Consistent filtering and pre-processing of a large population from various regions for the microeukaryotic meta-analysis

In this meta-analysis, eight published faecal shotgun metagenomics cohorts and one indoor cohort were included1–8. All published datasets contain at least two stages, CRC patients and healthy individuals; five published encompass the adenoma patients1,2,5,7,8 (table 1 and Supplementary Table 1). Our indoor cohort was generated with the new faecal metagenomic data from samples collected in Hong Kong from 2009 to 2012. Even though a subset of samples from this patient collective was published previously9, we have added complete follow-up clinical information (see Supplementary Table 2 and Methods). These nine studies were organized from eight countries and various sampling procedures, sample storage, and DNA extraction protocols. In the beginning, all raw sequencing data were reprocessed using the KneadData, Kraken210, and Bracken11 for taxonomic profiling (see Methods). 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 mEuk genome (figure 1a). And the median ratio of microeukaryotes to bacteria was 10e-2.80 (figure 1a), which is consistent with previous research12 that revealed that fungi occupy nearly 0.1% of the total enteric microbes. It acknowledges that our custom libraries, alignment, and results were reliable. Through the rarefaction curve (figure 1b), we could know that all cohort samples have reached or exceeded the plateau at 10,000. Hence, the minimum rarefies microeukaryotes counts of each individual were defined as 10,000 in the downstream analysis. We applied the strict criteria to remove a few samples to enhance outcomes rigor further and reduce the outlier effect (figure 1c). Because of the mEuk containing a low proportion, deep enough sequencing and free PCR were compulsory. Notably, one cohort5 whole-metagenomic-library preparation was employed 12 cycles of limited-cycle PCR; moreover, its sequencing size was five to ten times smaller than others. Therefore, we didn’t adopt this cohort. At last, after three main filters (figure 1c and Methods), a total of 1,329 samples (525 healthy control, 350 adenoma patients, and 454 CRC characters) were accepted for downstream analysis.

The enteric microeukaryotic composition was alterations in CRC

Consistent with previous studies and as a validation for our analysis, we observed bacteria phyla Bacteroidetes and Fusobacteria were enriched in the CRC group compared with healthy control. Conversely, Firmicutes and Actinobacteria were reduced (see supplementary figure 1). Among the microeukaryotic taxa, the phylum Ascomycota dominated the microbiota, while Basidiomycota was observed as the second most abundant phylum (figure 2a). It’s worth noting that each cohort would play a few variances in phylum level. For example, the second-largest abundance in Yachida’s cohort from Japan Asia was Mucoromycota instead of *Basidiomycota*. *Microsporidia* *contains* less proportion in Asians compared with non-Asians (figure 2b). In the downstream analysis, we normalized the data through healthy control median in each group and each feature to reduce these effects (see Methods). We also made the phylum comparison between CRC and healthy control. In bacteria phylum level, Fusobacteria performed significantly in 6 cohorts and enriched in all in CRC compared to healthy control (see supplementary figure 2). But none showed a steady trend or difference in each study like Fusobacteria in microeukaryotes phylum level (figure 2c). In agreement with the previous research showed distortion in microbiome diversity in the disease stage13, alpha diversity indices were reduced in patients with CRC compared to control individuals when compared all the samples together (figure 2d). Most cohorts showed diversity reduction by alpha diversity index, chao1. Even though the alteration in microeukaryotes is not as apparent as in bacteria level, it still offered some difference in CRC compared with healthy control. Although there was heterogeneity in different cohorts, overall, both microeukaryotes phyla composition and alpha diversity were significant differences in the CRC group compared with the healthy control.

Seventy-four microeukaryotic species were associated with CRC through univariate meta-analysis

As previously described, the factor ‘cohort’ has a predominant effect on species composition because of the different DNA extraction protocols, various races, etc. An analysis of microbial alpha diversity and beta diversity also revealed that cohort heterogeneity has a more significant effect on overall microbiome composition than CRC in our data (see supplementary figure 3 and supplementary figure 4). We normalized the data through the median for each species in different cohorts and accessed the relative median abundance (see Methods). It could reduce the impact of various studies and enhance the influence of other factors. We filtered the rarefied microeukaryotes (relative abundance < 0.1% of all the microeukaryote) and accessed 296 features (figure 3a and supplementary table 3) from 592 aligned species (see supplementary table 4). To determine potential enteric microeukaryotes shift in patients with CRC, we compared the selected 296 species relative median abundance between healthy control and CRC patients. We gained 74 and 33 candidates whose adjusted p-value is less than, respectively, 0.1 and 0.01 by the Mann-Whitney U test and Bonferroni adjustment (figure 3a). We identified the 74 candidates as the main set and 33 features as the core set. It’s worth remarking that the difference between CRC and healthy control of *Aspergillus* *rambellii* (-log10FDR = 17.29) was much more significant than others. In the meantime, we made the same comparison between adenoma and CRC patients (see supplementary table 4). Only six features also significantly differed (FDR < 0.01) in CRC compared with adenoma, namely *Aspergillus* *rambellii*, *Erysiphe* *pulchra*, *Thielaviopsis* *punctulata*, *Moniliophthora* *perniciosa*, *Sphaerulina* *musiva*, *Aspergillus* *ochraceoroseus*. Except for *Moniliophthora* *perniciosa* belonging to the *Basidiomycota*, the other five belong to the *Ascomycota*, the dominant feature in the phylum level. In summary, thirty-three species were selected as core-set for the downstream analysis from seventy-four significant different species around 296 none-rarefied microeukaryotes.

