Supplementary methods

Metagenome sequencing and processing

Whole-genome shot-gun sequencing of fecal samples were carried out on the Illumina Hiseq X Ten. All samples were paired-end sequenced with a 150-bp read length. Raw reads were firstly trimmed for adapter sequences and primer sequences. Reads containing more than 40bp of low-quality (quality value less than 38) bases were removed. Reads containing more than 10bp of 'N' bases (base not identified) were removed. Reads with an overlap of more than 15bp with the adapter sequences were removed. Reads mapped to the human hg38 genome were removed. Finally, unpaired reads were discarded. As a result, a total of 5,256,644,960 (43,087,254 on average) paired-end reads with 788,496,744,000bp (6,463,088,066bp on average) as the high-quality reads (**see online supplementary table S8**) used for the following analysis.

After quality control, the high-quality paired-end reads were assembled into contigs using MEGAHIT (version 1.2.6)[1] with the minimum contig length set at 500 bp. The open reading frames (ORFs) were predicted from the assembled contigs using Prodigal (version 2.6.3)[2] with default parameters. The ORFs of <100 bp were removed. The ORFs were then clustered to remove redundancy using Cd-hit (version 4.6.6)[3] with a sequence identity threshold set at 0.95 and the alignment coverage set at 0.9, which resulted in a catalog of 4,047,645 non-redundant genes. The non-redundant genes were then collapsed into metagenomic species (MGS)[4 5] and grouped into KEGG functional modules.[4]

Non-redundant gene catalogue construction

Reads were assembled into contigs using MEGAHIT[1] (version 1.2.6) with the minimum contig length set at 500 bp and default parameters for metagenomics. The open reading frames (ORFs) were predicted from the assembled contigs using Prodigal[2] (version 2.6.3) with default parameters. The ORFs of <100 bp were removed. The ORFs were clustered to remove redundancy using Cd-hit[3] (version 4.6.6) with a sequence identity threshold set at 0.95 and the alignment coverage set at 0.9, which resulted in a catalogue of 4,047,645 nonredundant genes.

Identification of metagenomic species

High-quality reads were mapped to the catalogue of nonredundant genes using Bowtie 2[6] (version 2.2.9) with default parameters. The abundance profile for each catalogue gene was calculated as the sum of uniquely mapped sequence reads, using 19M sequence reads per sample (downsized). The co-abundance clustering of the 4,047,645 genes was performed using canopy algorithm[5] (http://git.dworzynski.eu/mgs-canopy-algorithm), and 553 gene clusters that met the previously described criteria[5] and contained more than 700 genes were referred to as MGS. MGS that were present in at least 4 samples were used for the following analysis. The abundance profiles of MGS were determined as the medium gene abundance throughout the samples. MGS were taxonomically annotated by summing up the taxonomical annotation of their genes as described by Nielsen *et al*.[5] Each MGS gene was annotated by sequence similarity to the bacterial complete genome in NCBI Reference Sequence Database (BLASTN, E-value<0.001).

Annotation of KEGG modules

The catalogue of the nonredundant genes was functionally annotated to the KEGG database (release 94.0) by KofamKOALA (version 1.3.0).[7] The produced KEGG Orthologs (KOs) were mapped to the KEGG

modules annotation downloaded on 1 August 2020 from the KEGG BRITE database. KOs that were present in at least 4 samples were used for the following analysis. The KO abundance profile was calculated by summing the abundance of genes that were annotated to each of the KOs.

