Supplementary information

A hidden link in gut-joint axis: Gut microbes promote rheumatoid arthritis at early stage by enhancing ascorbate degradation

Yan Zhao,[1#] Mingyue Cheng,[2#] Liang Zou,[1] Luxu Yin,[1] Chaofang Zhong,[2] Yugo Zha,[2] Xue Zhu,[2] Lei Zhang,[3,4]* Kang Ning,[2]* Jinxiang Han[1]*

[1]Shandong Medicine and Health Key Laboratory of Rheumatism, Shandong Key Laboratory of Rheumatic Disease and Translational Medicine, Department of Rheumatology and Autoimmunology, Shandong Provincial Qianfoshan Hospital, First Affiliated Hospital of Shandong First Medical University. Jinan 250014, Shandong, China

[2]Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of AI Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

[3]Microbiome-X, National Institute of Health Data Science of China & Institute for Medical Dataology, Cheeloo College of Medicine, Shandong University, Jinan 250014, Shandong, China

[4]Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250014, Shandong, China

***Correspondence to** Dr Jinxiang Han, E-mail: jxhan@sdfmu.edu.cn; Dr Kang Ning, E-mail: ningkang@hust.edu.cn; Dr Lei Zhang, E-mail: zhanglei7@sdu.edu.cn.

[#]These authors contributed equally to this work

Sample description

A total of 122 fecal and 122 serum samples were collected from 122 outpatients from the Shandong Provincial Qianfoshan Hospital (Jinan, Shandong, China). These outpatients included 27 healthy individuals, 19 patients with osteoarthritis (OA), and 76 patients with rheumatoid arthritis (RA). Subsequently, the fecal samples were sequenced and the serum samples were used to examine serum metabolites and inflammatory cytokines. Serum inflammatory cytokines TNF-α and IL-6 were quantified by the MESO SCALE DISCOVERY (MSD®) Quick Plex S600MM multiplex assay. The cytokine levels of healthy individuals were extremely low and not available. In addition, 95 knee-joint synovial fluid samples were collected from the RA and OA patients to examine synovial fluid metabolites. Both serum and synovial fluid metabolites were examined by UHPLC-MS/MS.

All of the participants were at fasting status during sample collection in the morning. The participants were recruited in this study following the standards shown below:

1. Healthy individuals in good health condition with no gastrointestinal diseases, such as diarrhea, constipation, and hematochezia, in the recent one month, no hepatobiliary system diseases, no history of gastrointestinal tumors or inflammatory diseases, no serious heart, liver, kidney, lung, brain or other organ disorders, no infections, chronic diseases, or antibiotic treatment;
2. Healthy individuals had not taken any acid inhibitors, gastrointestinal motility drugs, antibiotics, or living bacteria products such as yogurt in the recent one month;
3. Healthy individuals with no history or family history of mental illness, and no history of gastrointestinal surgery;
4. RA/OA individuals with no other co-morbidity.

Metagenome sequencing and data 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. After quality control, the 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]

Identification of MGS

High-quality reads were mapped to the catalog of non-redundant genes using Bowtie 2 (version 2.2.9)[6] 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 (http://git.dworzynski.eu/mgs-canopy-algorithm),[5] and 553 gene clusters that met the previously described criteria[5] and contained more than 700 genes were referred to as MGS. MGS 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 as described by Nielsen *et al.*[5] and each MGS gene was annotated by sequence similarity to NCBI bacterial genome (BLASTN, E-value < 0.001)

Annotation of KEGG modules

The catalog of the non-redundant genes was functionally annotated to KEGG database (release 94.0) by KofamKOALA (version 1.3.0).[7 8] The produced KEGG Orthologies (KOs) were mapped to the KEGG modules annotation downloaded on August 1, 2020 from the KEGG BRITE database. KOs present in at least 4 samples were used for the following analysis. The KO abundance profile was calculated by summing the abundances of genes that were annotated to each KO.

Clustering of co-abundant metabolites

Co-abundant metabolites in serum or synovial fluid were identified using the R package WGCNA[9]. As recommended by Pedersen *et al.*,[4] a signed network and biweighted mid-correlation were used for clustering with the soft threshold $\beta = 8$ for both serum and synovial fluid metabolites. The minimum cluster size was set as 3. Similar clusters were subsequently merged if the biweight mid-correlation between the cluster's eigen vectors exceeded 0.8 for both serum and synovial fluid metabolites. The kIN of a metabolite was calculated by summing connectivity with all other metabolites in the given metabolite cluster. The kME was determined by the bicor-correlation between the metabolite profile and module eigenvector. Both kIN and kME were used to measure the intramodular hub-metabolite status.