Even though the main set of 74 microeukaryotes performed a significant difference between healthy control and CRC patients when combining all studies, we also wanted to realize their performance in each cohort. We analyze each study with the 74 candidates through SSTF and non-parameter tests (see Methods). Three of the main CRC-CTRL-associated species consisted of the same absolute trend in the comparison. *Aspergillus* *rambellii* and *Erysiphe* *pulchra* were enriched; simultaneously, *Trichophyton* *mentagrophytes* were decreased in CRC across eight cohorts (figure 3b and supplementary table 5). In addition, all these three belong to the core set. At seven from eight with the same trend, we identified the other five rose, and eleven reduced in CRC (see supplementary table 5). However, only *Aspergillus* *rambellii* was a significant difference (p-value < 0.05) in almost all the cohorts, excluding 2019\_Thomas research (figure 3d and supplementary table 6). Apart from the 2019\_Thomas and 2019\_Yachida, the other six cohorts performed the roughly synchronous trend, especially in the 33-core-set. Among the core set, ten species were enriched in CRC; meanwhile, the reduction was twenty-three (figure 3c). As for the cohort heterogeneous, we observed that the 2019\_Yachida research performance was dissimilar, and it seemed much cleaner. Excluding the *Aspergillus* *rambellii* and few species that have an apparent difference in fold change between CRC and healthy control, most features’ variance was weak and small. One more study, 2019\_Thomas, also behaved outlier performance in another section; most of its features were rich or no difference in CRC compared with the healthy control. Our results found that most cohorts performed the same trend among core-set, and most selections had at least three cohorts perform a significant difference between CRC and healthy control, but still existed cohort heterogeneity among them.

*Aspergillus* *rambellii* and *Aspergillus* *kawachii* showed opposite performance in CRC

We next increased the cutoff value to identify the most crucial candidate associated with CRC by filtering core-set using three strict criteria, FDR < 0.01, SSTF ≥ 6, and abs(log2FC) ≥ 1 (see Methods). After filtering, *Aspergillus* *rambellii* and *Aspergillus* *kawachii* were the only two microeukaryotes that meet these stricter criteria (figure 3c). A. rambellii was the only candidate with a significant difference among seven cohorts, excluding the 2019\_Thomas cohort (figure 3d). And *A. kawachii* was significantly different among 2014\_ZellerG, 2016\_VogtmannE, 2017\_JunY, and our indoor dataset (figure 3d). Although they belong to the same genus, *A. rambellii* was enriched, while *A. kawachii* was less in CRC. In previous research, *A. rambellii* has been acknowledged to accumulate both aflatoxins (AF) and the aflatoxin precursor sterigmatocystin (ST)14. And AF and ST are the most carcinogenic natural products known15. In contrast, *A. kawachii* was usually reported together with some anti-cancer or cancer-curing herbs, such as Korean mistletoe16 and fermented silkworm larvae17. Collectively, our meta-analysis revealed the key microeukaryote, *A.* *rambellii*, and *A. kawachii*, that was significantly correlated with CRC among multiple metagenomic studies.

Alteration in the CRC Microeukaryotic Ecological Association

Due to the complexity and multifactorial nature of CRC, we investigated the potential alternations of polymicrobial ecological interactions in CRC by estimating multiple microeukaryotes and microeukaryotic-bacterial selections correlations. We observed that the correlations within the microeukaryotic core-set network were stronger in CRC than in healthy control (figure 4). There were only four strong positive (correlation index ≥ 0.5) and three negative interactions (correlation index ≤ -0.15) in healthy control (figure 4a); meanwhile, nine high positive and four negative associations exhibit in CRC (figure 4b). In addition, nine close positives and one negative interrelationship were executed in adenoma. *Aspergillus* *rambellii*, *Rhizophagus* *irregularis*, *Rhizophagus* *clarus*, *Phytopythium* *vexans*, and *Edhazardia* *aedis* appeared in all three stages (figure 4a and supplementary figure 5). Still, their correlation was not consistent in various stages. The further exciting discovery was that co-occurrence interactions were observed among microeukaryotes *Aspergillus* *rambellii*, *Eysiphe* *pulchra*, *Thielaviopsis* *punctulata*, and *Sphaerulina* *musiva* in CRC. They were seen as a clustering of microeukaryotes that cooperates and symbiotically (figure 4a). However, these close relationships were disappeared in the adenoma or healthy control group (figure 4a and supplementary figure 5). In adenoma or healthy control, *T. punctulata*, *S. musiva*, and *E. pulchra* didn’t release the strong correspondence with any other candidates, but *A. rambellii* were a high positive connection with *Moniliophthora* *Pemiciosa* in these two stages. Thus, our analysis revealed that co-occurrence microeukaryotic relationships might be crucial for enteric homeostasis in a healthy gut. In contrast, microeukaryotic dysbiosis might break the balance and provide a suitable environment for the harmful microeukaryotes clustering developing, which might cause colorectal carcinogenesis.