Gavage experiments using ascorbate in the CIA mouse model

Nine healthy seven-week-old DBA/1 mice weighing 20g were fed in an ultra-clean animal laboratory (SPF grade) with a humidity of 55% and a temperature of 26°C. CIA models were constructed and established as described before.[8] Nine mice were then divided into three groups (three mice per group), including normal DBA/1 mice and two groups of DBA/1 mice with CIA. Three-month gavage (0.3ml/d) to three groups of mice was conducted, including 1) 0.9% normal saline to normal DBA/1 mice, 2) 0.9% normal saline to DBA/1 mice with CIA, and 3) 100ng/ul ascorbate to DBA/1 mice with CIA. After three-month gavage, the mice plasma TNF-α level and the IL-6 level were tested using ELISA kit (mlbio, China). Mice were then killed and preserved in 4% formalin for two days. Micro-CT (QuantumGX, PerkinElmer, UnitedStates) was used to perform scanning and three-dimensional structural reconstruction of the joints. The settings were set to 209m, 90kV X-ray tube voltage, 160uA of the current, and 3 minutes of the scan time. The angle of the X-ray scan rotated 180 degrees. We have adhered to standards articulated in the Animal Research: Reporting of In Vivo Experiments (ARRIVE).

Joint synovial fluid sampling

Synovial fluid samples were collected aseptically from knee joints during therapeutic aspiration. Synovial fluid samples were deposited in sterile tubes on ice and homogenized within five minutes of collection. A tube filled with sterile phosphate-buffered saline (PBS) was left open throughout the procedure and subsequently processed in parallel with the samples as a negative control. The entire experiment was conducted in a completely sterile atmosphere. Each sample was immediately frozen and kept without heparin or hyaluronidase at –135 °C. A total of 7ml synovial fluid was collected, of which 5ml was utilized for 16S rRNA gene sequencing, 1ml was used for bacteria isolation, and 1ml synovial fluid was prepared for scanning electron microscopy.

16S rRNA gene sequencing and processing

V1-V2 regions of 16S rRNA gene was sequenced on the Illumina Hiseq 2500. Raw reads were firstly trimmed for adapter sequences and primer sequences. Using split_libraries.py in QIIME (version 1.9.1),[9] reads were splitted according to the barcodes, and reads with average quality less than 25 were removed. We then obtained a total of 7,932,312 (92,236 on average) paired-end reads with 1,991,010,312bp (23,151,283bp on average). The paired-end sequences were then joined using the FLASH with default parameters.[10] In addition, Chimeric sequences were identified and removed using de novo chimera detection of USEARCH (version 6.1).[11] Finally, we obtained a total of 3,798,209 (44,165 on average) sequences for the following analysis. The remained sequences were clustered into operational taxonomic units (OTUs) at a 97% sequence similarity, using open-reference OTU picking protocol in QIIME (version 1.9.1).[9] Taxonomy of OTUs was annotated using the RDP classifier[12] against the Greengenes database (release 13_8) with 0.8 confidence.

Bacterial isolation of the joint synovial fluid

Due to that pH of joint synovial fluid was 7.6, we used Luria-Bertani (LB) broth medium with a pH of 7.6 for cultivation. The procedures were as follows: 1) sterilized LB medium (pH7.6) and inoculation

equipment were placed in an anaerobic glove box (Ruskinn Concept 400) for deaeration one day in advance; 2) 1ml synovial fluid samples per stage of RA were used for bacterial cultivation. 3) The synovial fluid sample was serially diluted with sterilized water, plated onto LB (0.5% yeast extract, 1% tryptone, 1% sodium chloride) agar medium, and then incubated at 37°C for 72 h to obtain single colonies. Then, three colonies per plate were randomly selected and streaked three consecutive times on LB agar medium to obtain a pure culture, which was named isolate SF1 to SF9, respectively. To identify isolate SF1 to SF9, a partial fragment of 16S rDNA was amplified with the primer pair 27F (5'-AGAGTTTGATCCTGGCTCAG-3') and 1492R (5'-GGTTACCTTGTTACGACTT-3'), and then DNA sequencing was performed for preliminary identification.

