Supplementary Materials

Microbiota studies

Bacterial DNA was isolated from mouse cecum digesta samples using the CTAB method according to Bousquet *et al.*[1]. DNA purity and concentration were monitored on 1% agarose gels, afterwards per DNA sample was diluted to 1 ng μl^{-1} with sterile water for serving as template for PCR amplicon.

The V3-V4 hypervariable regions of the bacteria 16 S ribosomal RNA gene was amplified by using special primers with barcode, PCR master mix with GC buffer, and high-efficiency and high-fidelity enzyme. All PCR reactions were carried out with Phusion® High-Fidelity PCR Master Mix (New England Biolabs) to ensure the efficiency and accuracy of amplification. PCR products were mixed with the same volume of 1X loading buffer (contained SYB green) and detected by electrophoresis on 2% agarose gel, thus the aim sample with bright main strip between 400 and 450 bp was chosen for further purification. Then, these PCR products were mixed in equimolar concentrations and purified using the Gel Extraction Kit according to the manufacturer's recommendations (Qiagen, Germany).

Sequencing libraries were generated using TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, USA) following manufacturer's instructions and index codes were added. The library quality was assessed on the Qubit@ 2.0 Fluorometer (Thermo Scientific) and Agilent Bioanalyzer 2100 system. At last, the library was sequenced on an IlluminaHiSeq2500 platform (USA MRDNA Molecular Research LP) based on standard protocols. As a consequence, 250 bp paired-end (PE) reads were generated.

PE reads assembly and quality control were performed based on the methods described previously [2-4]. Briefly, PE reads were assigned to each sample according to their unique barcode and truncated by cutting off the barcode and primer sequence. Using a computational software FLASH (version 1.2.7, http://ccb.jhu.edu/software/FLASH/), sequence assembly by overlapping PE reads from DNA fragment libraries of short length was designed for splicing and extending

the reads, thus raw tags were generated. According to the QIIME (version 1.7.0, http://qiime.org/index.html) quality controlled process, the high-quality clean tags were obtained by filtering the raw tags, and subsequently referred to the Gold database (http://drive5.com/uchime/uchime_download.html) using UCHIME algorithm (http://www.drive5.com/usearch/manual/uchime_algo.html) to generate the effective tags by detecting and removing chimera sequences [2].

The effective tags were assigned to operational taxonomic units (OTUs) cluster analysis with a 97% identity by Uparse software (version 7.0.1001, http://drive5.com/uparse/), and the sequence with the highest frequency was screened as the representative for each OUT[5]. Taxonomic data (kingdom, phylum, class, order, family, genus, species) were then assigned to each representative sequence against SSUrRNA Database (http://www.arb-silva.de/) of SILVA based on Mothur method algorithm (ID threshold: 0.8-1).

In order to study phylogenetic relationship of different OTUs, and the diversity of using the MUSCLE software [6] (version 3.8.31, http://www.drive5.com/muscle/). OTUs abundance information were normalized using a standard of sequence number corresponding to the sample with the least sequences.

Basing on the OTUs normalized data, α diversity and β diversity were performed. The *observed-species*, *chao1*, *shannon*, *simpson*, *ACE*, *good-coverage* were calculated with mothur project (Department of microbiology & immunology at the university of michigan, https://www.mothur.org/) as α diversity applied in analyzing species diversity for a sample. β diversity analysis on both weighted and unweighted unifrac was used to quantify differences of samples in species complexity by QIIME software (version 1.7.0) [7]. From β diversity analyses, principal component analysis (PCA) was generated by using the FactoMineR package and ggplot2 package in R software (version 2.15.3, https://www.r-project.org/) for demonstrating the clustering of different samples. Nonmetric multidimensional scaling (NMDS) diagrams were generated using the R package vegan to demonstrate the clustering of different samples. The linear discriminant analysis effect size (LEfSe) conducted by R vegan package (the default LDA score is 4) was used to select OTUs that exhibited significance in structural segregation among the grouping of samples [8].

For inter-relationship between bacterial species and TMA, TMAO production from choline-containing substrates, as well TMA lyase of all groups, the spearman correlation coefficients for environmental factors and significance of the correlation were performed using the corr. test function of the psych package in R software. Redundancy analysis (RDA) was performed in R using the vegan package with normalized OTUs abundance and environmental chemical data [9].

