, Adenoma, CRC) (figure 4 and supplementary figure 3). However, no characteristic pattern was identified in the changes of their correlations across different conditions. A more exciting discovery was that 4 of the CRC-associated fungi, *Aspergillus* *rambellii*, *Erysiphe* *pulchra*, *Thielaviopsis* *punctulata*, and *Sphaerulina* *musiva,* were found to be positively correlated with one another in CRC patients. However, these correlations disappeared in the adenoma or healthy individuals (figure 4a and supplementary figure 3). In these two conditions, only *A. rambellii* still showed correlations with other fungi but not *T. punctulata*, *S. musiva*, and *E. pulchra*. Interestingly, CRC-depleted fungi such as *R. clarus*, *E. aedis*, *Naumovozyma dairenensis* and *R. irregularis* are correlated with one another in healthy individuals. Altogether, our results suggested that different sets of fungi were strongly correlated in different conditions. While CRC-depleted fungi might be crucial for maintaining enteric homeostasis in a healthy gut, enrichment of CRC-associated fungi might break the homeostasis contributing to CRC carcinogenesis.

Next, we asked whether the correlations between the differentially abundant fungi and bacteria are associated with CRC. DGCA24 was performed to calculate the correlation between fungi and bacteria. We discovered that the fungal-bacterial correlations were weaker than inter-fungal correlations across all three conditions (Healthy, Adenoma, CRC) (supplementary figure 5 and supplementary table 11). However, the fungal-bacterial correlation was still more vigorous in CRC when compared to healthy individuals, which showed the same pattern as in inter-fungal correlations (figure 4, supplementary figure 5 and supplementary table 11). This suggested that although the fungal-bacterial interactions are weaker than inter-fungal interactions, they might still be associated with CRC tumorigenesis.

Alternative enteric microbiome in CRC and fungi were the important potential biomarkers or inducements for CRC

After determining the inter-bacteria, inter-fungal and fungal-bacterial correlations in CRC and healthy conditions, we next asked whether these correlations are significantly different between these two conditions. DGCA was used to perform the differential correlation analysis. Interestingly, inter-bacterial correlations were stronger in CRC patients than in healthy individuals, while inter-fungal correlations were stronger in healthy individuals. (figure 5a). When assessing fungal-bacterial correlations, two peaks at -5 and +5 were observed in the density graph with Z-score, indicating the strength of fungal-bacterial correlations do not show simple unidirectional changes across two conditions. While a group of fungal-bacterial interactions became stronger in CRC patients, another group of fungal-bacterial interactions became weaker. Collectively, our differential correlation analysis demonstrated distinct differences in the correlation changes among inter-fungal, inter-bacterial and fungal-bacterial interactions.

We also defined the nine cases in the pair correlation comparison (figure 5b left panel and see methods). Our results showed that the most significant correlations were '+/+', '+/0', and '0/+', indicating that the most meaningful comparisons (*pm Val* < 0.05) were based on the positive correlations; in other words, negative correlations were rare (figure 5c). Notably, only the intra-fungi had six '-/+' cases, which means the feature pair correlation in CRC was negative, while its association in healthy control was positive (figure 5c). It might reveal some potential markers or changes in the stage alteration.

Sixty-four microbes (31 bacteria and 33 fungi) were separated into six clusters with affinity propagation clusters (figure 5d). Among these, two clusters contained most of the candidates identified. We named the biggest one the Fun\_cluster because 18 of 22 microbes were fungi. In this cluster, the correlations between the enriched fungi were enhanced in CRC compared. We identified that *E. pulchra* and *A. rambellii* were the core fungi in the eEuk\_cluster. Three CRC enriched bacteria, including *F. nucleatum*, *F. periodonticum*, and *P. micra,* had strong correlations with these fungi (figure 5d). The most outstanding fungi, *A. rambellii*, and the cancer-related pathogens, *F. nucleatum*, were in the same cluster. Its z-score was -5.95, and it belonged to the '+/+' case (see supplementary table 11), indicating that fungi *A. rambellii* and bacteria *F. nucleatum* were both positive relative CRC and healthy control. Still, their pair association was more potent in CRC compared with healthy control. Also, *A. rambellii* showed a strong correlation with another CRC-enriched bacteria *P. micra*, with a z-score -5.07, belonged' +/0' (see supplementary table 11). In contrast, there was no direct correlation between another critical fungus, E. pulchra, and these three CRC-enriched bacteria were identified. It revealed that *A. rambellii* was an essential potential carcinogen and related to other reported CRC-pathogenes.

