1. Supplementary Methods

Contents

1.1. Sample collection and DNA preparation

1.1.1. Sample collection in China

The study included adult individuals undergoing colonoscopy at the Shaw Endoscopy Centre at the Prince of Wales Hospital, the Chinese University of Hong Kong. The Chinese cohorts C1 (**Table S1**) and C2 (**Table S16)** included individuals presenting symptoms such as change of bowel habit, rectal bleeding, abdominal pain or anaemia, and asymptomatic individuals aged 50 or above undergoing screening colonoscopy. The exclusion criteria were: 1) use of antibiotics within the past 3 months; 2) on a vegetarian diet; 3) had an invasive medical intervention within the past 3 months; 4) had a past history of any cancer, or inflammatory disease of the intestine. Subjects were asked to collect stool samples in standardized containers at home, and store the samples in their home freezer immediately. Frozen samples were then delivered to the hospital in insulating polystyrene foam containers and stored at -80ºC immediately until further analysis. The study protocol in Hong Kong was approved by the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (CUHK-NTEC CREC).

1.1.2. Sample collection in Denmark

Cohort D: Stool samples were collected from individuals referred to colonoscopy due to symptoms associated with CRC or from patients who had been diagnosed with CRC and referred to large bowel resection for their primary cancer disease (See **Table S18**). All individuals were included at their visit to the out-patient clinic either before colonoscopy or before the operation and always before bowel evacuation. The individuals received a stool collection set including a tube without stabilizing buffer and were instructed to collect a stool sample at home one or two days before initiation of large bowel evacuation. Every included individual kept the sample refrigerated at -18ºC and contacted a research nurse who collected the sample. At the laboratory stool samples were immediately snap frozen in liquid nitrogen and subsequently stored at -80ºC under 24/7 electronic surveillance until analysis.

All included individuals thus underwent complete colonoscopy either as the primary examination or after the subsequent operation. Exclusion criteria were previous adenoma, previous CRC and previous or present other malignant diseases.

The recording of data from the included individuals was performed according to the Helsinki II declaration. The protocol was approved by the Ethics Committee of the Capital Region of Denmark (H-3-2009-110) and the Danish Data Protection Agency (2008-41-2252).

1.1.3. DNA extraction

Chinese samples: Stool samples were thawed on ice and DNA extraction was performed using the Qiagen QIAamp DNA Stool Mini Kit (Qiagen) according to manufacturer's instructions. Extracts were treated with DNase-free RNase to eliminate RNA contamination. DNA quantity was determined using NanoDrop spectrophotometer, Qubit Fluorometer (with the Quant-iTTMdsDNA BR Assay Kit) and gel electrophoresis.

Danish samples: A frozen aliquot (200 mg) of each fecal sample was suspended in 250 µl of 4 M guanidine thiocyanate– 0.1 M Tris (pH 7.5) and 40 µl of 10% N-lauroyl sarcosine. Then, DNA extraction was conducted using bead beating method as previously described[24]. The DNA concentration and its molecular size were estimated by nanodrop (Thermo Scientific) and agarose gel electrophoresis.

1.1.4. DNA library construction and sequencing

DNA library construction for samples from cohort C1, C2 and D was performed following the manufacturer's instruction (Illumina) at the same facility. We used a previously described workflow to perform cluster generation, template hybridization, isothermal amplification, linearization, blocking and denaturation, and hybridization of the sequencing primers[25].

We constructed one paired-end (PE) library with insert size of 350bp for each sample, followed by high-throughput sequencing to obtain around 30 million PE reads of length 2x100bp. High-quality reads were obtained by filtering low-quality reads with ambiguous 'N' bases, adapter contamination and human DNA contamination from the Illumina raw reads, and by trimming low-quality terminal bases of reads simultaneously.

1.2. Gene profile analysis

1.2.1. Generating gene profiles

We mapped our high-quality reads to a published reference gut microbial gene catalogue derived from European and Chinese adults[25] (using sequence identity >= 90%). We then derived the gene profiles using previously described procedures[25].

1.2.2. Bio-diversity analysis

Based on the gene profiles, we calculated the within-sample (alpha) diversity to estimate the gene richness using Shannon index and Simpson index of alpha diversity[25], where larger value indicates a higher degree of diversity in the sample. To analyse effects of different phenotype factors, including

age, BMI, eGFR, TCHO, LDL, HDL, and TG, on gut microbial diversity, Pearson correlation coefficients between each factor and Shannon index were also calculated.

