Online supplementary material methods

Study cohorts and sample preparation

From 2017 to 2019, we collected 484 consecutive serum samples from individuals aged 45–75 years for this study; among them, stool samples of 44 individuals were also collected (Table S1). Our cross-sectional study was divided into 4 cohorts, namely the discovery cohort, the serum and feces matched cohort, the modeling cohort, and the validation cohort（Fig. 1B, Table S1). In brief, the discovery cohort was used for profiling CRC-associated serum metabolites by untargeted metabolomic analysis, and this cohort consisted of 92 individuals classified into three populations: Normal (N, n = 31), adenoma (A, n = 12), and colorectal cancer (C, n = 49). The serum and feces matched cohort was established for the integrated analysis of fecal metagenome and serum metabolomic analysis. Forty-four individuals, including 11 normal healthy volunteers and 33 patients with colorectal cancer or adenoma, were enrolled for the serum and feces matched cohort, and both the stool and serum were collected from the same individuals in this cohort. The modeling cohort was used to establish a diagnostic model and contained 192 individuals, including 72 normal and 120 colorectal abnormal patients. The validation cohort had 156 individuals, including 53 normal healthy volunteers and 103 colorectal abnormal patients, and the serum carcinoembryonic antigen (CEA) was recorded for each individual in this cohort.

Colorectal cancer was diagnosed by postoperative pathological examination and staged based on the tumor size, node, and the metastasis (TNM) staging system maintained by the American Joint Committee on Cancer and the International Union for Cancer Control. The blood or stool samples were taken before cancer treatment, and individuals who received preoperative radiation or chemotherapy treatment or had a previous history of CRC were excluded. The stool samples were

snap frozen by liquid nitrogen and stored at -80 °C until further usage. The baseline characteristics and clinical pathological features are shown in Table S1. Our study was approved by the Independent Ethics Committee of National Cancer Center/Cancer Hospital, Chinese Academy of Medical Science, and Peking Union Medical College.

Metabolite extraction

For untargeted metabolomics study, 240 μL of acetonitrile: isopropanol (3:1 by volume, Thermo Fisher) was added to 60 μL of serum, reaching a volume of 300 μL. To precipitate serum proteins, 60 μL of ammonium formate (0.5 g/mL, Thermo Fisher) was added along with 6 μL of aninternal standard solution containing 100 μg/mL L-Tyrosine-(phenyl-3,5-d2) (Sigma-Aldrich); 10 μg/mL 13C-Cholic Acid (Cambridge Isotope Laboratories) and 60 μg/mL Doxercalciferol (MedChem Express), vortexed for 4 min, and centrifuged at 17,949 g for 5 min. Then, 200 μL of the supernatant was transferred to another tube and dried by Centrivap cold-trap centrifugation at -60℃. Finally, the dried metabolite extracts were reconstituted with 75 μL of 55% methanol (Thermo Fisher) containing 0.1% formic acid (Thermo Fisher).

For targeted metabolomics analysis, similar sample extraction and preparation method was employed. Briefly, 6 μL of an internal standard solution (5 μg/mL 13C-Cholic Acid) was added to 60 μL of serum along with 240 μL of acetonitrile: isopropanol (4:1 by volume, Thermo Fisher) and 60 μL of ammonium formate (0.5 g/mL), followed by centrifugation at 17,949 g for 5 min. 80 μL of the supernatant was then diluted with 200 μL of water prior to sample analysis.

Quality control (QC) samples and QC matrix

An equal volume of serum derived from each individual in the normal population from the

discovery cohort was pooled together as the N-pool sample. The method for the generation of C-pool sample was the same as those used for the N-pool sample. A series of QC matrix were generated by mixing different volume of C-pool and N-pool samples. For each batch, 3 N-pool and 3 C-pool samples were used, serving as the quality control samples. A series of 7 additional QC matrix samples were prepared by mixing 10%, 20%, 30%, 40%, 50%, 75%, and 90% of C-pool samples with N-pool samples by volume, naming them as NC10, NC20, NC30, NC40, NC50, NC75, and NC90, respectively. The accuracy, precision, and linearity of the samples in the semi-quantitative untargeted metabolomic profiling were estimated as described in the data analysis section.

