OPEN ACCESS

TRANSLATIONAL SCIENCE

Stage-specific roles of microbial dysbiosis and metabolic disorders in rheumatoid arthritis

Mingyue Cheng ,[1,2] Yan Zhao,[1] Yazhou Cui,[1] Chaofang Zhong,[2] Yuguo Zha,[2] Shufeng Li,[1] Guangxiang Cao,[1] Mian Li,[1] Lei Zhang,[3] Kang Ning ,[1,2] Jinxiang Han [1]

Handling editor Josef S Smolen

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/ard-2022-222871).

For numbered affiliations see end of article.

Correspondence to
Dr Jinxiang Han, Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan, Shandong, China; jxhan@sdfmu.edu.cn, Dr Kang Ning, Huazhong University of Science and Technology, Wuhan, Hubei, China; ningkang@hust.edu.cn and Dr Lei Zhang, Shandong University, Jinan, Shandong, China; zhanglei7@sdu.edu.cn

MC and YZ contributed equally.

MC and YZ are joint first authors.
LZ, KN and JH are joint senior authors.

Received 31 May 2022
Accepted 7 August 2022
Published Online First 19 August 2022

To cite: Cheng M, Zhao Y, Cui Y, *et al*. *Ann Rheum Dis* 2022;**81**:1669–1677.

ABSTRACT

Objective Rheumatoid arthritis (RA) is a progressive disease including four stages, where gut microbiome is associated with pathogenesis. We aimed to investigate stage-specific roles of microbial dysbiosis and metabolic disorders in RA.

Methods We investigated stage-based profiles of faecal metagenome and plasma metabolome of 76 individuals with RA grouped into four stages (stages I–IV) according to 2010 RA classification criteria, 19 individuals with osteoarthritis and 27 healthy individuals. To verify bacterial invasion of joint synovial fluid, 16S rRNA gene sequencing, bacterial isolation and scanning electron microscopy were conducted on another validation cohort of 271 patients from four RA stages.

Results First, depletion of *Bacteroides uniformis* and *Bacteroides plebeius* weakened glycosaminoglycan metabolism ($p<0.001$), continuously hurting articular cartilage across four stages. Second, elevation of *Escherichia coli* enhanced arginine succinyltransferase pathway in the stage II and stage III ($p<0.001$), which was correlated with the increase of the rheumatoid factor ($p=1.35\times10^{-3}$) and could induce bone loss. Third, abnormally high levels of methoxyacetic acid ($p=1.28\times10^{-8}$) and cysteine-S-sulfate ($p=4.66\times10^{-12}$) inhibited osteoblasts in the stage II and enhanced osteoclasts in the stage III, respectively, promoting bone erosion. Fourth, continuous increase of gut permeability may induce gut microbial invasion of the joint synovial fluid in the stage IV.

Conclusions Clinical microbial intervention should consider the RA stage, where microbial dysbiosis and metabolic disorders present distinct patterns and played stage-specific roles. Our work provides a new insight in understanding gut–joint axis from a perspective of stages, which opens up new avenues for RA prognosis and therapy.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Rheumatoid arthritis (RA) is a progressive disease, clinically including four stages.
⇒ Intestinal microbiome is associated with the pathogenesis of RA.
⇒ Joint synovial fluid is generally considered as sterile.

WHAT THIS STUDY ADDS

⇒ This is the first study focusing on RA stages to report microbial and metabolic profiles and roles, particularly their enhancement of inflammation, bone loss and bone erosion in the stage II and stage III.
⇒ Joint synovial fluid is not sterile, where bacterial invasion happened in the stage IV.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The study provides microbial and metabolic targets for each stage of RA.
⇒ Further experiments and intervention on microbiota of joint synovial fluid are warranted for patients in the stage IV.

INTRODUCTION

Rheumatoid arthritis (RA) affects over tens of millions of people worldwide.[1] RA is recognised clinically as a progressive, inflammatory and autoimmune disease that primarily affects the joints and typically has four stages[2–5]: (1) In the first stage, the synovium of the joints is inflamed and most people have minor symptoms such as stiffness on awakening; (2) In the second stage, the inflamed synovium has caused damage to the joint cartilage and people begin to feel swelling, and have a restricted range of motion; (3) In the third stage, RA has proceeded to a severe state when bone erosion begins and the cartilage on the surface of the bones has deteriorated, resulting in the bones rubbing against one another and (4) In the fourth stage, certain joints are severely deformed and lose function. To inhibit RA progression, specific therapeutic strategies are necessary for people across different RA stages.

Gut microbial dysbiosis has been implicated in the pathogenesis of RA via a range of mechanisms such as metabolic perturbation and immune response regulation, which is known as the gut–joint axis,[6 7] for instance, increased abundance of *Prevotella* and *Collinsella* in patients with RA are correlated with the production of T_H17 cell cytokines.[8 9] Moreover, Gut microbes and their products were likely to be transited to the joint due to the increased gut permeability.[6] Metabolites have also been correlated to immunity regulation in RA: administration of Short-chain fatty acids to mice with collagen-induced arthritis (CIA) can reduce the severity of arthritis by modulation of IL-10.[6 10] Comprehensive metagenomic and metabolomic analyses could

Table 1 General characteristics at stool collection of multiomics cohort (122 participants)

	Patients with RA from four stages					
	RAS1 n=15	RAS2 n=21	RAS3 n=18	RAS4 n=22	OA n=19	HC n=27
Age (years), median (IQR)	52 (50–60)	64 (59–67)	59 (50–66)	60 (54–66)	66 (64–71)	56 (50–60)
Female sex, n (%)	12 (80)	16 (76)	12 (67)	21 (95)	15 (79)	19 (70)
Classification score						
A (IQR)	3 (3–3)	3 (3–3)	5 (5–5)	5 (5–5)	1 (0–1)	0
B (IQR)	2 (2–2)	3 (3–3)	3 (3–3)	3 (3–3)	0	0
C (IQR)	1 (1–1)	1 (1–1)	1 (1–1)	1 (1–1)	1 (1–1)	0
D (IQR)	1 (1–1)	1 (1–1)	0 (0–0)	1 (1–1)	1 (1–1)	0
Sum score (IQR)	7 (7–7)	8 (8–8)	9 (9–9)	10 (10–10)	3 (2–3)	0
ACPA positivity, n (%)	5 (33)	18 (86)	13 (72)	17 (77)	0	
ESR (IQR)	60.00 (30.50–87.00)	72.00 (45.00–116.00)	48.00 (29.50–60.00)	60.00 (40.50–80.00)	23.00 (15.75–41.50)	
CRP (IQR)	29.10 (9.20–53.40)	43.00 (15.95–56.55)	38.60 (12.50–44.40)	27.15 (20.32–64.45)	17.45 (4.62–65.75)	
RF (IQR)	52.00 (39.50–60.50)	421.00 (186.00–590.00)	34.00 (19.50–80.00)	279.50 (106.00–359.00)	26.00 (22.00–29.00)	

ACPA positivity was defined as a concentration of greater than 5 µL/mL.
Classification scores were summarised according to 2010 RA classification criteria: A, joint involvement; B, serology; C, acute-phase reactants; D, duration of symptoms.
ACPA, anticitrullinated protein antibody; CRP, C reactive protein; ESR, erythrocyte sedimentation rate; HC, healthy controls; OA, osteoarthritis; RA, rheumatoid arthritis; RAS1–4, the first to fourth stage of RA; RF, rheumatoid factor.

therefore enhance our understanding about the gut–joint axis. However, the role of the gut–joint axis across successive stages of RA is understudied,[6 11] where more examinations may provide an alternative approach to ameliorate RA progression.

Here, we aimed to investigate the stage-based profiles and roles of the gut–joint axis in RA pathogenesis, and whether or in which stage gut microbial invasion of the joint synovial fluid happened.

MATERIALS AND METHODS

Study design and sample collection

Data collection for this multiomics study was conducted in The First Affiliated Hospital of Shandong First Medical University (Jinan, Shandong, China), which was a provincial-level large-scale comprehensive tertiary first-class hospital and had tens of thousands of outpatients with arthritis per year. A total of 122 faecal and 122 plasma samples were collected from 122 outpatients of the The First Affiliated Hospital of Shandong First Medical University from 2017 to 2020.[7] These outpatients included 76 patients with RA, 19 patients with OA and 27 healthy individuals (table 1). Patients with RA were grouped into four RA stages including RAS1 (n=15), RAS2 (n=21), RAS3 (n=18) and RAS4 (n=22) according to the rheumatoid diagnostic score,[3] where RAS1, RAS2, RAS3 and RAS4 has a score of 6–7, 8, 9 and 10, respectively. The score was evaluated by the sum of four categories as summarised in the 2010 RA classification criteria.[3] Faecal samples were collected and sequenced and plasma samples were used to test the plasma metabolites, anticitrullinated protein antibody, erythrocyte sedimentation rate, C reactive protein, rheumatoid factor, cytokines and plasma metabolites.

To confirm the bacterial invasion of the joint synovial fluid, another cohort of 271 with RA of four distinct stages were recruited, including 52 patients in RAS1, 66 in RAS2, 67 in RAS3 and 86 in RAS4. Synovial fluid samples were collected aseptically from knee joints during therapeutic aspiration. The entire experiment was conducted in a completely sterile atmosphere. For each patient, a total of 7 mL synovial fluid was collected, of which 5 mL was used for 16S rRNA gene sequencing, 1 mL was used for bacteria isolation and 1 mL synovial fluid was prepared for scanning electron microscopy.

All of the participants were at fasting status during the sample collection in the morning. Only participants who met the standard were recruited in this study: Recruited individuals had not received treatment in the recent month and were in the active period, and had no malignant tumour, no other rheumatic diseases such as ankylosing spondylitis, psoriasis, gout, no gastrointestinal diseases such as diarrhoea, constipation and haematochezia in the recent month, no infections, no other comorbidity such as diabetes and hepatitis B.

Metagenomic sequencing and processing to analyse the faecal microbiome

Whole-genome shotgun sequencing and processing of faecal samples, non-redundant gene catalogue construction, identification of metagenomic species (MGS), functional annotation to Kyoto Encyclopaedia of Genes and Genomes (KEGG) were performed (details in online supplemental text). Two parallel processes were used for gut metagenomic data analysis: One was based on 4 million non-redundant genes and investigated the functional composition across RA stages and OA, as well as the MGS that most drove the correlation of these microbial functions with RA or OA and (2) The other reported the 232 classified microbial species composition across RA stages and OA, profiled by MetaPhlAn2[12] (V.2.7.8).

UHPLC-QTOF-mass spectrometry analysis of plasma metabolites

Untargeted plasma metabolome was examined by ultra-performance liquid chromatography-quadrupole time-of-flight (UHPLC-QTOF) mass spectrometry: liquid chromatography with tandem mass spectrometry on an UHPLC system (1290, Agilent Technologies) with a UPLC BEH Amide column (1.7 µm 2.1×100 mm, Waters) coupled to TripleTOF 6600 (Q-TOF, AB Sciex) and QTOF 6550 (Agilent) (details in online supplemental text).