1.2.3. Rarefaction analysis based on gene profile

Estimation of total gene richness in a set of metagenomics samples was performed by randomized sampling with replacement. This was done independently for cohort C1, CRC patients group in C1, and non-CRC control group in C1. In each set of size *n*, we randomly sampled *n* individual samples with replacement and then calculated the total number of genes that could be identified from these samples. Only genes with ≥1 mapping reads were considered to be present. This was repeated 100 times. The result showed that the control group had higher gene richness than the case group.

1.2.4. Analysis of factors influencing gut microbial gene profile

From the reference gene catalogue[25], we derived a subset of 2.1M genes that appeared in at least 6 samples in all 128 samples from cohort C1 (74 CRC and 54 control). We used the permutational multivariate analysis of variance (PERMANOVA) test to assess the effect of different characteristics, including age, BMI, eGFR, TCHO, LDL, HDL, TG, gender, DM, CRC status and location, on gene profiles of 2.1M genes (see Supplementary Table S1 for explanation of these factors). We performed the analysis using the implementation in the "vegan" package in R, and the permuted p-value was obtained by performing 10,000 permutations. We also adjusted for multiple testing using the function "p.adjust" in R with Benjamini-Hochberg method to get the corresponding q-values.

1.2.5. Identification of CRC associated genes

To identify the association between the metagenomic gene profiles and CRC, a two-tailed Wilcoxon rank-sum test was performed for each of the 2.1M genes. We obtained 140,455 gene markers which were enriched in either case or control with $P<0.01$. To control for colonoscopy as a confounding factor, we performed the independence test after stratifying by colonoscopy status, using the *kruskal_test* function implemented in COIN package in R.

1.2.6. Estimating the false discovery rate (FDR)

Instead of a sequential p-value rejection method, we applied the "qvalue" method proposed in a previous study[46] to estimate the FDR.

1.3. Taxonomic annotation of genes

1.3.1. Creating IMG genome database and species annotation of IMG genomes

Bacterial, archaeal and fungal genome sequences were extracted from IMG v400 reference database[27] downloaded from http://ftp.jgi-psf.org. In total, 522,093 sequences were obtained. For each IMG genome, using the NCBI taxonomy identifier provided by IMG, we identified the corresponding NCBI taxonomic classification at species and genus levels using NCBI taxonomy dump files. The genomes without corresponding NCBI species names were left with their original IMG names, most of which were unclassified at the genus and species levels.

1.3.2. Identification of CRC associated metagenomic linkage group (MLG) species

Based on the identified 140,455 CRC associated marker genes, we constructed the CRC associated MLGs using the method described in our previous study on type 2 diabetes[25]. All the above genes were aligned to the reference genomes of IMG database v400 to get genome level annotation. An MLG was assigned to a genome if >50% constituent genes were annotated to that genome, otherwise it was termed as unclassified. 86 MLGs consisting over 100 genes were selected as CRC associated MLGs. These MLGs were grouped based on the species annotation of these genomes to construct MLG species.

1.4. Data profile construction

1.4.1. Functional profiles based on KEGG database

Based on the gene profiles, we derived the KO profiles using previously described procedures[25]. Functional analysis was performed based on KEGG orthologous group (KO) abundance profiles. KEGG module and pathway (the KEGG Class Level 2) abundance profiles were calculated by summing the abundances of KOs belonging to each functional category.

1.4.2. Molecular operational taxonomic unit (mOTU) profiles

Clean reads were aligned to mOTU reference database (total 79268 sequences) with default parameters[26]. 549 species level mOTUs were identified, including 307 annotated species and 242 mOTU linkage groups (not to be confused with metagenomics linkage groups) without representative genomes. Most of the mOTU linkage groups were putatively Firmicutes or Bacteroidetes.

1.4.3. IMG-species and IMG-genus profiles

SOAP reference index was constructed for the IMG genome database based on 7 equal size chunks of the original file. Clean reads were aligned to reference using SOAP aligner[47] version 2.22, with parameters “-m 4 -s 32 -r 2 -n 100 -x 600 -v 8 -c 0.9 -p 3”. Then, SOAP coverage software was used to calculate read coverage of each genome, normalized with genome length, and further normalized to

relative abundance for each individual sample. The profile was generated based on uniquely mapped reads only.

1.4.4. MLG-species and MLG-genus profiles

To estimate the relative abundance of an MLG species, we estimated the average abundance of the genes of the MLG species, after removing the 5% lowest and 5% highest abundant genes. Relative abundance of IMG species was estimated by summing the abundance of IMG genomes belonging to that species. Genus abundances were estimated by analogously summing species abundances.