Untargeted metabolite profiling

Metabolites extracted from the serum of the discovery cohort and the feces matched cohort were analyzed by the Q Exactive mass spectrometer coupled with UltiMate3000 UPLC (Thermo Fisher). Data was acquired within the mass/charge ratio (m/z) range of 130 to 1200 Da at a resolution of 700,000 in the full MS-scan mode. The electrospray source conditions were set as follows: Sheath gas, 40 psi; capillary temperature, 320 °C; spray voltage, 3 kV (positive heating electrospray ionization (HESI)) and 3.2 kV (negative HESI). A CORTECS (Waters) 1.6 μm C18 2.1*100 mm column was used, with the oven temperature maintained at 35 °C. The flow rate was set at 0.3 ml/min and a 5 μL sample was injected. Mobile phase A (acetonitrile containing 0.1% formic acid) was applied as a gradient (from 5% to 45% at 0.5–14 min, 75% at 32 min, 80% at 42 min, 100% at 50–55 min, and back to 5% in next 5 min). Mobile phase B was the Merck Millipore water containing 0.1% formic acid. The resulting mass spectra were exported into the Progenesis QI Software (Nonlinear Dynamics, Durham, NC, USA) for further analysis.

Metabolomic data preprocessing

Peak extraction and alignment were performed using the Progenesis QI software. Parameters for mass and RT tolerance of metabolites were set as following: ± 5 ppm for mass tolerance, and 0.1 min for RT tolerance. To filter out background signals, metabolites with an abundance of less than 5000 in all individuals or with an equal to zero in more than 85% of the individuals were left out. To eliminate batch-to-batch differences, the R pre-process Core software package (v1.47.1) was used for Robust Multi-array Average (RMA) normalization and the abundance ratio of metabolites was calculated.

Metabolite annotation and inferring

Metabolite annotation was done as previously described with some modifications.[1] In brief, the MS-DIAL 4.24 program was applied for annotation. First, the QC MS1/MS2 spectrums were searched against the Human Metabolome Database (HMDB), MassBank of North America, and MassBank. Then, LipidBlast was performed using a default similarity cutoff score. Finally, a manual check with the reference database was done for confirmation and distinguishing between similar readouts. For metabolites whose MS/MS data could not be reliably acquired, their m/z were searched against the Metlin,HMDB and Bio-ML databases to determine their potential identities.[2] The confidence level for metabolite annotation was set as the following: MS/MS from referencing compounds in the current chromograph and MS condition > MS/MS from the library > m/z matching.

Metagenome Sequencing and taxonomic profiling

DNA extraction of the fecal samples was done by the QIAamp DNA Stool Mini Kit (QIAGEN), and 44 DNA samples in total passed the QC, including 11 normal and 33 colorectal abnormal individuals, and these individuals were involved in the serum and feces matched cohort. Whole-genome shotgun metagenome sequencing was used for the taxonomy and function analysis of the gut microbiome.[3] Library preparation and subsequent metagenomic sequencing were carried out on the HiSeq 4000 platform (Illumina) with 150 base pair paired-end reads at the Shanghai OE Biotech Co. Ltd, targeting >10 Gb of sequencing data per sample. The raw sequencing data was processed using the Trimmomatic V0.36 tool, including adapter trimming, depleting low quality reads or base pairs, and the Bowtie 2 tool was further used to remove host contaminations by mapping against the reference human genome (version hg38).[4 5] Subsequently, clean reads were constructed and further taxonomically profiled using the MetaPhlAn2 version 2.2.0 computational tool with default parameters.[6] In total, 12,445 microbiome species were profiled and among them, 640 species with relative abundances greater than 0.1% in at least one individual were considered for further microbiome-metabolome co-relation analysis.

Targeted metabolite profiling

Without using pure standards, we optimized the pseudotargeted method, described by Fujian Zheng [7], to determine the relative level of all metabolites in the identified panel by using the same reference pool for normalizing abundances of each individual. The ExionLC AC system was connected to a 6500 QTrap Mass Spectrometer (Sciex) and run in separate ion modes (positive and negative). The mobile phase and the column used for reversed-phase liquid chromatography were the same as those used for the untargeted metabolite profiling. The injection volume was 10 μL for each sample. The dwell time for each transition was 10 ms with the medium collision gas,

the curtain gas was 40 psi, the ion spray voltage was 5,000 V and -4,500 V, and the source temperature was 450 °C. Metabolites were eluted from the column at a flow rate of 0.3 ml/min with a gradually increasing concentration of mobile phase B, 12% of mobile phase B initially, to 60% of the mobile phase B after 2.5 min. A linear 60%-85% and 85%-100% phase B gradient was set at 6 min and 8.5 min. The quality control samples of the targeted analysis were pooled as follows: N-pool: C pool (1:1). Declustering potentials and collision energies were optimized from the quality control samples of the control group. Metabolite peaks were integrated using the Sciex Analyst 1.6.3 software.

