Study inclusion and data attainment

We used PubMed and Google scholar to search for CRC-related research containing at least CRC patients and healthy controls with faecal shotgun metagenomic data. And seven published studies and one of our previous research were included. We downloaded six public faecal shotgun CRC datasets from European Nucleotide Archive (ENA) using the following ENA identifiers: ERP005534 for Zeller et al.1, ERP008729 for Feng et al.2, PRJEB12449 for Vogtmann et al.3, PRJNA389927 for Hanningan et al.4, PRJEB27928 for Wirbel et al.5, and SRP136711 for Thomas et al.6. And the eighth cohort was downloaded from the DNA Data Bank of Japan (DDBJ) with the Accession numbers: DRA006684, DRA008156 for Yachida et al.7.

,

Hong Kong study recruitment and sequencing

This clinical study performed here was approved by the relevant ethics committees (Ethics Committee of Prince of Wales Hospital, Hong Kong, China, protocol NO. \*\*\*). Inform consent was obtained from all participants.

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. Deep freezing at –80°C within 24 hours of stool collection was done for long-term storage. According to the manufacturer's instructions, DNA was extracted using Qiagen (Hilden, Germany) QIAamp DNA Stool Mini Kit. All subjects had intact colonic lesions at the time of stool collection. An independent Chinese cohort of 112 control subjects, 111 patients with CRC, and 197 patients with colorectal adenomas were recruited. Part of the samples had been published in the previous research8.

Sample filter criteria

In the beginning, we included 2,052 individuals from eight countries and four continents among nine cohorts. We have three primary filtering sections (figure 1c). According to previously published meta-information, some outlier characters, such as history surgery patients, IBD patients, or other disease patients, would be filtered, and 1,986 samples were left after the first step. And then, 77 individuals would be disused because of the ambiguous stage. Free PCR is usual in the general whole-metagenomic-library preparation, but Hannigan’s research4 applied the 12 cycles of limited-cycle PCR in whole-metagenomic-library. The samples in this cohort would be filtered, and 1,837 cases would leave. One sample was filtered because of the low reads’ alignment (alignment reads number < 1,000,000). In the second filtering section, we intended to exclude the suspected contamination and outlier samples. Following the previous research9, microeukaryotes account for around 0.1% of the total intestinal flora. So, we discarded 19 high-microeukaryotes-abundance (RelAbuneuk > 1%) and 78 low-microeukaryotes-abundance (RelAbuneuk < 0.01%) samples, respectively. We recognized the samples whose one species accounted for more than 50% were contaminated. Therefore, we reduced the 69 large proportion of microeukaryotes and 45 large proportion of bacterial cases. Collectively, 221 samples were filtered in this section. Through the rarefaction curve (figure 1b), we could know that all cohort samples have reached or exceeded the plateau at 10,000. We abandoned the low microeukaryotes sequencing depth sample (RawReadseuk < 10,000) in the last part, and 296 cases were filtered. In summary, we move 216 cases for the sample sequence quality in the first section, 211 cases for reducing the outlier and contamination samples effect, and 296 cases for removing the low-microeukaryotes sequencing depth samples.

Sequence preprocessing 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 microeukaryotes 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. 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, and dividing the median of the control group of each feature in various cohorts with the script (https://github.com/ifanlyn95/multi-CRC-fungi).

Feature selections criteria

We had three criteria to select the potential candidates, whether it is bacteria or microeukaryotes. In the most beginning, we excluded the rarefied candidates with an average abundance of less than 0.1% in all microeukaryotes. 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 microeukaryotes 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 larger than 0.5. In addition, we ignored the unclassified strain of bacteria because we could not explain it.

Association calculation and comparison

Co-occurrence and co-exclusion relationships within microeukaryotes and between microeukaryotes and bacteria were estimated using the DGCA algorithm10, which is the methodology for systematical 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 larger 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 significant different correlation between different stages.

Additional validation experiments on cancer cell line

TBA