16S rRNA gene sequencing and processing to analyse the synovial fluid microbiota

Bacterial DNA was extracted from 271 5 mL synovial fluid samples. The tube containing PBS serves as environmental

control. Only a total of 86 synovial fluid samples from patients in RAS4 had enough bacteria DNA content (≥10 ng) (Bacterial DNA Kit, TIANGEN) for bacteria 16S rRNA gene high-throughput sequencing. The V1/V2 hypervariable regions of the 16S ribosomal RNA gene were sequenced using the Illumina HiSeq platform. The 16S sequence paired-end data set was joined and quality filtered using the FLASH as previously described.[13] Taxonomic annotation was then performed (details in online supplemental text).

Bacterial isolation and scanning electron microscopy

Six synovial fluid samples (1 mL) per RA stage were used for bacteria isolation, and the obtained isolated colonies were identified using 16S rRNA gene sequencing (details in online supplemental text). For the samples from which bacterial can be isolated, synovial fluid samples (1 mL) of the same individuals were then filtered and imaged with scanning electron microscope (ZEISS Sigma 300, details in online supplemental text).

Statistical analysis

Samples were divided into three groups including the healthy group, the OA group and the RA group. Samples of the RA group were further divided into four subgroups including RAS1, RAS2, RAS3, RAS4. For comparisons of vectors across groups or subgroups, such as microbial species abundance, KO abundance, metabolite intensity. Mann-Whitney-Wilcoxon test (p values) with Benjamini and Hochberg correction (q values) was used to test the significance. A threshold for statistical significance was $p<0.05$, and for multiple testing the threshold was $p<0.05$ and $q<0.1$.

For correlations between KEGG modules and clinical phenotypes including arthritis (healthy=0, OA=1, RA=2), cytokine levels and rheumatic factor level, owing to that a KEGG module contained multiple KOs, Spearman correlation coefficients (SCC) between abundances of KOs and clinical phenotypes were first calculated. Subsequently, Mann-Whitney-Wilcoxon test (p values) with Benjamini and Hochberg correction (q values) was used to test if SCC between the KOs in a given KEGG module and phenotypes were different from that between all the other KOs out of the KEGG module and phenotypes. In this process, the KEGG module with statistical significance was viewed as significantly correlated with the clinical phenotypes. A threshold for statistical significance was $p<0.05$ and $q<0.1$. Considering that sex and age might have potential effects on gut microbiome,[14] partial SCCs with age and gender adjusted were also calculated and compared, and a threshold for statistical significance was $p_{partial}<0.05$ and $q_{partial}<0.1$.

Leave-one-out analysis was used to test which MGS was driving the observed correlations between KEGG modules and arthritis. Owing to that one MGS contained multiple genes that were mapped to KOs, if one MGS was excluded in the dataset, the overall profiles of the KO abundance would change, resulting in the change of the correlations between KEGG modules and arthritis. Therefore, to determine the driving effects of each of MGS, the calculation of the KO abundance was iterated excluding the genes from a different MGS in each iteration, and the correlations between each KEGG module and arthritis were recalculated. Finally, the driving effects of a given MGS on a specified correlation was defined as the change in median SCC between KOs and arthritis when genes from the respective MGS were left out.

To determine the diagnostic potential of RA stages using multiomics features, random forest algorithm was performed on 6,224 KOs, 232 microbial species and 277 plasma metabolites, using the R package 'randomForest'. Function 'trainControl' in R package 'caret' was used to perform 10 repeats of 10-fold cross-validation for each data set. Function 'train' in R package 'caret' was used to fit models over different tuning parameters to determine the 'mtry' for random forest algorithm. Gini coefficients were used to measure how each variable contributed to the homogeneity of the nodes and leaves in the resulting random forest.

RESULTS

Stage-specific microbial taxonomic profiles

We obtained a total of 231 classified microbial species from metagenomic data, and tested their alterations in each stage of RA, as compared with healthy controls (see online supplemental figure S1, table S1–S5). The elevated species in RA progression were mostly from the phyla Firmicutes and Actinobacteria, while the depleted species were predominantly from the phylum Bacteroides ($q<0.1$). We found certain microbes did not remain altered across RA stages, as compared with healthy controls. *Bifidobacterium dentium*, for instance, was reported to be associated with the development of dental caries and periodontal disease, both of which were particularly prevalent in patients with RA.[15 16] Compared with healthy controls, it remained elevated across RA stages except for RAS1 (RAS2: $p=7.16\times10^{-3}$, RAS3: $p=3.70\times10^{-3}$, RAS4: $p=9.15\times10^{-4}$). Moreover, we noticed that 29 species that were altered exclusively in a specific stage (see online supplemental table S1–S5). We found that *Collinsella aerofaciens* was elevated exclusively in RAS1 ($p=0.043$). *C. aerofaciens* was previously reported to generate severe arthritis when inoculated into CIA-susceptible mice, and an in vitro experiment showed that *C. aerofaciens* could increase gut permeability and induce IL-17A expression, a key cytokine involved in RA pathogenesis.[9] The elevation of *C. aerofaciens* in RAS1 might contribute to the early breach in gut barrier integrity, through which the translocation of microbial products would then trigger the subsequent clinical arthritis.[6] Moreover, *Veillonella parvula*, whose infection could cause osteomyelitis,[17] was found elevated exclusively in RAS3 ($p=0.027$). *Eggerthella lenta* ($p=0.018$) and *Bifidobacterium longum* ($p=0.022$) were found elevated exclusively in RAS4. The gavage of *E. lenta* were reported to increase gut permeability and produce proinflammatory cytokines.[18] We also recognised species altered exclusively in OA, such as elevated *Dialister invisus* ($p=0.041$) that was positively correlated with spondyloarthritis severity.[19] These stage-specific altered species had the potential to serve as the targets for intervention in a given RA stage.

Stage-specific microbial functional profiles

Next, we sought to detect the microbial dysfunction across stages of RA. We grouped 4 047 645 metagenomic genes into 6,224 KOs and 404 KEGG modules. We identified 12 KEGG modules that were significantly correlated with RA or OA ($q<0.1$ or $q_{partial}<0.1$, see online supplemental figure S2) and presented their variation across stages (figure 1A). We then used leave-one-out analysis to identify the MGS that most drove the correlations of these KEGG modules with RA or OA (figure 1B, online supplemental figure S3).

We found an evident decrease in glycosaminoglycan (CAG) metabolism across four RA stages and OA. It was mainly reflected by the significant decrease in K01197 (hyaluronoglucosaminidase) of dermatan sulfate (DS) degradation and the significant decrease in K10532 (heparan-alpha-glucosaminide

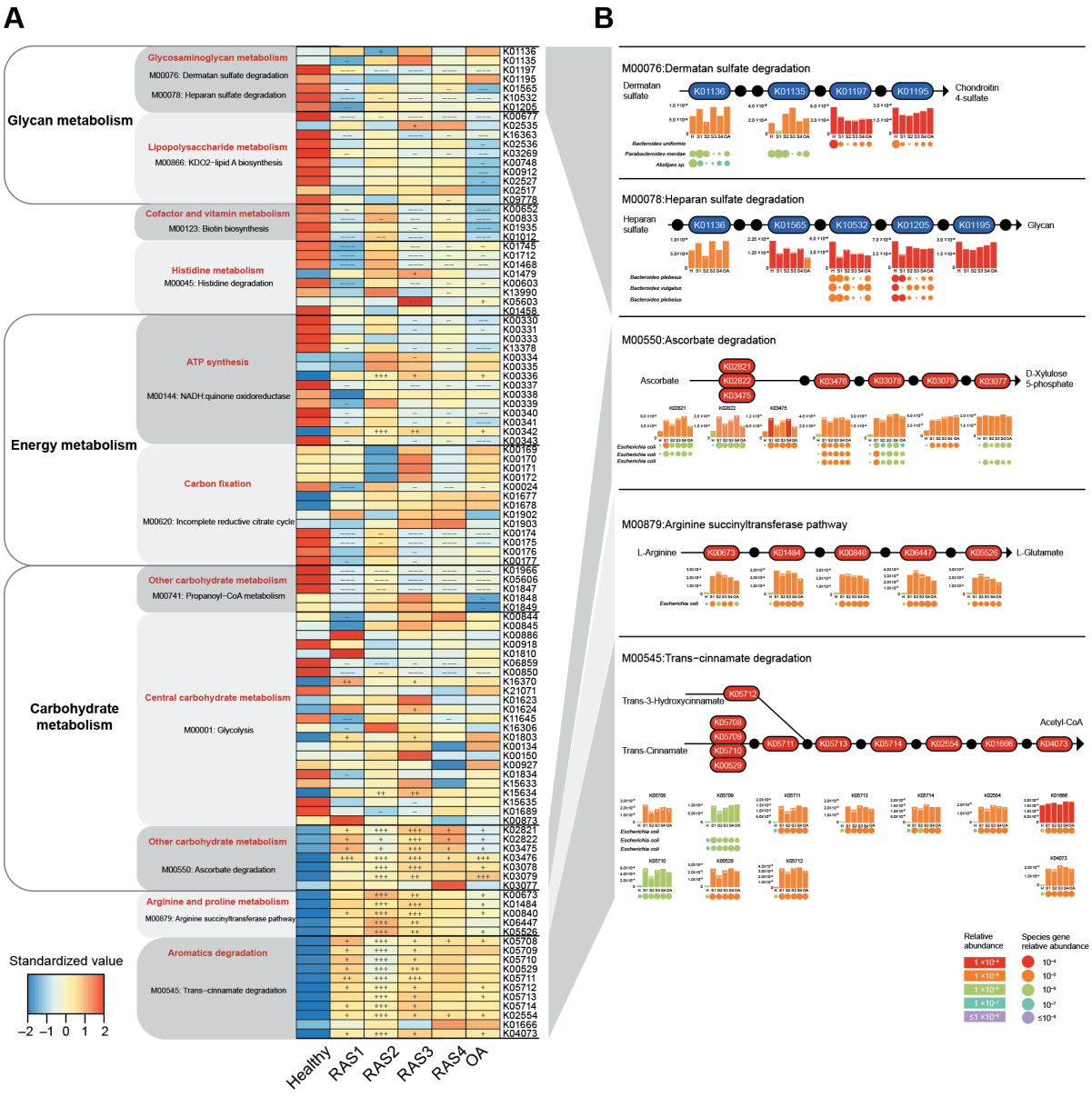


Figure 1 Stage-specific microbial functional profiles. Gene abundances were assessed for elevation or depletion in each of the arthritis stages, RAS1 (n=15), RAS2 (n=21), RAS3 (n=18), RAS4 (n=22) and OA (n=19) compared with the healthy individuals (n=27). (A) Relative abundance of KO genes in the KEGG modules that were significantly correlated with arthritis ($q<0.1$ or $q_{partial}<0.1$, see online supplemental figure S1). KO genes with a prevalence of 5% or higher are shown. (B) KO genes involved in specific KEGG pathway modules in (A) are shown in the KEGG pathway maps. Each box in a pathway represents a KO gene and is marked in red for elevation or in blue for depletion at any of the stages compared with healthy individuals. Bar plots show relative gene abundances averaged over samples within each of the five groups (healthy (H), RAS1 (S1), RAS2 (S2), RAS3 (S3), RAS4 (S4) and OA) and are coloured according to the values. Each KO gene is composed of MGS genes represented by circles. The sizes and colours of the circles are proportional to the relative abundances of the MGS genes. MGS genes are grouped into one row and indicated by the taxonomic name. The three MGS that most drove the correlation of the KEGG modules with arthritis types are shown. In all panels, significant changes are denoted as follows: +++, elevation with $p<0.005$; ++, elevation with $p<0.01$; +, elevation with $p<0.05$; −−−, depletion with $p<0.005$; −−, depletion with $p<0.01$; −, depletion at $p<0.05$; Mann-Whitney-Wilcoxon test. KEGG, Kyoto Encyclopaedia of Genes and Genomes; KO, KEGG ortholog; MGS, metagenomic species; OA, osteoarthritis.