1.5. Biomarker discovery analysis

1.5.1. Minimum Redundancy Maximum Relevance (mRMR) framework

To establish CRC classification only using gut metagenomic markers, we adopted the mRMR method[28] to perform feature selection. We used the "sideChannelAttack" package from R to perform an incremental search and found 128 sequential marker sets. For each sequential set, we estimated the error rate by leave-one-out cross-validation (LOOCV) of a linear discrimination classifier. The optimal selection of marker sets was the one corresponding to the lowest error rate. In the present study, we made the feature selection on a set of 102,514 CRC associated gene markers. Since it was computationally prohibitive to perform mRMR using all genes, we derived a statistically non-redundant gene set. Firstly, we pre-grouped the 102,514 CRC associated genes that are highly correlated with each other (Kendall correlation > 0.9). Then we chose the longest gene as representative gene for the group, since longer genes have a higher chance of being functionally annotated, and will attract more reads during the mapping procedure. This generated a non-redundant set of 11,128 significant genes. Subsequently, we applied the mRMR feature selection method[28] to the 11,128 significant genes and identified an optimal set of 20 gene biomarkers that are strongly associated with CRC for classification.

1.5.2. Definition of CRC index

To evaluate the risk of CRC from the gut metagenome, we defined and computed a CRC index for each individual on the basis of the 20 gene markers identified by mRMR procedure. For each individual sample, the CRC index of sample *j* that denoted by I_j was computed by the formula below:

$$I_j = \left[\frac{\sum_{i\in N}\log_{10}\left(A_{ij}+10^{-20}\right)}{|N|} - \frac{\sum_{i\in M}\log_{10}\left(A_{ij}+10^{-20}\right)}{|M|}\right]$$

where A_{ij} is the relative abundance of marker i in sample j. N is a subset of all CRC-enriched markers in these 20 genes. M is a subset of all control-enriched markers in these 20 genes. And $|N|$ and $|M|$ are

the sizes of these two sets. The ability of the CRC index to distinguish CRC patient microbiomes from non-CRC microbiomes was examined using Wilcoxon rank-sum test. P-values estimated by these tests were adjusted for multiple testing using Benjamini-Hochberg method, when comparing CRC samples in cohort C1 with several other sample sets.

1.5.3. Receiver Operator Characteristic (ROC) analysis

We applied the ROC analysis to assess the performance of CRC classification based on metagenomic markers. We used the "Daim" package in R to draw the ROC curve.

1.5.4. Functional signatures associated with CRC

Wilcoxon rank-sum test with Benjamini-Hochberg adjustment was employed to identify KEGG KOs, modules and pathways associated with CRC.

1.5.5. Gut microbial species associated with CRC

Out of the 86 MLG species consisting over 100 genes, 85 MLGs were associated with CRC at a significance level of $q<0.05$ according to Wilcoxon rank-sum tests with Benjamini-Hochberg adjustment. This higher number is expected as the MLGs were constructed with genes that are associated with CRC in the first place. Using the same procedure at the same significance level, 28 IMG species and 21 mOTU species were associated with CRC.

1.5.6. Identifying gut microbial species that can classify CRC microbiomes

To evaluate the classification potential of the gut microbial species associated with CRC (identified by three methods: 85 MLG-species, 28 IMG species, and 21 mOTU species), we used "randomForest 4.5-36" package in R vision 2.10 based on these species profiles. For each method, firstly, we sorted all the *N* species by the importance given by the "randomForest" method. Then we created incremental marker sets by creating subsets of the top ranked species, starting from top 1 species and ending at *N* species. For each marker set, we calculated the false prediction ratio in Chinese cohort C1. Species from the marker set with lowest false prediction ratio were considered to have high potential for classification of CRC microbiomes from control microbiomes. Furthermore, we drew the ROC curve using the probability of illness based on these selected species markers.

1.5.7. Species co-occurrence network construction

Co-occurrence networks were constructed for the 85 MLGs, 28 IMG species and 21 mOTUs associated with CRC ($q<0.05$) using Spearman's correlation coefficient (>0.5 or <-0.5), as described previously[25]. Cytoscape[48] v3.0.2 was used to construct the three networks.

1.6. References

46 Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 2003;**100**:9440-5.
47 Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, *et al.* SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009;**25**:1966-7.
48 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 2003;**13**:2498-504.