Scanning electron microscopy for bacteria in the joint synovial fluid

The synovial fluid sample was filtered through the membrane (special membrane for flow cytometry) to remove solids and large particles (Filtrate A). Filtrate A was then filtered through a 0.45 μmum membrane to remove most human cells (Filtrate B). Filtrate B was then filtered through a 0.22 μmum filter membrane to enrich bacteria and subsequently washed with sterilized ultra-pure water to obtain Filtrate C. Filtrate C was soaked in 2.5% glutaraldehyde for 4 hours at room temperature. Filtrate C was then washed three times with 0.1 M PBS buffer and treated with 1% osmic acid for 4h. Filtrate C was then dehydrated in ethanol, vacuum dried by tert-butyl alcohol, coated with gold, and imaged with a scanning electron microscope (ZEISS Sigma 300).

UHPLC-QTOF-MS analysis of plasma metabolites

Plasma samples were thawed at 4°C on ice, and 100μL of the sample was then placed in an EP, extracted with 300μL of the extraction solvent (methanol with internal standard of 2μL/mL), followed by vortex for 30s, treated with ultrasound for 10min (incubated in ice water), and incubation for 1h at –20°C to precipitate proteins. Then samples were centrifuged at 12,000rpm for 15 min at 4°C. Subsequently, 100μL of the supernatant was transferred into a fresh LC/MS glass vial, and 20μL of the supernatant of each sample was pooled as QC samples, and 300μL of the supernatant was used for following analysis.

Liquid chromatography with tandem mass spectrometry (LC-MS-MS) analysis was performed using an UHPLC system (1290, Agilent Technologies) with an ACQUITY UPLC BEH Amide column (1.7μm 2.1*100mm, Waters) coupled to TripleTOF 6600 (Q-TOF, AB Sciex) & QTOF 6550 (Agilent). The mobile phase consisted of 25mM NH_4Ac and 25 mM NH_4OH in water (pH=9.75) (A) and acetonitrile (B), which was carried with elution gradient as follows: 0 min, 95% B; 0.5min, 95% B; 7min, 65% B; 8 min, 40% B; 9 min, 40% B; 9,1 min, 95% B; 12 min, 95% B, delivered at 0.5mL/min. The injection volume was 2μL. The Triple TOF mass spectrometer was used for its ability to acquire MS/MS spectra on an information-dependent bases (IDA) during an LC/MS experiment. In this mode, the acquisition software (Analyst TF 1.7, AB Sciex) continuously evaluated the full scan survey MS data as it collected and triggered the acquisition of MS/MS spectra depending on preselected criteria. In each cycle, 12 precursor ions whose intensity greater than 100 were chosen for fragmentation at collision energy (CE) of 30V (15 MS/MS events with product ion accumulation time of 50 msec each). ESI source conditions were set as following: Ion source gas 1 as 60 Psi, Ion source gas 2 as 60 Psi, Curtain gas as 35 Psi, source temperature 600°C, Ion Spray Voltage Floating (ISVF) 5000 V or –4000 V in positive or negative modes, respectively. MS raw data files (.wiff) were convered to the mzXML format using ProteoWizard, and processed by R package XCMS. The preprocessing results generated a data matrix that consisted of the

retention time (RT), massto-charge ratio (m/z) values, and peak intensity. R package CAMERA was used for peak annotation after XCMS data processing. In-house MS2 database was applied in metabolites identification. The relative intensity was determined by peak area normalization and was used for the following analysis.

References

1. Li D, Luo R, Liu CM, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 2016;102:3–11.
2. Hyatt D, Chen GL, Locascio PF, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform* 2010;11:119.
3. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22(13):1658–9.
4. Pedersen HK, Forslund SK, Gudmundsdottir V, et al. A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links. *Nat Protoc* 2018;13(12):2781–800.
5. Nielsen HB, Almeida M, Juncker AS, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 2014;32(8):822–8.
6. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9(4):357–9.
7. Aramaki T, Blanc-Mathieu R, Endo H, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 2020;36(7):2251–52.
8. Song G, Feng T, Zhao R, et al. CD109 regulates the inflammatory response and is required for the pathogenesis of rheumatoid arthritis. *Ann Rheum Dis* 2019;78(12):1632–41.
9. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;7(5):335–6.
10. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 2011;27(21):2957–63.
11. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;26(19):2460–1.
12. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;73(16):5261–7.