Statistical analysis

Using ANOVA with Tukey's HSD test, we selected metabolites with an adjusted *p*-value of < 0.005 as significantly altered. Based on this result, metabolites with fold changes between the abnormal and normal populations being less than 1.2 and greater than 0.8 were filtered. The coefficient of variance (CV%) for each metabolic feature was calculated based on its abundances in the C-pool samples in each batch (5 batches of untargeted metabolite detection, with 3 C-pool samples enrolled in each batch). Metabolites with coefficient of variances (CV%) greater than 15% between batches were also filtered.

Estimating the accuracy, precision, and linearity of semi-quantitative untargeted metabolomic profiling

The QC samples including the C-pool and N-pool, as well as the mixed pool samples (NC10, NC20, NC30, NC40, NC50, NC75, and NC90) were built as described in the QC samples and QC matrix section, and used to calculate the accuracy of 13,666 negative and 14,758 positive ion

mode metabolites with normalized relative abundances above the background blank cut-off. Their accuracy was estimated by comparing their mixing ratio derived from the measured abundance with the expected mixing ratio between the C-pool and N-pool samples. The R^2 values of the linear regression mode between the expected mixing ratio and measured mixing ratio for each metabolite in both the negative and positive ion mode were distributed respectively, as shown in Fig. 1C. The R^2 values for more than 50% of the metabolites were greater than 0.9 in both the negative and positive ion mode, suggesting that not only a majority of metabolites can be measured with significant accuracy, but also indicating that the relative abundance of these metabolites show a robust linearity when ranging within 10% to 100% of different concentrations between the C-pool and N-pool samples. The precision of the metabolite profiling was evaluated by the root mean squared error (RMSE) of their linear regression model. The distribution of RMSE values for all the metabolite features shows that more than 50% of metabolites have RMSE values less than 0.2 in both the negative and positive ion mode (Fig. S1A–B). In conclusion, these metabolite profiling samples can be precisely and repeatedly measured in a semi-quantitative manner with high accuracy.

Selecting metabolites for the CRC GMSM panel to detect colorectal cancer

To select the metabolites features for the CRC GMSM panel, we implemented the LASSO algorithm with 10-fold cross validation for feature selection from the gut microbiome associated serum metabolomics data, as reported previously.[8] 322 metabolite features, which were significantly altered between the normal and the CRC or adenoma samples (adj. p < 5E-3), exhibited significant correlations with the gut microbiome (p <= 1E-3, FDR <= 18%). Using the panel voting approach, the features selected more than 75% of the time by 200 LASSO runs were

identified as the metabolite features for CRC GMSM panel. The chemical structure annotation, including MS2 ion pairs, if identifiable, was established by MS/MS spectrum matching as described previously.[9] Thus, 8 metabolites were identified as the members of the CRC GMSM panel. The chemical characteristics of the 8 inferred metabolites were first evaluated by targeted metabolomic analysis with the same discovery cohort, from which the corresponding features were initially uncovered by the un-targeted metabolomic analysis. As Fig. 3C shows that the AUC of the ROC for the same sample cohort based on the untargeted metabolomic analysis is 0.95, and that the AUC of the ROC based on the targeted metabolomic detection of these metabolites also reaches this value (Fig.3E), suggesting that the inferred metabolites can stably distinguish the normal samples from the CRC or adenoma samples.

Reference:

1. Liang L, Rasmussen M-LH, Piening B, et al. Metabolic Dynamics and Prediction of Gestational Age and Time to Delivery in Pregnant Women. *Cell* 2020;181(7):1680-92.e15. doi: https://doi.org/10.1016/j.cell.2020.05.002
2. Poyet M, Groussin M, Gibbons SM, et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nature Medicine* 2019;25(9):1442-52. doi: 10.1038/s41591-019-0559-3
3. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;464(7285):59-65.
4. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170
5. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012;9(4):357-59. doi: 10.1038/nmeth.1923
6. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* 2015;12(10):902-03. doi: 10.1038/nmeth.3589
7. Zheng F, Zhao X, Zeng Z, et al. Development of a plasma pseudotargeted metabolomics method based on ultra-high-performance liquid chromatography-mass spectrometry. *Nature protocols* 2020;15(8):2519-37. doi: 10.1038/s41596-020-0341-5
8. Wilmanski T, Rappaport N, Earls JC, et al. Blood metabolome predicts gut microbiome α-diversity in humans. *Nature Biotechnology* 2019;37(10):1217-28. doi: 10.1038/s41587-019-0233-9
9. Domingo-Almenara X, Montenegro-Burke JR, Benton HP, et al. Annotation: a computational solution for streamlining metabolomics analysis. *other* 2018;90(1)