N-acetyltransferase) of heparan sulfate (HS) degradation ($p<0.05$, figure 1A). Chondroitin 4-sulfate is a major component of the extracellular matrix of many connective tissues, such as cartilage, bone and skin.[12] We found that the significant depletion of DS degradation would inhibit the production of chondroitin 4-sulfate (figure 1B), which might hurt the mechanical properties of the articular cartilage.[12] Moreover, the significant depletion of HS degradation might be a potential cause of the higher plasma level of HS observed in RA and OA patients,[20 21] which could promote arthritis progression by regulating protease activity.[22] The most driving species of DS degradation and HS degradation were MGS *Bacteroides uniformis* and MGS *Bacteroides plebeius*, respectively. The genes of MGS *B. uniformis* related to K01197 were found most depleted in RAS2, while the genes of MGS *B. plebeius* related to K10532 were found most depleted in RAS3 and RAS4 (figure 1B). These results indicated that the depleted microbial function in DS degradation and HS degradation driven by *B. uniformis* and *B.*

plebeius, respectively, could promote RA and OA in a way of hurting articular cartilage.

We also identified elevated microbial functions that were related to inflammation such as the previously reported ascorbate degradation.[7] Here, we found most of the KOs related to ascorbate degradation retained a higher level across RA stages and OA, especially in RAS2 and RAS3 ($p<0.05$, figure 1A). Genes of K02821 (phosphotransferase system) in RAS1, K03475 (phosphotransferase system), K03476 (L-ascorbate 6-phosphate lactonase), and K03479 (L-ribulose-5-phosphate 3-epimerase) were mostly driven by MGS *Escherichia coli* (figure 1B). The enhanced ascorbate degradation might contribute to the deficiency of the ascorbate reported in patients with RA[23] and were found positively correlated with multiple plasma cytokines ($q<0.1$ or $q_{partial}<0.1$, see online supplemental table S6), such as IL-1β ($p=5.44\times10^{-4}$), TNF-α ($p=6.59\times10^{-4}$) and IL-6 ($p=1.12\times10^{-3}$). Moreover, to confirm the effects of ascorbate on RA progression, we examined the plasma TNF-α level and IL-6 level, bone CT scans, and bone density of (1) normal DBA/1 mice, (2) DBA/1 mice with CIA and (3) DBA/1 mice with CIA and gavage of ascorbate. We found that the 3-month gavage of ascorbate to CIA mice can prevent the increase of TNF-α and IL-6 levels by half, inhibit bone destruction, and maintain bone density ($1.58\pm0.0034\,g/cm^3$), as compared with the CIA mice without ascorbate ($1.53\pm0.013\,g/cm^3$), and the normal group ($1.61\pm0.021\,g/cm^3$, see online supplemental figure S4).

For other elevated microbial functions, the trans-cinnamate degradation driven by MGS *E. coli*, where most KOs were notably elevated in RAS2, was also correlated with multiple cytokines ($q<0.1$ or $q_{partial}<0.1$, see online supplemental table S6), such as IL-13 ($p=1.63\times10^{-5}$), IL-1β ($p=2.87\times10^{-5}$) and IL10 ($p=4.10\times10^{-5}$). Moreover, the arginine succinyltransferase pathway driven by MGS *E. coli* was found significantly elevated mainly in RAS2 and RAS3 (figure 1). L-arginine is able to prevent bone loss induced by zinc oxide nanoparticles or by ciclosporin A, through anti-inflammatory mechanism[24] or nitric oxide production, respectively.[25] Both arginine succinyltransferase pathway and trans-cinnamate degradation was positively correlated with the elevation of rheumatoid factor ($p=1.35\times10^{-3}$). Taken together, these results suggested that microbial dysfunction could promote RA progression mainly by hurting bone tissue and strengthening inflammation. The inflammation-related microbial dysfunction was extremely active in RAS2 and RAS3 and largely driven by *E. coli*.

Microbial invasion of the joint synovial fluid

Next, we investigated whether or in which stage microbial invasion of the joint synovial fluid happened. Enhanced gut permeability may render it possible for microbes and their products to translocate, triggering an immune response.[6 26] We thus speculated that gut microbes might invade the joint synovial fluid of patients with RA through the gut–joint axis. To test this, we performed 16S rRNA gene sequencing on the synovial fluid samples from another cohort of 271 patients in four RA stages, including RAS1 (n=52), RAS2 (n=66), RAS3 (n=67) and RAS4 (n=86). Notably, we were not able to obtain enough bacterial DNA for sequencing in samples of RAS1, RAS2 or RAS3, however, we could identify many microbes in samples of RAS4 (see online supplemental figure S5). We found that most of the microbes in joint synovial fluid were from phyla Proteobacteria and Firmicutes, and a total of 98 genera could also be detected in faecal metagenomic data (see online supplemental table S7). Moreover, we could recognise *E. lenta* and *B. longum* in most

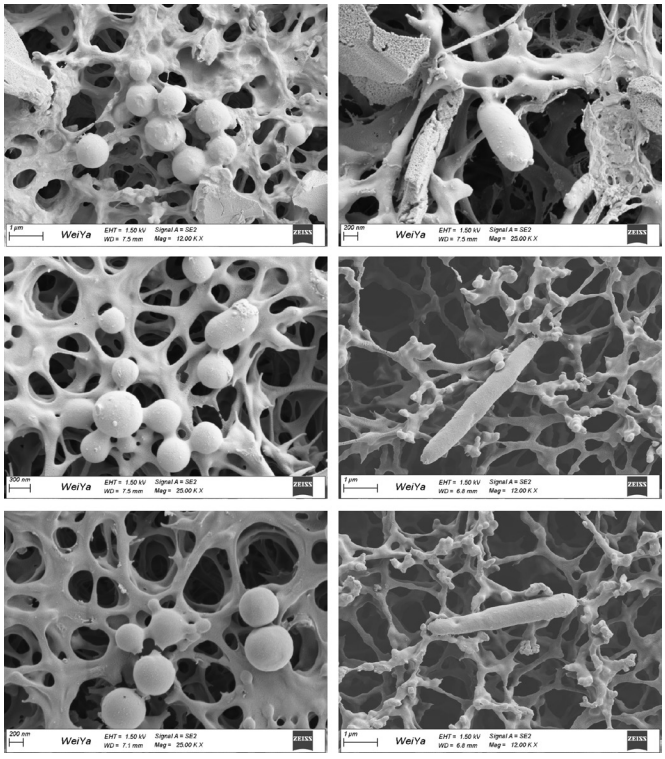


Figure 2 Scanning electron microscopy of the joint synovial fluid.

of the synovial fluid samples, both of which were observed to be exclusively elevated in faecal metagenome of patients in RAS4 from the multiomics cohort (see online supplemental table S4). In addition, *Prevotella copri* that has been reported highly correlated with RA[8 27] was also found abundant in most synovial fluid samples of patients in RAS4. We then randomly selected six synovial fluid samples per RA stage for bacteria isolation. Only from three synovial fluid samples of RAS4 can we separate bacteria. We then picked and sequenced three single colonies per synovial fluid sample. Five of the nine colonies were identified as *Clostridium sporogenes* strain, and three were identified as *Enterococcus gallinarum* strain, and one was identified as *Citrobacter freundii* strain (see online supplemental table S8). Interestingly, *Enterococcus gallinarum* and *Citrobacter freundii* could also be detected in faecal metagenomic data of 18% of patients with RA. We subsequently observed the corresponding synovial fluid samples using scanning electron microscopy, and found substances shaped like bacteria in rod-like or spherical forms (figure 2). Taken together, this multifaceted investigation has provided unprecedented evidence to support the existence of microbial invasion of the joints in the fourth stage of RA.

Stage-specific metabolomic profiles

We then introduced metabolomic data, and performed a random forest algorithm on 232 microbiome species, 6224 KOs and 277 metabolites to test their diagnostic potential for each stage of RA and OA (figure 3A–E). Metabolites exhibited the best area under the receiver operating characteristic curve (AUROC) in discriminating samples of four RA stages or OA from healthy samples, with AUROC ranging from 0.974 to 0.998. Other characteristics at the species and KO levels exhibited weaker discriminant ability, with AUROC ranging from 0.760 to 0.838 and from 0.799 to 0.852, respectively. The most prominent changes in metabolites were the significant increase of DL-lactate and gly-glu in RAS1 ($p=2.15\times10^{-6}$, $p=4.70\times10^{-4}$), the decrease of