The second clustering included twenty-one candidates, and 17 of 21 were bacteria, namely Bac\_cluster. It had the most reported probiotics or potential probiotics, including *S. thermophilus*, *S. salivarius*, *A. hadrus*, and *E. eligens* (figure 5d). While some cancer enriched such as *D. pneumosintes*, *S. anginosus*, and *P. intermedia* were also included in this clustering (figure 5d). And these results were consistent with many previous studies reporting the correlations among bacteria. Interestingly, *P. kudriavzevii* was the only one correlated with three of four probiotics. In previous research51, colon cancer cell proliferation, induction of endogenous and exogenous apoptosis inhibition would be inhibited by metabolites from *P. kudriavzevii*. Collectively, these findings revealed that the correlations among intra-fungi, fungi-bacteria and bacteria-bacteria were significantly different in CRC. The specific relationships between *A. rambellii* with CRC-carcinogen and *P. kudriavzevii* with CRC-probiotics revealed that fungi were the important potential biomarkers or inducement for CRC.

*Aspergillus rambellii* and its conditioned medium promote the viability of colon cancer cells.

[TBA]

Discussion

Our study performed the first comprehensive multi-cohort meta-analysis of enteric fungi shotgun metagenomics in CRC. Samples from eight publicly available cohorts and one unpublished dataset were assembled in this study. We were able to demonstrate the distinct mycobiota alteration in CRC patients versus healthy individuals. Using robust statistical methods, we also identified differentially abundant fungi present in CRC patients, which could be used together with bacterial markers to improve the accuracy of distinguishing CRC patients from tumor-free healthy individuals based on fecal metagenomes. Our DAGC analysis further inferred the correlations among fungal markers and bacterial markers identified. We further calculated the correlation differences in CRC patients versus healthy controls, which revealed the potential fungi-bacterial interactions associated with CRC pathogenesis. **[TBA]**

To improve credibility and accuracy, we adopted strict criteria to gain 1,329 from 2,052, those ratio consisted of the previous study23,52. Previous research mainly focused on the relationship between gut bacteria and host pathology. A few studies were investigated in microbiota studies with fungi for an extended period because of their low proportion in the enteric microbiome53. According to the alpha diversity rarefaction curve, our research suggested that future studies on enteric fungi should be conducted at a sequencing depth of at least ten thousand fungi in each case. We have supplemented the gaps in sequencing depth for fungi and provided reliable support for follow-up fungal research based on the meta-analysis of more than 2,000 samples.

Cohort heterogeneity of the enteric mycobiota was observed across different studies. From our principal component analysis, there were significant p-values for the fungal compositional differences across cohorts. This is consistent with previous studies showing the crucial roles of genetic background, age, dietary habits, lifestyle, and local environments in microbiota composition across different populations54. Similar to previous research studying gut bacterial composition, we also observed the altered mycobiota composition in CRC versus healthy controls. We found that the chao1 index of the intestinal fungi in CRC was weaker than controls, which was stable with gut bacteria in previous studies55,56.

*Ascomycota* and *Basidiomycota* dominated in both CRC and control subjects in almost all cohorts. The alpha diversity in CRC was decreased compared with healthy controls. It was similar to other intestinal diseases such as inflammatory bowel diseases (IBD)8. *Saccharomyces cerevisiae* and *Candida albicans* were enriched in IBD patients, which may be the primary reason for low alpha diversity in IBD57. However, the present study showed no apparent difference between CRC and healthy control. This revealed that even though the alpha diversity of the various disease would be disturbed, their mechanisms may be different.

The meta-analysis approach has been used to evaluate and combine results from comparable studies58 with significant advantages of reducing the influence by cohort-specific bias and increasing statistical power. Using the rank-sum test and SSTF in our meta-analysis, we identified 33 fungi and 31 bacteria that were associated with CRC across eight cohorts. Our results suggested that *A. rambellii* was the most significant CRC-enriched fungi, which showed universal associations with CRC in seven of eight cohorts. This finding was supported by previous studies showing the ability of *A. rambellii* in synthesizing carcinogenic products, aflatoxin and aflatoxin precursor sterigmatocystin29,30. Interestingly, the most significant CRC-depleted fungi were *A. kawachii*, also from the genus *Aspergillus*. Even though both fungi were from the same genus, they play precisely opposite roles in CRC. The crude enzyme extract derived from *A. kawachii* could enhance the antioxidative activities of Viscum album var. coloratum31 (Korean mistletoe), a promising agent for immunomodulation, treating colon cancer59 and hepatoma. Another study also reported that the fermented silkworm produced by *A. kawachii* solid-state fermentation could inhibit the human hepatocellular carcinoma cells32. *R.* *irregularis* was the second most significant CRC-depleted fungi. Supporting evidence from previous studies showed that *A. Officinalis*-*R. irregularis* symbiosis can lead to the production of salvianolic acid, which has anticancer effects60,61. Furthermore, other beneficial effects of *A. Officinalis*-*R. irregularis* symbiosis included the production of rosmarinic acid, ferulic acid and caffeic acid, presenting several health-related properties, such as antioxidant, anti-inflammatory, and antimicrobial effects62,63. These observations support their roles in the promotion or inhibition of colorectal carcinogenesis. Since only very few studies have been done to investigate the fungi associated with CRC, we have to ensure the statistical methods we used to identify differentially abundant fungi in CRC were robust and accurate. Therefore, the same statistical methods were used to identify the CRC-associated bacteria from the metagenomics sequencing data. Our results were supported by more than half of the CRC-related bacteria we identified were reported previously to be cancer-related or commonly used probiotics. Four out of the top five were well-known CRC-associated microbes, namely, *F. nucleatum*64, *P. micra*17, *G. morbillorum*65,66, and *A. hadrus*67. This proved that the statistical methods used in our study are reliable and that the CRC-associated fungi we identified might be potential pathogenic fungi or probiotics.

