Supplementary Methods

Tissue collection

Colon tissues were then collected and either fixed by formalin or snap-frozen in liquid nitrogen, followed by storage at -80℃. Stool samples were snap-frozen in liquid nitrogen, followed by storage at -80℃ for metagenomic sequencing and metabolomics analyses.

Histological examination

Colon tissues fixed by formalin were embedded in paraffin, sectioned and stained with hematoxylin and eosin. Microscopic examination of the sections was performed by pathologist blind to the groups.

Western blot

Total protein of colonic epithelium was extracted by tissue extraction reagent II (Thermo Fisher Scientific) followed by SDS-PAGE gel and transfer to nitrocellulose membrane. After blocking with 5% BSA, membranes were incubated by primary antibodies zo-1 (Thermo Fisher Scientific), claudin-3 (Thermo Fisher Scientific), β-actin (Thermo Fisher Scientific), ρ-ERK1/2 (Cell Signaling Technology, Waltham, MA), t-ERK1/2 (Cell Signaling Technology), and Proliferating cell nuclear antigen (PCNA) (Cell Signaling Technology) separately. The membranes were then incubated with secondary antibody and ECL Plus Western Blotting Detection Reagents (GE Healthcare, Chicago, IL). The protein band intensities were calculated by Image Lab software. Zo-1, claudin-3 and PCNA were normalized by the intensity of β-actin while ρ-ERK1/2 was normalized with intensity of t-ERK1/2.

Bai X, *et al. Gut* 2022; 71:2439–2450. doi: 10.1136/gutjnl-2021-325021

Quantitative reverse-transcription PCR (qRT-PCR)

Total RNA in colon tissue was extracted using Trizol Reagent (Thermo Fisher Scientific), then reverse-transcribed to complementary DNA using PrimeScript RT Reagent Kit with gDNA Eraser (Takara, Shiga, Japan). The relative expression level of gene was detected by QuantStudio™ 7 Flex Real-Time PCR System (Thermo Fisher Scientific) and normalized to the expression level of β-actin separately.

Quantitative polymerase chain reaction (qPCR) was performed to detect the *Eggerthella lenta* level by using 20 ng genomic DNA in 20 μL universal SYBR Green PCR Master Mix (Takara) on the QuantStudio 7 Flex Real-Time PCR System. Specific bacteria quantitation was measured relative to the universal 16s gene.

Primers used are listed in **Supplementary Table 1**. RT2 Profiler PCR Array Mouse Inflammatory Response and Autoimmunity (PAMM-077Z; QIAGEN, Hilden, Germany) and RT2 Profiler PCR Array Mouse Cancer Pathway Finder (PAMM-033Z; QIAGEN) were used to detect the changes in expression levels of inflammation-related and cancer-related genes.

Serum LPS Quantification

The serum LPS level was measured with an ELISA kit (Cusabio Technology Co., Ltd., Wuhan, China). All testing procedures were performed according to the manufacturer's instructions.

Bai X, *et al. Gut* 2022; 71:2439–2450. doi: 10.1136/gutjnl-2021-325021

16S ribosomal RNA gene sequencing and sequence curation and annotation

We extracted the DNAs of mice stools at three time points (initial time point, week8 and end time point). DNA library preparation and 16S ribosomal RNA gene sequencing were performed by NovoGene[1], Tianjin, China. The V3-V4 regions of 16S rRNA genes were amplified using specific primer (341F [CCTAYGGGRBGCASCAG] and 806R [GGACTACNNGGGTATCTAAT]) together with the barcode. The 16S rRNA gene sequence data were quality-filtered and analyzed using QIIME2 (version 2019.4.0) software[2]. The sequencing errors and replicated sequences were detected by Deblur algorithm. Before dereplicating sequences that encoded the amplicon sequence variants (ASV), paired reads were joined and trimmed to 404 base pairs. After filtering chimera sequences, the dereplicated sequences were classified taxonomically using Greengenes 16S rRNA gene reference database at a 99% identity cut-off by VSEARCH software. Beta diversity was measured by Arrhenius z distance, and Principal Coordinates Analysis (PCoA) was used for ordination analysis. Community dissimilarities were tested by permutational multivariate analyses of variance (PERMANOVA) with 1,000 iterations.

Detection of 3β-HSDH enzymes in metagenomic sequencing samples

We first extracted the gene sequences encoding 3β-HSDH enzymes from the *Eggerthella lenta* genome using samtools (version 1.9). The genome position of the genes was retrieved from MetaCyc or UniProt database and the corresponding *Eggerthella lenta* genome was downloaded from ENA database. Upon aligning the sequence reads to genes by bowtie2 (version 2.3.4.3), we extracted read count of genes and normalized the

Bai X, *et al. Gut* 2022; 71:2439–2450. doi: 10.1136/gutjnl-2021-325021

gene count by the library size.

Immunohistochemistry assay

Paraffin-embedded colon slides were deparaffinized, antigen-retrieved, blocked and incubated with anti-Ki67 antibody (Abcam, Cambridge, UK). The slides were counterstained with hematoxylin after secondary antibody incubation, enzyme conjugation, and DAB chromogen staining. The proliferation index was determined by the proportion of Ki-67 positive cells divided by total cells under the microscopic field. Five random fields were selected and examined for each sample.