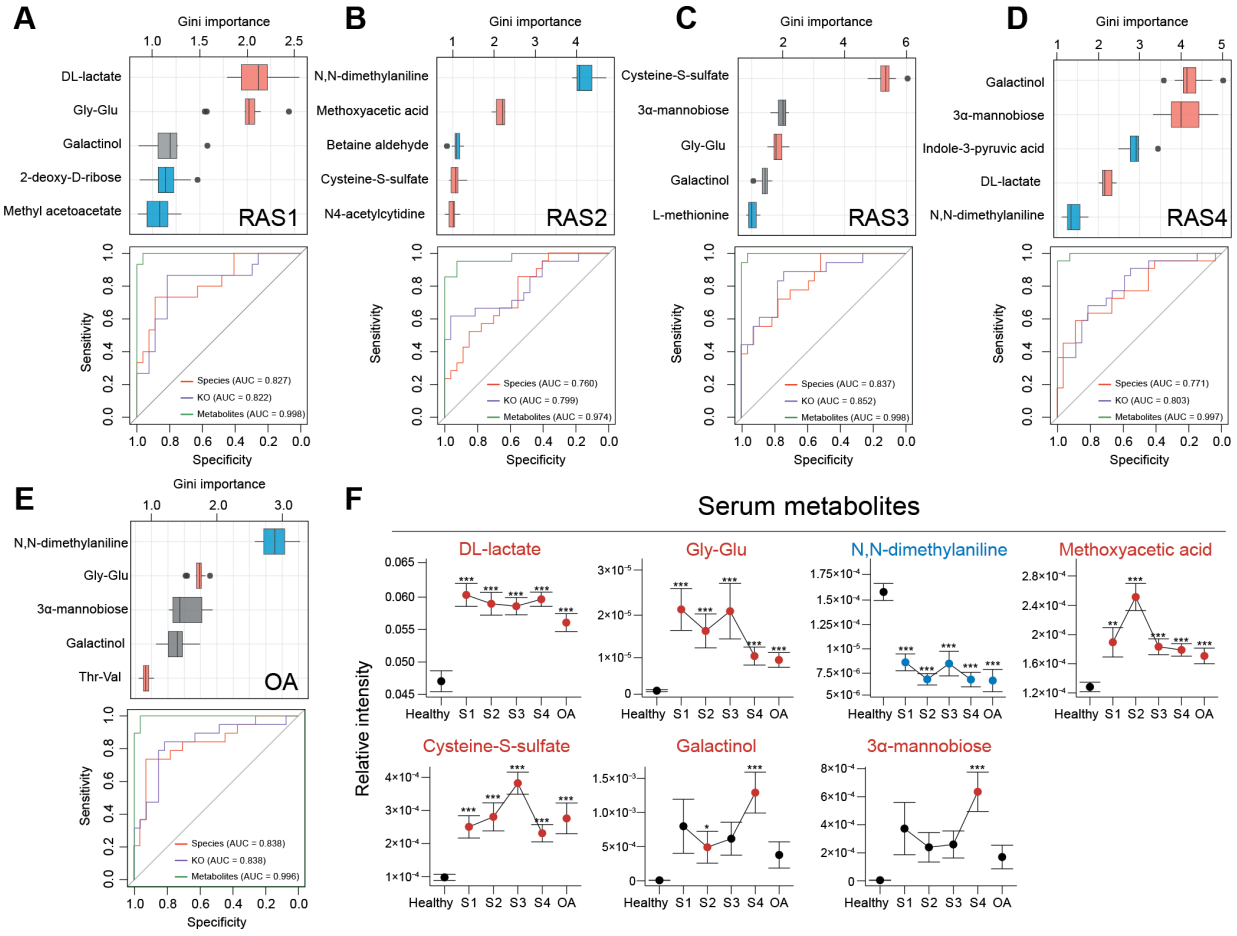


Figure 3 Multiomics diagnostic potential for the RA stage. A random forest algorithm was performed on 6224 KOs, 232 microbial species and 277 plasma metabolites in RAS1 (A), RAS2 (B), RAS3 (C), RAS4 (D) and OA (E). The Gini importance of the top five most discriminant metabolites are displayed. Boxes represent the IQR between the first and third quartiles and the line inside represents the median. Whiskers denote the lowest and highest values within the 1.5×IQR from the first and third quartiles, respectively. Boxes are marked in a specific colour to show the significant elevation (p<0.05, red, Mann-Whitney-Wilcoxon test) or depletion (p<0.05, blue, Mann-Whitney-Wilcoxon test) of the features in each of the arthritis stages compared with the healthy group. The ROC curves of the random forest model using microbial species, KOs, or metabolites were plotted, with AUC calculated by 10 randomised 10-fold cross-validation. The colour of the curve represents the category of the used features. (F) The dot plots show stage-specific abundance or concentration (mean±SE) of plasma metabolites, which are specified in (A–E). Four RA stages are connected to display the variance. Dots are coloured differently if the features are significantly elevated (red) or significantly depleted (blue), as compared with those of the healthy group. *p<0.05, **p<0.01, ***p<0.005; Mann-Whitney-Wilcoxon test. AUC, area under curve; KEGG, Kyoto Encyclopaedia of Genes and Genomes; KO, KEGG ortholog; OA, osteoarthritis; RA, rheumatoid arthritis; ROC, receiver operating characteristic.

N,N-dimethylaniline and the increase of methoxyacetic acid in RAS2 ($p=4.60\times10^{-8}$, $p=1.28\times10^{-8}$), the increase of cysteine-S-sulfate ($p=4.66\times10^{-12}$) in RAS3, the increase of galactinol and 3α-mannobiose in RAS4 ($p=5.71\times10^{-5}$, $p=5.68\times10^{-4}$), and the decrease of N,N-dimethylaniline and increase of gly-glu ($p=2.00\times10^{-7}$, $p=1.74\times10^{-3}$) in OA, as compared with a healthy state. The predominant metabolic disorders implicated a critical involvement in pathogenesis and a great diagnostic potential for RA stages.

Moreover, metabolic disorders could distinguish a given RA stage from not just healthy controls but also other RA stages or OA (figure 3F): Methoxyacetic acid in RAS2 ($p=1.68\times10^{-4}$) or cysteine-S-sulfate in RAS3 ($p=2.42\times10^{-4}$) or Galactinol and 3α-mannobiose in RAS4 ($p=9.37\times10^{-3}$, $p=4.89\times10^{-3}$), respectively, was higher than that in all the other RA stages and OA. Methoxyacetic acid was reported to have inhibitory effects on osteoblasts and could cause reductions in bone marrow cellularity.[28–30] Additionally, cysteine-S-sulfate was a structural analogue of glutamate, acting as an agonist of N-methyl-D-aspartate receptor (NMDA-R) whose expression and function in osteoclasts engaged in bone resorption.[31] Therefore, notable elevations of methoxyacetic acid in RAS2 might hinder osteoblasts, whereas notable elevations of cysteine-S-sulfate in RAS3 might encourage osteoclasts. The imbalance between osteoblasts and osteoclasts would promote the bone erosion that occurred clinically in the third stage of RA. Moreover, DL-lactate in OA was less than that in all RA stages ($p=0.037$), which might improve clinical differentiation of early RA from OA.

DISCUSSION

Our findings reveal dynamic shifts in gut microbiome and plasma metabolome, and their continuous roles in pathogenesis of RA across four successive stages (figure 4). Moreover, we demonstrate that microbial invasion of the joint synovial fluid happens in the fourth stage of RA.

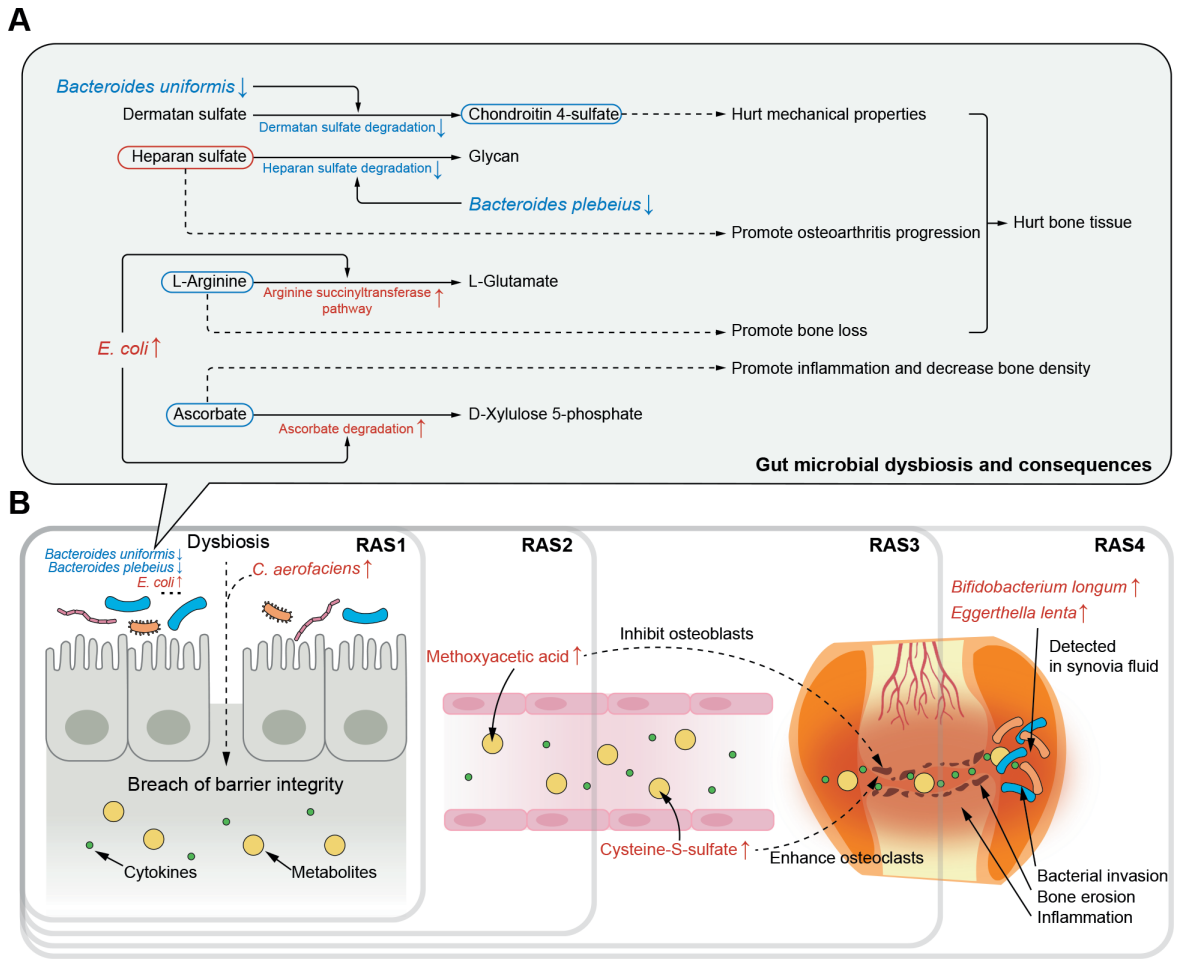


Figure 4 Potential pathogenesis across successive RA stages from multiomics perspective. (A) Potential mechanisms by which gut microbial dysbiosis play roles in RA pathogenesis through hurting hone tissue and increasing inflammation. The driving species, microbial dysfunction and related metabolites were extracted from figure 1B. The red or blue box of metabolites represents their speculated elevation or depletion according to the KEGG map. The dotted line represents the speculated effects of microbial and metabolic variation on arthritis pathogenesis. (B) The most representative effects of microbial dysbiosis and metabolic disorders on RA progression across successive stages. KEGG, Kyoto Encyclopaedia of Genes and Genomes; RA, rheumatoid arthritis.

The samples used for this study can fairly represent gut microbiome of each RA stage. Our hospital had tens of thousands of outpatients with arthritis per year and we have kept collecting samples from patients diagnosed with each stage of RA from 2017 to 2020. Considering the potential effects of clinical intervention on gut microbiome,[32] in this study, we only recruited samples of those patients who had not received treatment within 1 month and were in the active period. Therefore, the microbial dysbiosis and metabolic disorders depicted here could serve as a profound reference for future studies in each stage of RA.