Metagenomic analysis of metabolic pathway and gene function

Metagenomic DNA preparation from digesta samples, metagenomic library construction and sequencing were performed as described in section "microbial analysis". The clean data for subsequent analysis was generated by removing low-quality bases (threshold value ≤ 38, length ≥ 40 bp) or high-percentage N base (length ≥ 10 bp) sequences and adapters (length ≥ 15 bp) from the raw data obtained from the Illumina HiSeq sequencing platform using Readfq (version 8, https://github.com/cjfields/readfq). These clean data were screened for host reads according to the host database and filtered with Bowtie software (version 2.2.4, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) using '--end-to-end, --sensitive, -I 200, -X 400' to remove host contaminants [10].

The set of high-quality clean data were assembled and analyst using SOAP denovo software (version 2.04) available on Sourceforge at http://soap.genomics.org.cn/soapdenovo.html, with the parameters '-d 1, -M 3, -R, -u, -F, -K 55' and k-mer length of 55 [11]. The assembled Scaffolds were interrupted from N connection, which yielded the Scaftigs without N that was compared with all samples' clean data to acquire the PE reads not used by the above-mentioned Bowtie software and parameters. The reads not used in the forward step of all samples were combined and submitted to the above-mentioned SOAP denovo with the same parameters for mixed assembly Scaffolds. Similarly, these mixed assembled Scaffolds were broken from N connection and obtained the Scaftigs [12]. The length of the Scaftigs derived from single and mixed assembly were trimmed and the reads shorter than 500 bp were removed with a custom R script [13].

The results of Scaftigs trimming with an average minimum of 500 bp were used to predict the Open Reading Frame (ORF) by MetaGeneMark software (version 2.10,

http://topaz.gatech.edu/GeneMark/), and subsequently the predicted result were filtered out the readers shorter than 100 nt with default parameters. For ORF predicted, the redundant gene catalogue were removed adopting CD-HIT software (version 4.5.8, http://www.bioinformatics.org/cd-hit) with the parameters option '-c 0.95, -G 0, -aS 0.9, -g 1, -d 0', and the unique initial gene catalogue (refers to the nucleotide sequences coded by unique and continuous genes) were acquired [14]. The sequence generating hits with 95% identity and 90% coverage were used to assign clones to broad phylogenetic groups, and the longest one was selected as the representative [15]. Reads mapping and the number of reads were prepared by aligning the clean data to initial gene catalogue with Bowtie 2.2.4 and parameter as mentioned above. The gene catalogue (Unigenes) was obtained via cutting off the gene with reads number ≤ 2 and used for subsequently analysis [15]. The abundance of a gene was calculated by counting the number of reads that aligned to gene map via the total number (r) and the length of gene (L) with the following format[16]:

$$G_K = \frac{r_k}{L_K} \cdot \frac{1}{\sum_{i=1}^{n} \frac{r_i}{L_i}}$$

Then the basic information statistic and correlation analysis of samples could be obtained based on the abundance of each gene in each sample in gene catalogue.

The predicted Unigenes were annotated based on KEGG (version 2018-01-01, http://www.kegg.jp/kegg/), CAZy (version 201801, http://www.cazy.org/), eggNOG (version 4.5, http://eggnogdb.embl.de/#/app/home), database using an *e*-value cutoff of 1e-15, by BLASTP program (http://www.ncbi.nlm.nih.gov/BLAST/) as part of the DIAMOND software package [17-19]. For each sequence's blast result, the best Blast Hit is used for subsequent analysis. The relative abundance of each functional hierarchy equal the sum of relative abundance of genes annotated to that functional level. Based on the abundance table of each taxonomy hierarchy, the abundance cluster heat map and comparative analysis of metabolic pathways were performed.

Transcriptome sequencing of intestinal epithelial tissue

A total amount of 1 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using NEBNext® UltraTM

RNA Library Prep Kit for Illumina® (NEB, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5X). First strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H^-). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, NEBNext Adaptor with hairpin loop structure were ligated to prepare for hybridization. In order to select cDNA fragments of preferentially 250~300 bp in length, the library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA). Then 3 μl USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated cDNA at 37°C for 15 min followed by 5 min at 95 °C before PCR. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer (NEB, USA). At last, PCR products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system (Agilent Technologies, USA).