Immunofluorescence assay

Paraffin-embedded colon sections were deparaffinized, antigen-retrieved, blocked and incubated with antibody zo-1 (Thermo Fisher Scientific, Waltham, MA) and claudin-3 (Thermo Fisher Scientific, Waltham, MA). The slides were then incubated with DAPI and examined under laser scanning confocal microscope (LEICA TCS SP8, Wetzlar, Germany).

Transmission electron microscope (EM)

Colon tissues from AOM/smoking mice and germ-free mice were collected and fixed in 2.0% glutaraldehyde. Ultrathin sections were prepared by ultramicrotome. The ultrastructure of tissues was examined using a transmission EM Philips CM100 (Philips, Amsterdam, Holland) at an acceleration voltage of 100 kV.

Bai X, *et al. Gut* 2022; 71:2439–2450. doi: 10.1136/gutjnl-2021-325021

Metabolomics analyses and metabolites profiling

Metabolite extraction from stool, non-targeted LC-MS/MS analysis and data preprocessing and annotation were performed by BIOTREE, Shanghai, China. Briefly, 100mg of stool sample was used for the UHPLC-QTOF-MS analysis. Non-targeted LC-MS/MS analyses were performed using UHPLC system (1290, Agilent Technologies, Santa Clara, CA) with a UPLC BEH Amide column (1.7μm 2.1*100mm, Waters Corporation, Milford, MA) coupled to TripleTOF 6600 (Q-TOF, AB Sciex, Redwood City, CA). The Triple TOF mass spectrometer was used to acquire MS/MS spectra on an information-dependent basis (IDA) during LC/MS experiment. MS raw data files were converted to mzXML format using ProteoWizard and processed by R package XCMS (version 3.2). The preprocessing results generated a data matrix that consisted of retention time (RT), mass-to-charge ratio (m/z) values, and peak intensity. R package CAMERA was used for peak annotation after XCMS data processing. MS2 database was applied in metabolites identification. The metabolomic data was analyzed using MetaboAnalystR R package. Significantly altered metabolites were determined by 2-tailed Mann-Whitney U test, and adjusted *P* (*FDR*) values < 0.05 were considered statistically significant. The association of differentially bacteria with metabolites were computed using Partial's Spearman correlation and heat maps were generated using ComplexHeatmap R package.

PCR array Enrichment analysis

The enrichment analysis was performed by hyper geometric test as follow:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where, *N* is the number of molecules (genes or metabolites) in the background, *K* is the number of molecules of a pathway, *n* is the number of significantly altered molecules and *k* is the number of shared molecules between the altered molecules (*n*) and the pathway (*K*). In particular, the background of PCR array is the detectable genes. The enrichment score was calculated by $k/(K \times n/N)$. *P* value of the hypergeometric distribution were calculated by the cumulative probability $P(X \geq k)$.

Targeted mass spectrometry assay for TDCA

Stock solution was prepared by diluting standard substance, TDCA, to give a final concentration of 10 mmol/L. A series of calibration standard solutions was then prepared by stepwise dilution of this standard solution (containing isotopically-labelled internal standard mixture in identical concentrations with the samples). A 25 mg aliquot of each individual stool sample was precisely weighed and transferred to an Eppendorf tube. After addition of 1000 μL of extract solution (acetonitrile-methanol-water, 2:2:1, containing 0.1% formic acid and isotopically-labelled internal standard mixture), the samples were vortexed, homogenized at 35 Hz, and sonicated in ice-water bath, followed by incubation and centrifugation. The resulting supernatants were transferred to LC-MS vials for UHPLC-MS/MS analysis (BIOTREE, Shanghai, China).

Data availability

The datasets generated in the current study are available in the Genome Sequence

Archive (GSA) at the National Genomics Data Center, Beijing Institute of Gemonics, Chinese Academy of Sciences / China National Center for Bioinformation (GSA: CRA006099)[3] and are publicly accessible at https://bigd.big.ac.cn/gsa/.

Statistical analysis

Data are shown as mean±standard deviation (SD) and compared using unpaired Student's t test or Mann-Whitney U test between two groups as appropriate. Fisher's exact test was used to evaluate the incidence of tumor variables between groups. Network analysis was performed by SparCC[4], which is commonly used to estimate correlations from compositional data. All differences were considered statistically significant if *P* values < 0.05. To account for multiple-testing, *P* values were adjusted using Benjamini-Hochberg false discovery (*FDR*) rate correction. GraphPad Prism 8.0 and open-source R software (version 3.5.2) was used to perform statistical analyses.

References

1. Chen K, Luan X, Liu Q, et al. Drosophila Histone Demethylase KDM5 Regulates Social Behavior through Immune Control and Gut Microbiota Maintenance. Cell Host Microbe 2019;25(4):537-52 e8. doi: 10.1016/j.chom.2019.02.003 [published Online First: 2019/03/25]
2. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 2019;37(8):852-57. doi: 10.1038/s41587-019-0209-9 [published Online First: 2019/07/26]
[dataset] 3. Bai X, Wei H, Liu W, et al. Data from: Smoking-exposed and smoking-free mice (Accession ID: CRA006099). Genome Sequence Archive (GSA), Feburay 17, 2022. https://bigd.big.ac.cn/gsa/
4. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol 2012;8(9):e1002687. doi: 10.1371/journal.pcbi.1002687 [published Online First: 2012/10/03]