Clinical microbial intervention should take into account the stage of RA. We found each RA stage had its special elevated or depleted microbes that played a role in RA pathogenesis (figure 4A). Hence, it may not be adequate for clinical guidance to generally report microbial alterations in RA without information of the stage, as many studies have done.[6 11] For instance, early inhibition of *C. aerofaciens* that was elevated exclusively in the first stage could help prevent increasing of gut permeability.[9] Additionally, inhibition of *E. coli* in the second and third stage could help maintain the content of L-arginine that acted as an inhibitor of bone lose,[24 25] as well as the content of anti-inflammatory ascorbate.[33] Moreover, certain species may need intervention across stages owing to its depletion during the whole RA progression. A cross-stages restoration of *B. uniformis* could help maintain the content of chondroitin 4-sulfate to keep mechanical properties of the articular cartilage.[12]

Moreover, metabolic alterations kept considerable throughout RA progression, in spite of which we found that certain of these metabolites need a higher priority of intervention in a specific stage. In the second stage of RA, the aberrant elevation of methoxyacetic acid might have inhibitory effects on osteoblasts and cause reductions in bone marrow cellularity[28–30] (figure 4B). The inhibited osteoblasts then drew the foreshadowing for the bone erosion that happened in the next stage. In the third stage, the considerable elevation of cysteine-S-sulfate might enhance the osteoclasts by NMDA-R interaction.[31] The imbalance between osteoblasts and osteoclasts would then promote bone erosion that happened in the third stage and persisted in the late RA stages. Thus, methoxyacetic acid may be a targeted metabolite for treatment to patients in the second stage of RA and serve as a precaution against the upcoming third stage.

Our findings suggested that bacterial invasion of joint synovial fluid happened in the fourth stage of RA (figure 4B). Joint synovial fluid was generally considered sterile, and indeed, we failed to either extract enough DNA or isolate bacteria from the synovial fluid in the first three stages. However, in the fourth stage,

we succeeded to obtain bacterial 16S reads, isolate bacteria and observe substances shaped like bacteria in rod-like or spherical forms under scanning electron microscopy. Moreover, in the multiomics cohort, we found two faecal microbes elevated exclusively in the fourth stage of RA, *E. lenta* and *B. longum*, and their existence in the joint synovial fluid was validated by the other cohort. It might due to the buildup of the continuous damages in gut barrier and microbes and microbial metabolites would then be transferred to the joints via blood.[6] Hence, for patients in the fourth stage of RA, in addition to routine medical therapies, specific treatments to the microbes in the joint synovial fluid may ameliorate the joint micro-environment to decrease synovial inflammation and inhibit potential bacterial effects.

This study also has limitations and prospects. First, a long-term follow-up investigation on a single individual throughout his/her RA development may reinforce the conclusions of this study. Second, it remains unclear how bacterial genetic materials are transferred from intestine to joint. It might be realised by bacteria transmission through blood or by means of extracellular vesicles or both. Third, the proposed links between microbial dysbiosis/metabolic disorders and RA can serve as a guidance for future experiments on RA pathogenesis. Lastly, additional researches into the synovial fluid microbiome and metabolome have the potential to reveal more sophisticated mechanisms underlying RA pathogenesis.

In conclusion, this study demonstrates microbial and metabolic roles in RA pathogenesis across four successive stages. A stage-specific intervention of microbial dysbiosis and metabolic disorders is warranted for prognosis and prevention of RA.

Author affiliations
[1]First Affiliated Hospital of Shandong First Medical University, Institute of Medical Genomics, Biomedical Sciences College & Shandong Medicinal Biotechnology Centre, NHC Key Laboratory of Biotechnology Drugs (Shandong Academy of Medical Sciences), Key Lab for Rare & Uncommon Diseases of Shandong Province, Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan, 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, Hubei, China
[3]Microbiome-X, National Institute of Health Data Science of China & Institute for Medical Dataology, Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, Shandong, China

Correction notice This article has been corrected since it was first published. The open access licence has been updated to CC BY. 25th May 2023.

Contributors MC, YaZ, LZ, KN and JH designed the study, reviewed, and verified the data. MC, YaZ, YC, CZ, YuZ, SL, GC and ML collected samples and conducted experiments. MC, YaZ and KN conducted data analysis and produced the figures and tables. MC, YaZ, KN and JH wrote the manuscript. All authors revised the manuscript. MC, YaZ, LZ, KN and JH supervised the study. MC and YaZ are joint first authors. LZ, KN, and JH are joint senior authors. All authors approved the final version of the article. JH accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish.

Funding This work was partially funded by the National Natural Science Foundation of China (grant numbers 31871334, 82003766, 32071465, and 31671374), the Academic Promotion Programme of Shandong First Medical University (grant number 2019LJ001) and the Key Research and Development project of Shandong Province (grant number 2021ZDSYS27).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the Ethics Committee of The First Affiliated Hospital of Shandong First Medical University (NO.2017-02 and NO.2020-011). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Whole-genome shotgun sequencing data are available in the Genome Sequence Archive (GSA) section of the National Genomics Data Center (project accession number CRA004348). 16S rRNA gene sequencing data are available in the Genome Sequence Archive (GSA) section of the National Genomics Data Center (project accession number CRA005811). Plasma metabolomic data are available in the MetaboLights (project accession number MTBLS5297).

ORCID iDs
Mingyue Cheng http://orcid.org/0000-0003-1243-5039
Kang Ning http://orcid.org/0000-0003-3325-5387
Jinxiang Han http://orcid.org/0000-0002-2507-9611

REFERENCES

1 Almutairi K, Nossent J, Preen D, *et al*. The global prevalence of rheumatoid arthritis: a meta-analysis based on a systematic review. *Rheumatol Int* 2021;41:863–77.
2 Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. *Lancet* 2016;388:2023–38.
3 Aletaha D, Neogi T, Silman AJ, *et al*. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League against rheumatism collaborative initiative. *Arthritis & Rheumatism* 2010;62:2569–81.
4 Steinbrocker O, Traeger CH, Batterman RC. Therapeutic criteria in rheumatoid arthritis. *J Am Med Assoc* 1949;140:659–62.
5 Barhum L. What does rheumatoid arthritis progression look like? 2021. Available: https://www.verywellhealth.com/rheumatoid-arthritis-stages-of-progression-4768891
6 Zaiss MM, Joyce Wu H-J, Mauro D, *et al*. The gut-joint axis in rheumatoid arthritis. *Nat Rev Rheumatol* 2021;17:224–37.
7 Zhao Y, Cheng M, Zou L, *et al*. Hidden link in gut-joint axis: gut microbes promote rheumatoid arthritis at early stage by enhancing ascorbate degradation. *Gut* 2022;71:1041–3.
8 Pianta A, Arvikar S, Strle K, *et al*. Evidence of the immune relevance of Prevotella copri, a gut microbe, in patients with rheumatoid arthritis. *Arthritis Rheumatol* 2017;69:964–75.
9 Chen J, Wright K, Davis JM, *et al*. An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med* 2016;8:43.
10 Smith PM, Howitt MR, Panikov N, *et al*. The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science* 2013;341:569–73.
11 Maeda Y, Takeda K. Role of gut microbiota in rheumatoid arthritis. *J Clin Med* 2017;6:60.
12 Henrotin Y, Mathy M, Sanchez C, *et al*. Chondroitin sulfate in the treatment of osteoarthritis: from in vitro studies to clinical recommendations. *Ther Adv Musculoskelet Dis* 2010;2:335–48.
13 Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 2011;27:2957–63.
14 de la Cuesta-Zuluaga J, Kelley ST, Chen Y, *et al*. Age- and sex-dependent patterns of gut microbial diversity in human adults. *mSystems* 2019;4:e00261–19.
15 Lugli GA, Tarracchini C, Alessandri G, *et al*. Decoding the Genomic Variability among Members of the *Bifidobacterium dentium* Species. *Microorganisms* 2020;8. doi:10.3390/microorganisms8111720. [Epub ahead of print: 03 11 2020].
16 Silvestre-Rangil J, Bagán L, Silvestre FJ, *et al*. Oral manifestations of rheumatoid arthritis. A cross-sectional study of 73 patients. *Clin Oral Investig* 2016;20:2575–80.
17 Marriott D, Stark D, Harkness J. Veillonella parvula discitis and secondary bacteremia: a rare infection complicating endoscopy and colonoscopy? *J Clin Microbiol* 2007;45:672–4.
18 Balakrishnan B, Luckey D, Taneja V. Autoimmunity-Associated gut commensals modulate gut permeability and immunity in humanized mice. *Mil Med* 2019;184:529–36.
19 Tito RY, Cypers H, Joossens M, *et al*. Brief report: Dialister as a microbial marker of disease activity in spondyloarthritis. *Arthritis Rheumatol* 2017;69:114–21.

20 Jura-Półtorak A, Komosinska-Vassev K, Kotulska A, *et al*. Alterations of plasma glycosaminoglycan profile in patients with rheumatoid arthritis in relation to disease activity. *Clin Chim Acta* 2014;433:20–7.
21 Shamdani S, Chantepie S, Flageollet C, *et al*. Heparan sulfate functions are altered in the osteoarthritic cartilage. *Arthritis Res Ther* 2020;22:283.
22 Severmann A-C, Jochmann K, Feller K, *et al*. An altered heparan sulfate structure in the articular cartilage protects against osteoarthritis. *Osteoarthritis Cartilage* 2020;28:977–87.
23 Abrams E, Sandson J. Effect of ascorbic acid on rheumatoid synovial fluid. *Ann Rheum Dis* 1964;23:295–9.
24 Abdelkarem HM, Fadda LH, El-Sayed EM, *et al*. Potential role of L-arginine and vitamin E against bone loss induced by nano-zinc oxide in rats. *J Diet Suppl* 2018;15:300–10.
25 Fiore CE, Pennisi P, Cutuli VM, *et al*. L-Arginine prevents bone loss and bone collagen breakdown in cyclosporin A-treated rats. *Eur J Pharmacol* 2000;408:323–6.
26 Manfredo Vieira S, Hiltensperger M, Kumar V, *et al*. Translocation of a gut pathobiont drives autoimmunity in mice and humans. *Science* 2018;359:1156–61.
27 Scher JU, Sczesnak A, Longman RS, *et al*. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* 2013;2:e01202.
28 Brown NA, Holt D, Webb M. The teratogenicity of methoxyacetic acid in the rat. *Toxicol Lett* 1984;22:93–100.
29 Miller RR, Carreon RE, Young JT, *et al*. Toxicity of methoxyacetic acid in rats. *Fundam Appl Toxicol* 1982;2:158–60.
30 Sparks NR. The embryotoxic effects of harm reduction tobacco products on osteoblasts developing from human embryonic stem cells. University of California, Riverside 2018; chapter 4:73–104.
31 Li P, Sundh D, Ji B, *et al*. Metabolic Alterations in Older Women With Low Bone Mineral Density Supplemented With *Lactobacillus reuteri*. *JBMR Plus* 2021;5:e10478.
32 Schwartz DJ, Langdon AE, Dantas G. Understanding the impact of antibiotic perturbation on the human microbiome. *Genome Med* 2020;12:82.
33 Carr AC, Maggini S. Vitamin C and immune function. *Nutrients* 2017;9:1211.