We developed fungal-bacterial machine learning models to distinguish CRCs from healthy control with an average performance of 0.83 AUC when validated on the cohort excluded from the model's training, while AUC of pure bacterial and fungi classifiers were 0.81 and 0.73, respectively. The combined classifier improved 1.44% - 10.60% compared with the traditional bacterial classifier in seven out of eight cohorts. It revealed that the fungi also contributed to CRC diagnosis, and fungi might also play essential roles in colorectal carcinogenesis as bacteria. The performance of fungal classifiers in 2016\_VogtmannE and 2019\_WirbelJ was better than the bacterial one disclosed that bacteria wasn't the critical CRC-related factor in all situations.

The fungal internal relationship in CRC was significantly different from the healthy control and adenoma. Most previous studies have focused on one key microorganism or one metabolite effect on CRC development68–70, but the human gut microbiome was a complicated and mutually microbial community. Multiple strong positive or negative correlations in CRC were disappeared or weaken in adenoma or healthy control groups. Except for the transformed fungal abundance, alteration in the intra-fungal relationship and fungi-bacteria association could partially explain colorectal tumorigenesis. The most substantial relationship clustering in these three communities was also significantly different, but *Aspergillus* *rambellii*, *Rhizophagus* *irregularis*, *Rhizophagus* *clarus*, *Phytopythium* *vexans*, and *Edhazardia* *aedis* appeared in all stages. It revealed that they might play a vital role in the stability of the entire intestinal ecology.

We found two clusters, Bac\_Cluster and Fun\_Cluster, contained cross-kingdoms carcinogen and tumor suppressor groups, respectively. Our results showed that the fungal internal correlations were interrupted in CRC, while bacterial were enhance. The disruption from the fungal interrelationship may break the healthy intestinal environment and induce colorectal carcinogenesis. On the other hand, the new increased bacterial correlations in CRC may potentially contribute to colorectal carcinogenesis. We also observed some interesting comparisons of the relationship between fungi and bacteria. Our results showed that the less changed differential fungal-bacterial correlations (|z-score| < 2) contained low proportions. The CRC strengthen and weaken correlations performed the primary and secondary ratios in fungal-bacterial correlation comparisons, respectively. It revealed that the internal-kingdom associations and external-kingdom correlations were significantly different. This suggested that bacterial kingdom dysbiosis may cause the fungi to tremble rapidly, which was not similar to the warm alteration of internal-kingdom relationships as previously described14,71. All reported potential probiotics, excluded *R. intestinalis*, were separated in the Bac\_Cluster.

We disclosed that *P. kudriavzevii* owned multiple correlations with these probiotics. And itssecretion metabolites exert anticancer effects by inhibiting cell proliferation and inducing intrinsic and extrinsic apoptosis in colon cancer cells51. Another interesting finding, there were strong correlations among *A. rambellii*, *F. nucleatum*33,64,72, and *P. micra*17, and the latter two were the famous CRC-related pathogens. The previous study disclosed that the altered trans-kingdom association between bacteria and virome in CRC21. We supposed that the trans-kingdom interactions also were happened between bacteria and fungi, while it suggested the fungal-bacterial synergistic correlations could inhibit or promote colorectal carcinogenesis. But this discovery was explored only in metagenome sequencing, and more experiments are needed to verify and prove it. It may indicate colorectal carcinogenesis under their synergistic effect. Collectively, this discovery exposed that may exist other potential probiotics or carcinogens in these clusters.

TBA (in vitro)

Contribution

**Yufeng Lin**, study design, dry lab analysis, wet lab validation, and written.

**Yali Liu**, wet lab validation.

**Nick Ding**, written.

**Thomas Kwong**, written and web lab technology guidance.

**Yiwei Wang**, dry lab guidance (SBS, CUHK).

**Harry Lau**, written guidance.

**Han Jing**, fungal database establishment (Med-X, XJTU).

**Jason Kang**, web lab technology guidance

**Changan Liu**, dry lab guidance.

**Weixin Liu**, dry lab guidance.

**SK**, dry lab guidance.

**Yanqiang Ding**, dry lab guidance.

**Bisi**, dry lab guidance.

**Sunny Wong**, clinical guidance.

**Jun Yu**