Correlation between CRC-related bacteria and selected microeukaryotes was Perturbed in CRC

As we know, our research was the first study about microeukaryotic meta-analysis in CRC, and we want to know whether our methodologies were corrected or not. We tend to utilize these criteria to select the bacterial candidates. Compared with microeukaryotes, the difference between bacteria in healthy controls and CRC is more significant; we gained 31 features through two filters (see Methods). At least half of the bacterial candidates have informed cancer-related18–23,23–29 or the reported probiotics30–35, including some well-known cancer-related bacteria, such as *Fusobacterium* *nucleatum*, *Parvimonas* micra, and *Gemella* *morbillorum*; and some famous probiotics, such as *Roseburia* *intestinalis*, *Bifidobacterium* *bifidum*, and *Streptococcus* *thermophilus*. This result revealed that our discovery was validated the previous analysis selection were credible. And next, we aimed to explore the associations between the microeukaryotes and bacteria. We utilized the same method, DGCA36, as the internal correlation of microeukaryotes. We discovered the relationship of microeukaryote-bacteria was weaker than internal microeukaryotes. However, we explored that the associations in CRC were much more potent than in healthy control, same with internal microeukaryotes correlations (figure 4 and supplementary table 7).

Alternative enteric microbiome in CRC and *A. rambellii* interacted domains with *F. nucleatum* and *P. micra*

Our previous work only compared the distribution, counts, or value of correlation index in different groups. Still, we intended to use a more reliable method, DGCA36, to judge the differences in the correlation of enteric microbiome between CRC and healthy control (figure 4). DGCA identified the z-score to represent the relative strength of differential association (see methods). The positive z-score refers to the correlation in CRC is weaker than the healthy control, while the negative z-score means the correlation is more positive in CRC (figure 5b).

Our results revealed a marked difference in correlations among microeukaryotes and bacteria in CRC and healthy (figure 5a). Correlations among microeukaryotes were higher in healthy controls compared to CRC (figure 5a). In contrast, correlations among bacteria were increased in CRC (figure 5a). When assessing microeukaryotes-bacteria correlations, two peaks at -5 and 5 were observed, indicating the microeukaryotes-bacteria correlations didn’t exist the gentle relationship. Collectively, our results implicated that correlations among microeukaryotes were weakened in CRC, while bacteria-bacteria correlations were utterly opposite. And associations of microeukaryotes-bacteria existed above two situations.

We also defined the nine cases in the pair correlation comparison (figure 5b and see methods). Our results showed that the most significant correlations were ‘+/+’, ‘+/0’, and ‘0/+’, indicating that the most meaningful comparisons (pmVal < 0.05) were based on the positive correlations; in other words, negative correlations were rare (figure 5c). Notably, only the internal microeukaryotes 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 candidates (31 bacteria and 33 microeukaryotes) were separated into six clusterings with affinity propagation cluster (figure 5d). Among these, two clusterings contained most of the candidates were identified. For the first main clustering, 22 candidates were involved, and most of them were in Eukaryota. We, therefore, named this clustering as mEuk\_Cluster. In this clustering, the correlation between the enriched microeukaryotes was enhanced in CRC compared. We identified that *E. pulchra* and *A. rambellii* were the core microeukaryotes in the eEuk\_cluster. We also identified three CRC enriched bacteria, including *F. nucleatum*, *F. periodonticum*, and *P. micra* had strong correlations with these microeukaryotes (figure 5d). We revealed that the most outstanding microeukaryotes, *A. rambellii*, and the most reported cancer-related pathogens, *F. nucleatum*, were in the same clustering. Its z-score was -5.95, and it belonged to the ‘+/+’ case (see supplementary table 8), indicating that microeukaryotic *A. rambellii* and bacterial *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 8). In contrast, no direct correlation between another key microeukaryote *E. pulchra* and these three CRC-enriched bacteria was identified.

The second clustering included twenty-one candidates, and most of them 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. Collectively, these findings revealed that the correlations among microeukaryotes-microeukaryotes, microeukaryotes-bacteria, and bacteria-bacteria were quite different in CRC and identified that *A. rambellii* had a strong direct correlation with *F. nucleatum* and *P. micra*.

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

(TBD)