**Consistent filtering criteria and pre-processing of nine cohorts for the micro-Eukaryotic meta-analysis.** In this meta-analysis, eight published fecal 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 (see table 1 and Supplementary Table 1). Our indoor cohort was generated with the new fecal 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 (see figure 1a). And the median ratio of micro-Eukaryot to bacteria was 10e-2.80 (see 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 (see figure 1b), we could know that all cohort samples have reached or exceeded the plateau at 10,000. Hence, the minimum rarefies micro-Eukaryota 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 (see 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 (see 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.

**Enteric micro-Eukaryotic composition alterations in CRC compared to healthy control.** Consistent with previous studies and as a validation for our analysis, we observed bacteria phyla Bacteroidetes and Fusobateria were enriched in the CRC group compared with healthy control. Conversely, Firmicutes and Actinobacteria were reduced (see supplementary figure 1). Among the micro-Eukaryotic taxa, the phylum Ascomycota dominated the mycobiota, while Basidiomycota was observed as the second most abundant phylum (see 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 (see 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, Fusobateria performed significantly in 6 cohorts and enriched in CRC in all (see supplementary figure 2). But none showed a steady trend or difference in each study like Fusobateria in micro-Eukaryota phylum level (see 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 (see figure 2d). Most cohorts showed diversity reduction by alpha diversity index, chao1. Even though the alteration in micro-Eukaryotic is not as apparent as in bacteria level, it still offered some difference in CRC compared with healthy control.

**Univariate meta-analysis of micro-Eukaryotic species associated with CRC.** Our study's first and primary task was to identify the enteric micro-Eukaryota that are enriched or depleted in CRC metagenomes in a consistent manner across the eight populations. However,

**Alteration ecological association between micro-Eukaryotic and Bacteria in different stages.** (TBD)

**Comparison and clustering of the modified correlations between CRC and healthy control.** (TBD)

**Sub-title.** (TBD)

The first aim of the meta-analysis was to determine the gut microbial species that are enriched or depleted in CRC metagenomes in a consistent manner across the five study populations. However, since these studies differed from one another in many biological and technical aspects, we first quantified the effect of study-associated heterogeneity on microbiome composition. We contrasted this with other potential confounders (patient age, body mass index (BMI), sex, sampling after colonoscopy, and library size; additionally, smoking status, type 2 diabetes comorbidity, and vegetarian diet where available; Extended Data Fig. 1 and Supplementary Table 3). This analysis revealed the factor ‘study’ to have a predominant impact on species composition, which is supported by a recent comparison of DNA extraction protocols, since these typically differ between studies. An analysis of microbial alpha and beta diversity showed that study heterogeneity also had a larger effect on overall microbiome composition than CRC in our data (Extended Data Fig. 2).

Parametric effect size measures are not well established for the identification of microbial taxa significantly differing in abundance in CRC because microbiome data is characterized by non-Gaussian distributions with extreme dispersion; thus, we used a generalization of the fold change (Extended Data Fig. 3) and non-parametric significance testing. In this permutation test framework (herein referred to as blocked (univariate) Wilcoxon tests), differential abundance in CRC can be assessed while accounting for ‘study’ as a confounding effect that is treated as a blocking factor; additionally, motivated by our confounder analysis, we also blocked for ‘colonoscopy’ in all analyses (Methods and Extended Data Fig. 1). To rule out spurious associations due to the compositional nature of microbial relative abundance data, we additionally compared the results of this test with a method that employs log-ratio transformation and found highly correlated results (Supplementary Fig. 1 and Supplementary Table 4).

At a meta-analysis FDR of 0.005, we identified 94 microbial species to be differentially abundant in the CRC microbiome out of 849 species consistently detected across studies (Supplementary Table 4 and Methods). Among these, we focused on a core set of the 29 most significant markers (FDR < 1 × 10−5; Fig. 1a) for further analysis. The latter included members of several genera previously associated with CRC, such as Fusobacterium, Porphyromonas, Parvimonas, Peptostreptococcus, Gemella, Prevotella, and Solobacterium (Fig. 1b), and 8 additional species without genomic reference sequences (meta-mOTUs; Milanese et al.; see Methods) mostly from the Porphyromonas and Dialister genera and the Clostridiales order (see Extended Data Fig. 4 and Supplementary Table 4 for genus-level associations). Collectively, these 29 core CRC-associated species show a previously underappreciated diversity of 11 Clostridiales species to be enriched in CRC (Fig. 1b). In contrast to the majority of species that are more strongly affected by study heterogeneity than by CRC status, 26 out of the 29 CRC-associated species varied more according to disease status (Fig. 1d).

All of the core CRC-associated species were enriched in patients and were often undetectable in metagenomes from non-neoplastic CTRLs. While previous studies were contradictory in the reported proportion of positive versus negative associations, our meta-analysis results are more easily reconciled with a model in which—potentially many—gut microbes contribute to or benefit from tumorigenesis than with the opposing model where a lack of protective microbes contributes to CRC development (Fig. 1c). Although these core taxonomic CRC associations were highly significant and consistent, individual studies showed marked discrepancies in the species identified as significant (Fig. 1b). Retrospective examination of the precision and sensitivity with which individual studies detected this core of CRC-associated species showed relatively low sensitivity for the United States study (consistent with the original report) and low precision of the Austrian study due to associations that were not replicated in other studies (Supplementary Fig. 2).

Analyzing patient metagenomes for co-occurrences among the core set of 29 species that are strongly enriched in the CRC micro-biome revealed four species clusters with distinct taxonomic composition (Fig. 2a and Extended Data Fig. 5; Methods). Two of them showed strong taxonomic consistency: cluster 1 exclusively comprised Porphyromonas species and cluster 4 only contained members of the Clostridiales order. In contrast, the other two clusters were taxonomically more heterogeneous, with cluster 3 grouping together the species with the highest prevalence in CRC cases (all among the ten most highly significant markers), consistent with a co-occurrence analysis of one of the data sets included here. Cluster 2 contained species with intermediate prevalence.

Investigating whether these four clusters were associated with different tumor characteristics, we found the Porphyromonas cluster 1 to be significantly enriched in rectal tumors (Fig. 2b), consistent with the presence of superoxide dismutase genes in Porphyromonas genomes possibly conferring tolerance to a more aerobic milieu in the rectum (Extended Data Fig. 5). The Clostridiales cluster 4 was significantly more prevalent in female CRC patients. All species clusters showed a slight tendency toward late-stage CRC (that is, American Joint Committee on Cancer stages 3 and 4), but this was only significant for cluster 3. Associations with patient age and BMI were weaker and not significant (Extended Data Fig. 5). To rule out secondary effects due to differences in patient characteristics among studies, all of these tests were corrected for study effects (by blocking for ‘study’ and ‘colonoscopy’; see Methods). At the level of individual species, significant stage-specific enrichments could not be detected, suggesting CRC-associated microbiome changes to be less dynamic during cancer progression than previously postulated, although fecal material may be less suitable to address this question than tissue samples.