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.

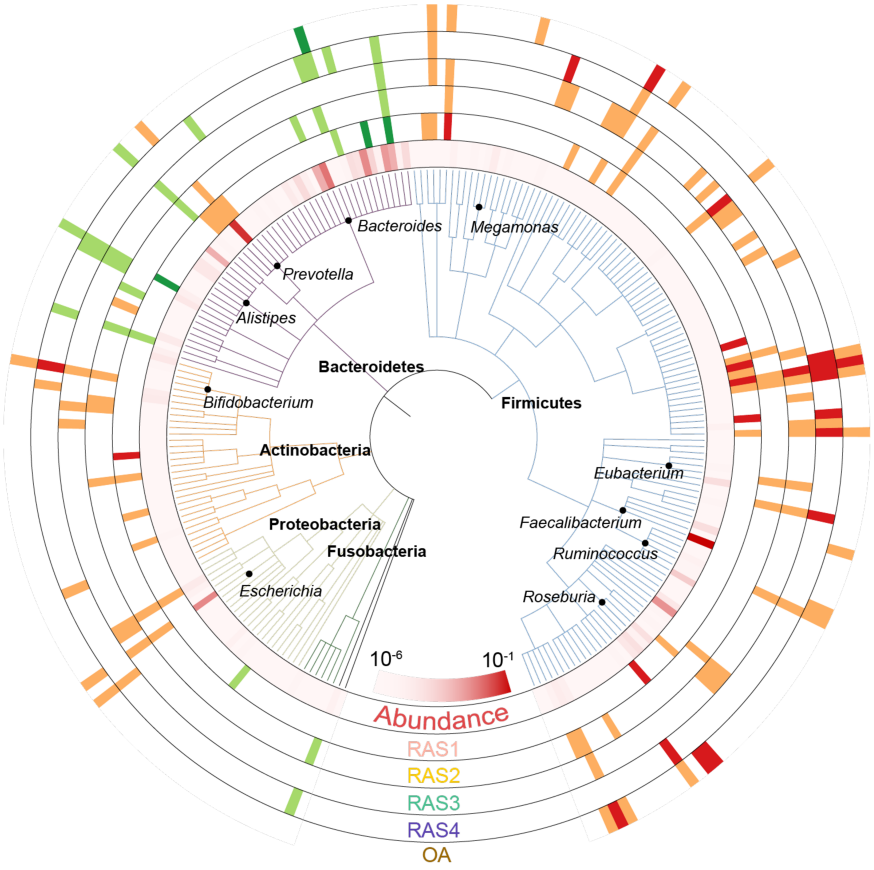
Liquid chromatography with tandem mass spectrometry (LC-MS-MS) analysis was performed using an UHPLC system (1290, Agilent Technologies) with an ACQUITY UPLC BEH Amide column (1.7μm 2.1*100mm, Waters) coupled to TripleTOF 6600 (Q-TOF, AB Sciex) & QTOF 6550 (Agilent). The mobile phase consisted of 25mM 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.

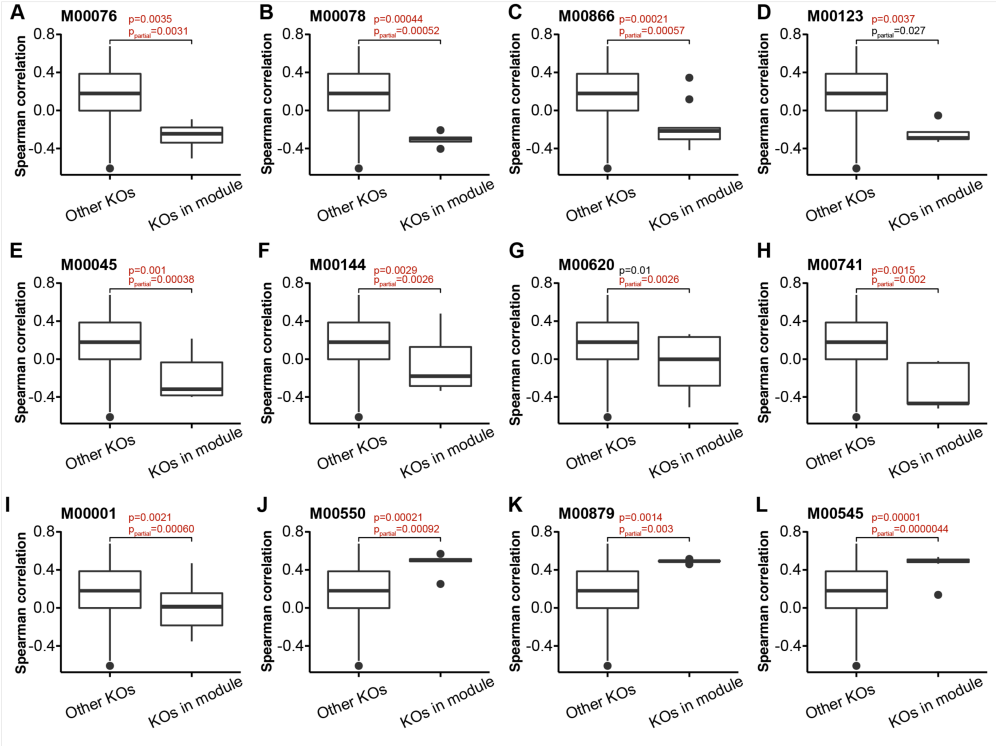
References

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.

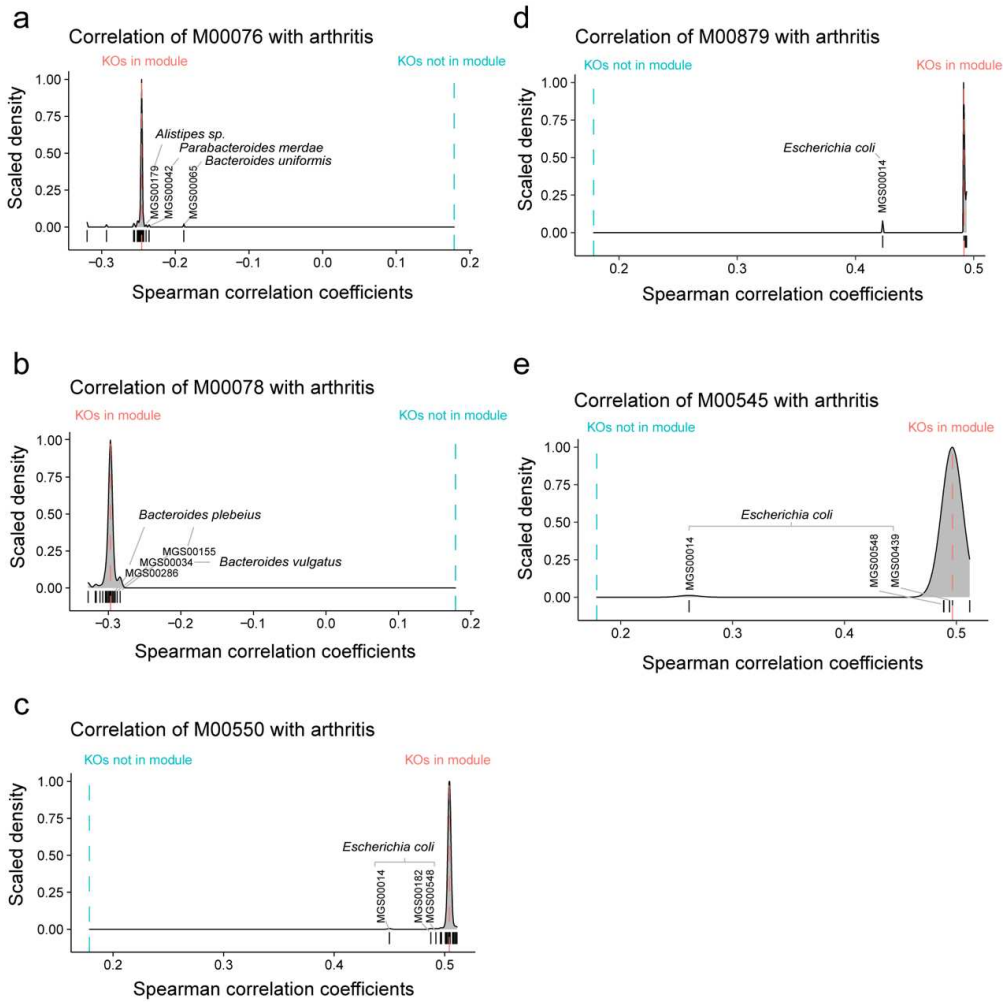
Supplementary figures



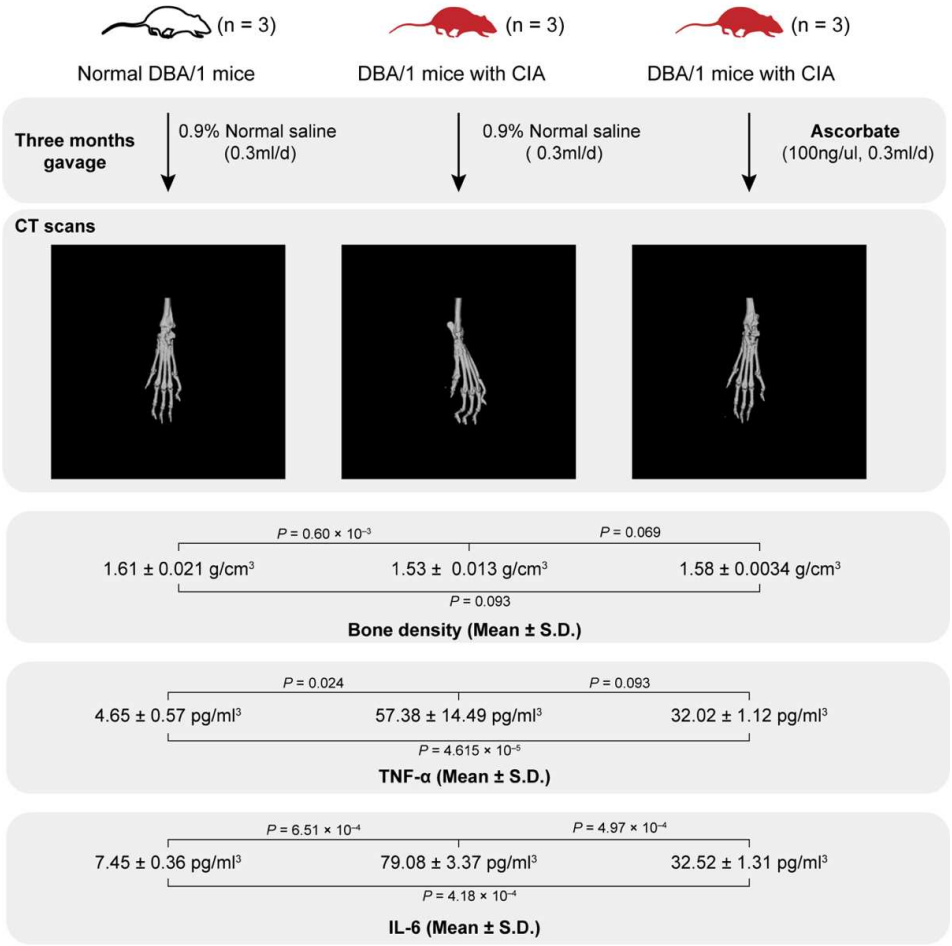
Supplementary figure 1 Stage-specific microbial taxonomic profiles. In total, 232 classified species whose relative abundance was more than 1×10^{-6} are shown in a phylogenetic tree, grouped in the phyla Fusobacteria, Proteobacteria, Actinobacteria, Bacteroidetes, and Firmicutes. The top 10 abundant genera were annotated. In the outer circles, species at each arthritis stage are marked for significant elevation ($p<0.05$, orange; $q<0.1$, red) or depletion ($p<0.05$, light green; $q<0.1$ dark green) in abundance, as compared to those of healthy individuals. The innermost circle shows species relative abundances averaged over all samples. The p values were produced using Mann-Whitney-Wilcoxon test and q values were produced using Benjamini and Hochberg corrections.