Cross-domain association analyses

The clinical phenotypes, including types of arthritis (Healthy = 0, OA = 1, RA = 2) and the levels of pro-inflammatory cytokines TNF-α and IL-6, were used in the association analysis. TNF-α and IL-6 were selected based on their potentials to act as the therapeutic targets for RA treatment.[10 11] The associations between clinical phenotypes and KEGG modules/metabolites clusters were determined through evaluating if the Spearman correlations of the phenotype with the abundances of KOs/metabolites in the given KEGG module/metabolite clusters were significantly higher or lower (Mann–Whitney U-test FDR < 0.1) than with the abundances of all other KOs/metabolites. The phenotypes adjusted by age and gender were also tested. Moreover, the union set of the significant associations between KEGG modules and phenotypes/phenotypes adjusted by age and gender, and the intersect set of the significant associations between metabolites clusters and phenotypes/phenotypes adjusted by age and gender, were used for the following association analysis. The associations between metabolite clusters and

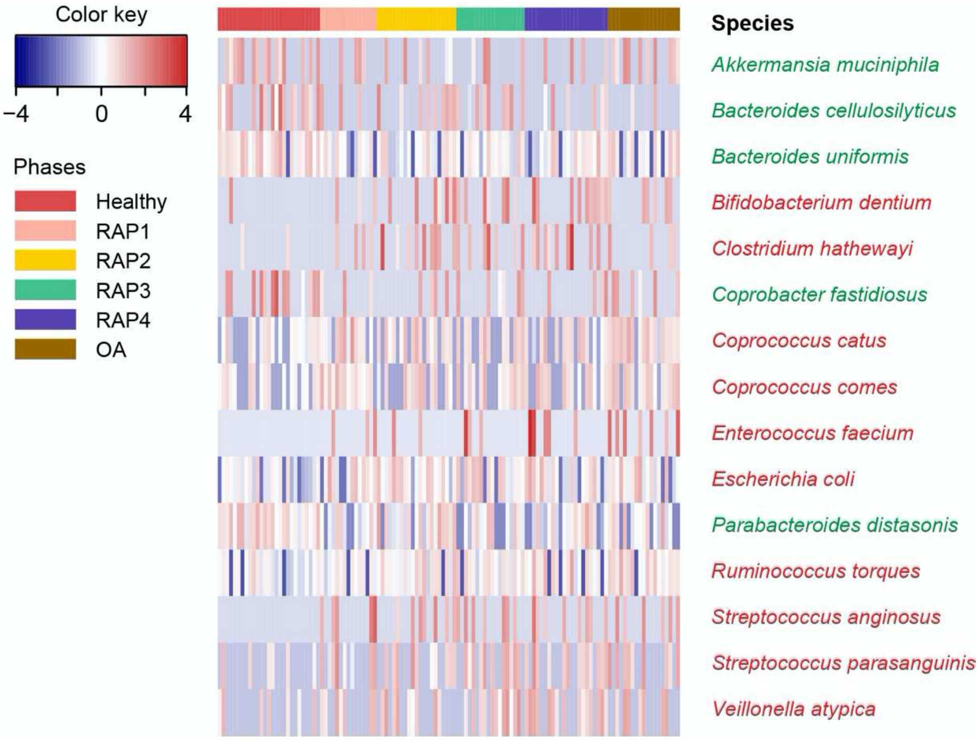
KEGG modules were determined through evaluating if the Spearman correlations of the eigen vectors of the metabolite clusters with the abundances of KOs in the given KEGG module were significantly higher or lower (Mann–Whitney U-test FDR < 0.1) than with the abundances of all other KOs/metabolites.

Leave-one-out analysis

Leave-one-out analysis was used to identify the specific MGS driving the observed associations between KEGG module M00550 and the clinical phenotypes, including the types of arthritis or the levels of pro-inflammatory cytokines TNF-α and IL-6. The calculation of the KO abundance was iterated excluding the genes from a different MGS, in each iteration. The effect of a given MGS on a specified association was defined as the change in median Spearman correlation coefficient between KOs and clinical phenotypes when genes from the respective MGS were left out, as previously described.[4 12]

Taxonomic identity of differentially present microbes across conditions

MetaPhlAn2[13] was used to generate species profiles. Species that were present in less than 10% samples were excluded. Supplementary Figure 1 displays the union set of the species (n=15) with significantly different abundances (Mann–Whitney U-test FDR < 0.05) between the healthy and RA groups or between the healthy and OA groups.



Supplementary figure 1 Taxonomic identity of differentially present microbes across conditions. Each row represents a species with significantly different abundances (Mann–Whitney U-test FDR < 0.05) between the healthy and RA groups or between the healthy and OA groups. Each column represents a sample from one of the groups including the healthy, RAP1, RAP2, RAP3, RAP4, and OA groups. Color of each heatmap unit represents the scaled abundance of a certain species in a specific sample. Species are colored for significantly elevation (red) or depletion (green) in the arthritis groups, in comparison with the healthy groups.

Data accession

Whole-genome shot-gun sequencing data are available in the Genome Sequence Archive (GSA) section of National Genomics Data Center (project accession number CRA004348) at https://bigd.big.ac.cn/gsa/browse/CRA004348.

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 Bioinformatics* 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;22:1658–9.

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;13:2781–800.

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;32:822–8.

6. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9(4):357–9.

7. Kanehisa M, Goto S, Sato Y, et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;42:D199–205.

8. 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.

9. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559.

10. Ceccarelli F, Massafra U, Perricone C, et al. Anti-TNF treatment response in rheumatoid arthritis patients with moderate disease activity: a prospective observational multicentre study (MODERATE). *Clin Exp Rheumatol* 2017;35(1):24–32.

11. Nakahara H, Nishimoto N. Anti-interleukin-6 receptor antibody therapy in rheumatic diseases. *Endocr Metab Immune Disord Drug Targets* 2006;6(4):373–81.

12. Pedersen HK, Gudmundsdottir V, Nielsen HB, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* 2016;535(7612):376–81.

13. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015;12:902–3.