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumia, USA) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Novaseq platform and 150 bp paired-end reads were generated.

Raw data (raw reads) of fastq format were firstly processed through in-house perl scripts. In this step, clean data (clean reads) were obtained by removing reads containing adapter, reads containing ploy-N and low quality reads from raw data. At the same time, Q20, Q30 and GC content the clean data were calculated. All the downstream analyses were based on the clean data with high quality.

Reference genome and gene model annotation files were downloaded from genome website directly. Index of the reference genome was built using Hisat2 v2.0.5 and paired-end clean reads were aligned to the reference genome using Hisat2 v2.0.5.

We selected Hisat2 as the mapping tool for that Hisat2 can generate a database of splice junctions based on the gene model annotation file and thus a better mapping result than other non-splice mapping tools.

Quantification of gene expression level feature Counts v1.5.0-p3 was used to count the reads numbers mapped to each gene. And then FPKM (expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced) of each gene was calculated based on the length of the gene and reads count mapped to this gene.

Differential expression analysis of groups (three biological replicates per group) was performed using the DESeq2 R package (version 1.16.1, http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html) [20]. DESeq2 provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting *P*-values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. Genes with an adjusted *P*-value < 0.05 found by DESeq2 were assigned as differentially expressed.

Gene Ontology (GO) enrichment analysis of differentially expressed genes was implemented by the clusterProfiler R package (http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html), in which gene length bias was corrected [21]. GO terms with corrected *P*-value less than 0.05 were considered significantly enriched by differential expressed genes. For understanding high-level functions and utilities of the biological system, from molecular-level information, especially large-scale molecular datasets, we used clusterProfiler R package to test the statistical enrichment of differential expression genes in KEGG pathways (http://www.genome.jp/kegg/) [22].

References

[1] Bousquet J , S. L., Lalonde M, DNA amplification from vegetative and sexual tissues of trees using polymerase chain reaction. *Can. J. For. Res.* 1990, *20*, 254-257.

[2] Bokulich, N. A., et al., Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 2013, *10*, 57-59.

[3] Caporaso, J. G., et al., QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 2010, *7*, 335-336.

[4] Tanja, M.; Salzberg, S. L., FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 2011, *27*, 2957-2963.

[5] Desantis, T. Z., et al., NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* 2006, *34*, W394-W399.

[6] Bing, L., et al., Characterization of tetracycline resistant bacterial community in saline activated sludge using batch stress incubation with high-throughput sequencing analysis. *Water Res.* 2013, *47*, 4207-4216.

[7] Lozupone, C., et al., UniFrac: an effective distance metric for microbial community comparison. *Isme J.* 2011, *5*, 169-172.

[8] Segata, N., et al., Metagenomic biomarker discovery and explanation. *Genome Biol.* 2011, *12*, R60.

[9] Sheik, C. S., et al., Exposure of soil microbial communities to chromium and arsenic alters their diversity and structure. *PLoS One* 2012, *7*, e40059.

[10] Karlsson, F. H., et al., Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat. Commun.* 2012, *3*, 1245.

[11] Nan, Q., et al., Alterations of the human gut microbiome in liver cirrhosis. *Nature* 2014, *513*, 59-64.

[12] HBjϕRn, N., et al., Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 2014, *32*, 822-828.

[13] Sunagawa, S., et al., Structure and function of the global ocean microbiome. *Science* 2015, *348*, 1261359.

[14] Fu, L., et al., CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012, *28*, 3150-2.

[15] Junhua, L., et al., An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 2014, *32*, 834-841.

[16] Emilie, V., et al., Ocean plankton. Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* 2015, *348*, 1261447.

[17] Kanehisa, M., et al., KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017, *45*, D353-D361.

[18] Huerta-Cepas, J., et al., eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 2016, *44*, D286-D293.

[19] Cantarel, B. L., et al., The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 2009, *37*, D233-8.

[20] Trapnell, C., et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 2012, *7*, 562-78.

[21] Young, M. D., et al., Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 2010, *11*, R14.

[22] Mao, X., et al., Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 2005, *21*, 3787-3793.