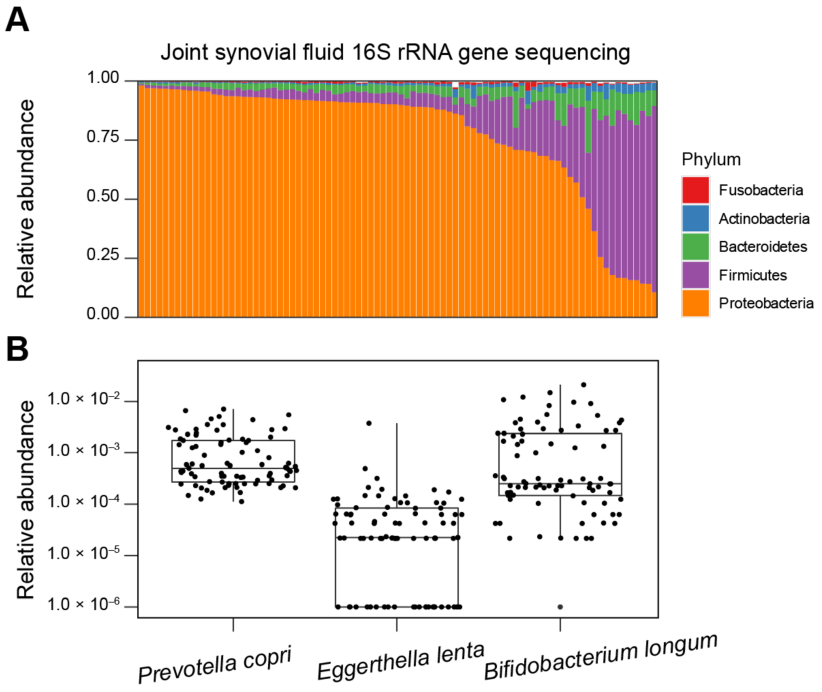
Supplementary figure 2 KEGG modules that were identified as correlated with arthritis. The correlations between KEGG modules and arthritis (Healthy=0, OA=1, RA=2) were determined by if SCC between the KOs in a given KEGG module and arthritis were different from that between all the other KOs out of the KEGG module and arthritis ($p<0.05$ and $q<0.1$, marked in red). Partial SCCs were also calculated and compared with age and gender adjusted ($p_{partial}<0.05$ and $q_{partial}<0.1$, marked in red). Boxes represent the interquartile range between first and third quartiles and the line inside represents the median. Whiskers denote the lowest and highest values within 1.5 × interquartile range from the first and third quartiles, respectively. The p and $p_{partial}$ were produced using Mann-Whitney-Wilcoxon test and q and $q_{partial}$ were produced using Benjamini and Hochberg corrections. SCC, Spearman correlation coefficient.



Supplementary figure 3 The MGS that particularly contributed to the observed linkage between functional modules and arthritis. Dashed line represents the median SCC between the arthritis (Healthy=0, OA=1, RA=2) and KOs in the specific module (red) or all the other KOs out of the module (blue). Density plot shows the median SCC between the arthritis and KOs in the specific module, when a given MGS (indicated by short vertical lines) has been excluded from the analysis. Age and gender were adjusted in SCC calculation. SCC, Spearman correlation coefficient; KO, KEGG ortholog; MGS, metagenomic species.



Supplementary figure 4 Ascorbate ameliorate collagen-induced arthritis in mice model. Three groups of mice (three mice per group) were used, including normal DBA/1 mice and two groups of DBA/1 mice with CIA. After three-month gavage, the CT scans, bone density, plasma TNF-α level, and plasma IL-6 level were examined. The statistical test is performed using t-test. IL, interleukin; TNF, tumour necrosis factor.



Supplementary figure 5 Microbial composition in the joint synovial fluid of patients in RAS4. (**A**) Relative abundances of the five most abundant phyla in the joint synovial fluid of patients in RAS4. (**B**) Relative abundances of the three microorganisms in the joint synovial fluid of patients in RAS4. The boxes represent 25th–75th percentiles, black lines indicate the median and whiskers extend to the maximum and minimum values within 1.5 × the interquartile range. RAS4, the fourth stage of rheumatoid arthritis.

Supplementary tables

Supplementary table 1 Altered microbial species in the first stage of RA.

Species	RAS1 (n=15)		HC (n=27)		p	q
	Mean	SEM	Mean	SEM		
Streptococcus sanguinis↑	8.48E-05	4.02E-05	1.52E-06	1.20E-06	5.37E-04	6.43E-02
Streptococcus mitis oralis pneumoniae↑	1.63E-04	5.73E-05	9.09E-06	4.72E-06	1.21E-03	6.43E-02
Bacteroides uniformis↓	9.98E-03	3.43E-03	7.83E-02	2.30E-02	1.36E-03	6.43E-02
Actinomyces graevenitzii↑	1.72E-05	7.22E-06	0.00E+00	0.00E+00	1.77E-03	6.43E-02
Streptococcus anginosus↑	8.26E-04	6.36E-04	0.00E+00	0.00E+00	1.77E-03	6.43E-02
Parabacteroides distasonis↓	1.18E-03	5.50E-04	4.97E-03	1.37E-03	1.95E-03	6.43E-02
Coprococcus comes↑	1.58E-03	6.26E-04	2.96E-04	8.87E-05	3.28E-03	7.92E-02
Streptococcus infantis↑†	6.18E-05	2.76E-05	5.93E-06	5.63E-06	3.43E-03	7.92E-02
Holdemania filiformis↑	3.45E-05	2.21E-05	3.23E-06	3.23E-06	3.60E-03	7.92E-02
Bacteroides ovatus↓†	1.22E-03	4.45E-04	1.60E-02	8.91E-03	4.21E-03	8.34E-02
Streptococcus cristatus↑	6.46E-06	3.15E-06	0.00E+00	0.00E+00	5.83E-03	9.62E-02
Streptococcus vestibularis↑	7.11E-05	4.24E-05	0.00E+00	0.00E+00	5.83E-03	9.62E-02
Streptococcus gordonii↑	5.69E-05	2.36E-05	1.81E-06	1.81E-06	7.49E-03	1.14E-01
Actinomyces odontolyticus↑	2.71E-05	1.75E-05	0.00E+00	0.00E+00	1.87E-02	1.95E-01
Atopobium parvulum↑†	1.58E-05	8.60E-06	0.00E+00	0.00E+00	1.87E-02	1.95E-01
Prevotella stercorea↑	3.84E-04	2.38E-04	0.00E+00	0.00E+00	1.87E-02	1.95E-01
Enterococcus faecium↑	4.24E-05	2.58E-05	0.00E+00	0.00E+00	1.87E-02	1.95E-01
Weissella confusa↑	6.79E-05	4.60E-05	0.00E+00	0.00E+00	1.87E-02	1.95E-01
Solobacterium moorei↑	6.03E-05	3.67E-05	0.00E+00	0.00E+00	1.87E-02	1.95E-01
Bacteroides cellulosilyticus↓	5.37E-04	3.89E-04	7.73E-03	4.77E-03	2.33E-02	2.16E-01
Parasutterella excrementihominis†	2.80E-04	2.62E-04	1.47E-03	6.70E-04	2.44E-02	2.16E-01
Granulicatella adiacens↑†	3.08E-05	1.94E-05	6.59E-07	6.59E-07	2.51E-02	2.16E-01
Turicibacter sanguinis↑†	8.02E-05	4.57E-05	1.11E-06	1.11E-06	2.51E-02	2.16E-01
Prevotella copri↑	1.46E-01	5.86E-02	1.04E-02	8.24E-03	3.46E-02	2.75E-01
Streptococcus parasanguinis↑	9.14E-04	5.85E-04	3.92E-05	2.56E-05	3.47E-02	2.75E-01
Dorea formicigenerans↑	8.21E-04	2.26E-04	3.85E-04	9.07E-05	3.81E-02	2.82E-01
Klebsiella pneumoniae↑	1.26E-02	1.02E-02	1.14E-03	8.56E-04	3.84E-02	2.82E-01
Coprobacter fastidiosus↓	8.14E-05	7.79E-05	2.14E-03	1.41E-03	4.10E-02	2.90E-01
Collinsella aerofaciens↑†	2.00E-03	9.48E-04	3.56E-04	1.73E-04	4.34E-02	2.96E-01

p values produced by Mann-Whitney-Wilcoxon test are displayed. q values are calculated using Benjamini and Hochberg corrections. Only results with p<0.05 are displayed here.

↑↓Species that are increased/decreased in RAS1 compared to HC.

†Species that are altered exclusively in RAS1, with p<0.05 only observed in comparisons between RAS1 and HC.

RAS1, individuals in the first stage of rheumatoid arthritis; HC, healthy controls; SEM, standard error of the mean.

Supplementary table 2 Altered microbial species in the second stage of RA.

Species	RAS2 (n=21)		HC (n=27)		p	q
	Mean	SEM	Mean	SEM		
Escherichia coli↑	5.59E-02	2.37E-02	4.74E-03	1.97E-03	1.27E-03	1.31E-01
Streptococcus anginosus↑	1.45E-04	7.82E-05	0.00E+00	0.00E+00	1.46E-03	1.31E-01
Clostridium nexile↑	8.02E-04	2.26E-04	5.64E-05	3.86E-05	1.89E-03	1.31E-01
Clostridium hathewayi↑	6.77E-04	5.60E-04	9.63E-07	6.71E-07	2.53E-03	1.31E-01
Bacteroides uniformis↓	1.81E-02	5.88E-03	7.83E-02	2.30E-02	5.35E-03	2.13E-01
Klebsiella pneumoniae↑	1.99E-02	1.80E-02	1.14E-03	8.56E-04	6.30E-03	2.13E-01
Bifidobacterium dentium↑	4.91E-04	2.37E-04	3.70E-05	3.70E-05	7.16E-03	2.13E-01
Blautia gnavus↑	5.71E-03	2.24E-03	1.22E-03	7.29E-04	8.76E-03	2.28E-01
Bacteroides salyersiae↓	5.63E-05	4.28E-05	2.21E-03	1.08E-03	1.23E-02	2.80E-01
Prevotella copri↑	8.60E-02	3.53E-02	1.04E-02	8.24E-03	1.35E-02	2.80E-01
Holdemania filiformis↑	1.15E-04	6.21E-05	3.23E-06	3.23E-06	1.62E-02	2.83E-01
Blautia torques↑	5.16E-03	1.48E-03	1.46E-03	3.78E-04	1.63E-02	2.83E-01
Clostridium clostridioforme↑†	3.94E-04	2.77E-04	1.12E-05	1.12E-05	1.95E-02	2.83E-01
Scardovia wiggsiae↑	1.35E-05	1.13E-05	0.00E+00	0.00E+00	2.04E-02	2.83E-01
Weissella confusa↑	3.25E-04	3.19E-04	0.00E+00	0.00E+00	2.04E-02	2.83E-01
Coprobacter fastidiosus↓	1.25E-04	7.51E-05	2.14E-03	1.41E-03	2.38E-02	3.09E-01
Bacteroides coprocola↓†	8.26E-03	4.20E-03	7.47E-02	2.01E-02	3.57E-02	3.67E-01
Streptococcus parasanguinis↑	4.92E-04	3.40E-04	3.92E-05	2.56E-05	3.63E-02	3.67E-01
Eubacterium ramulus↑	9.24E-04	5.07E-04	7.81E-05	2.84E-05	3.64E-02	3.67E-01
Parabacteroides johnsonii↓†	2.31E-06	2.31E-06	4.67E-04	2.35E-04	4.37E-02	3.67E-01
Paraprevotella clara↓	1.16E-03	6.88E-04	3.77E-03	1.72E-03	4.56E-02	3.67E-01
Akkermansia muciniphila↓	1.96E-03	1.66E-03	3.43E-03	1.33E-03	4.61E-02	3.67E-01
Actinomyces odontolyticus↑	1.03E-05	9.64E-06	0.00E+00	0.00E+00	4.76E-02	3.67E-01
Scardovia inopinata↑†	5.68E-06	3.95E-06	0.00E+00	0.00E+00	4.76E-02	3.67E-01
Porphyromonas somerae↑†	1.26E-05	9.64E-06	0.00E+00	0.00E+00	4.76E-02	3.67E-01
Lactobacillus salivarius↑	1.29E-04	1.12E-04	0.00E+00	0.00E+00	4.76E-02	3.67E-01
Streptococcus cristatus↑	1.66E-05	1.28E-05	0.00E+00	0.00E+00	4.76E-02	3.67E-01

p values produced by Mann-Whitney-Wilcoxon test are displayed. q values are calculated using Benjamini and Hochberg corrections. Only results with $p<0.05$ are displayed here.

↑↓Species that are increased/decreased in RAS2 compared to HC.

†Species that are altered exclusively in RAS2, with $p<0.05$ only observed in comparisons between RAS2 and HC.

RAS2, individuals in the second stage of rheumatoid arthritis; HC, healthy controls; SEM, standard error of the mean.

Supplementary table 3 Altered microbial species in the third stage of RA.

Species	RAS3 (n=18)		HC (n=27)		p	q
	Mean	SEM	Mean	SEM		
Lactobacillus salivarius↑	1.14E-03	8.44E-04	0.00E+00	0.00E+00	5.56E-04	5.45E-02
Streptococcus anginosus↑	1.81E-04	7.46E-05	0.00E+00	0.00E+00	5.56E-04	5.45E-02
Streptococcus parasanguinis↑	6.52E-04	2.07E-04	3.92E-05	2.56E-05	2.04E-03	1.33E-01
Bifidobacterium dentium↑	4.16E-04	2.47E-04	3.70E-05	3.70E-05	3.70E-03	1.39E-01
Parabacteroides distasonis	2.07E-03	9.12E-04	4.97E-03	1.37E-03	4.85E-03	1.39E-01
Dorea formicigenerans↑	1.74E-03	4.57E-04	3.85E-04	9.07E-05	4.85E-03	1.39E-01
Veillonella atypica↑	4.93E-04	2.15E-04	6.54E-05	3.46E-05	5.57E-03	1.39E-01
Clostridium hathewayi↑	3.46E-04	2.89E-04	9.63E-07	6.71E-07	5.92E-03	1.39E-01
Bacteroides uniformis↓	1.65E-02	6.81E-03	7.83E-02	2.30E-02	7.20E-03	1.39E-01
Escherichia coli↑	6.34E-02	2.08E-02	4.74E-03	1.97E-03	9.70E-03	1.39E-01
Klebsiella pneumoniae↑	7.06E-03	3.31E-03	1.14E-03	8.56E-04	1.09E-02	1.39E-01
Eubacterium ramulus↑	4.67E-04	1.48E-04	7.81E-05	2.84E-05	1.17E-02	1.39E-01
Actinomyces odontolyticus↑	2.69E-05	2.31E-05	0.00E+00	0.00E+00	1.21E-02	1.39E-01
Parascardovia denticolens↑†	1.25E-05	8.26E-06	0.00E+00	0.00E+00	1.21E-02	1.39E-01
Scardovia wiggsiae↑	6.48E-05	4.22E-05	0.00E+00	0.00E+00	1.21E-02	1.39E-01
Lactobacillus gasseri↑†	1.13E-04	6.15E-05	0.00E+00	0.00E+00	1.21E-02	1.39E-01
Weissella confusa↑	3.35E-05	2.35E-05	0.00E+00	0.00E+00	1.21E-02	1.39E-01
Ruminococcus 39BFAA↑	3.86E-03	1.36E-03	3.62E-04	2.07E-04	1.62E-02	1.73E-01
Blautia torques↑	4.26E-03	1.12E-03	1.46E-03	3.78E-04	1.70E-02	1.73E-01
Streptococcus gordonii↑	6.55E-05	2.88E-05	1.81E-06	1.81E-06	1.77E-02	1.73E-01
Clostridium nexile↑	1.06E-03	9.42E-04	5.64E-05	3.86E-05	1.88E-02	1.75E-01
Streptococcus pneumoniae↑	4.71E-05	1.69E-05	9.09E-06	4.72E-06	2.11E-02	1.86E-01
Holdemania filiformis↑	3.41E-05	1.64E-05	3.23E-06	3.23E-06	2.18E-02	1.86E-01
Veillonella parvula↑†	1.49E-03	6.11E-04	2.66E-04	8.53E-05	2.73E-02	2.17E-01
Enterococcus faecium↑	1.01E-03	9.96E-04	0.00E+00	0.00E+00	3.21E-02	2.17E-01
Lactobacillus crispatus↑†	4.82E-05	2.97E-05	0.00E+00	0.00E+00	3.21E-02	2.17E-01
Lactobacillus oris↑†	3.00E-05	1.84E-05	0.00E+00	0.00E+00	3.21E-02	2.17E-01
Streptococcus mutans↑†	3.72E-05	2.02E-05	0.00E+00	0.00E+00	3.21E-02	2.17E-01
Solobacterium moorei↑	1.57E-05	1.12E-05	0.00E+00	0.00E+00	3.21E-02	2.17E-01
Streptococcus salivarius↑	3.50E-03	1.51E-03	4.31E-04	1.64E-04	3.47E-02	2.23E-01
Paraprevotella clara↓	4.08E-04	1.75E-04	3.77E-03	1.72E-03	3.66E-02	2.23E-01
Dorea longicatena↑†	3.21E-03	1.61E-03	9.42E-04	3.77E-04	3.74E-02	2.23E-01
Enterobacter cloacae↑†	4.51E-03	3.67E-03	4.13E-03	3.98E-03	3.75E-02	2.23E-01
Parabacteroides merdae↓	3.07E-03	1.02E-03	7.83E-03	1.79E-03	4.67E-02	2.69E-01

p values produced by Mann-Whitney-Wilcoxon test are displayed. q values are calculated using Benjamini and Hochberg corrections. Only results with p<0.05 are displayed here.

↑↓Species that are increased/decreased in RAS3 compared to HC.

†Species that are altered exclusively in RAS3, with p<0.05 only observed in comparisons between RAS3 and HC.

RAS3, individuals in the third stage of rheumatoid arthritis; HC, healthy controls; SEM, standard error of the mean.

Supplementary table 4 Altered microbial species in the fourth stage of RA.

Species	RAS4 (n=22)		HC (n=27)		p	q
	Mean	SEM	Mean	SEM		
Actinomyces odontolyticus↑	2.46E-05	1.10E-05	0.00E+00	0.00E+00	7.88E-04	6.34E-02
Streptococcus anginosus↑	1.86E-04	1.36E-04	0.00E+00	0.00E+00	7.88E-04	6.34E-02
Bifidobacterium dentium↑	5.48E-04	3.90E-04	3.70E-05	3.70E-05	9.15E-04	6.34E-02
Veillonella atypica↑	3.71E-04	1.10E-04	6.54E-05	3.46E-05	1.25E-03	6.51E-02
Streptococcus sanguinis↑	4.20E-05	1.47E-05	1.52E-06	1.20E-06	1.60E-03	6.52E-02
Streptococcus vestibularis↑	6.59E-04	6.19E-04	0.00E+00	0.00E+00	1.90E-03	6.52E-02
Clostridium citroniae↑	8.92E-05	4.80E-05	4.85E-06	3.50E-06	2.29E-03	6.52E-02
Streptococcus parasanguinis↑	6.04E-04	3.99E-04	3.92E-05	2.56E-05	2.51E-03	6.52E-02
Clostridium hathewayi↑	2.00E-03	1.92E-03	9.63E-07	6.71E-07	2.96E-03	6.85E-02
Streptococcus gordonii↑	4.55E-05	2.40E-05	1.81E-06	1.81E-06	3.37E-03	7.01E-02
Bacteroides uniformis↓	3.06E-02	1.54E-02	7.83E-02	2.30E-02	5.71E-03	1.08E-01
Bacteroides cellulosilyticus↓	3.11E-04	2.67E-04	7.73E-03	4.77E-03	6.24E-03	1.08E-01
Parabacteroides distasonis↓	3.27E-03	1.94E-03	4.97E-03	1.37E-03	1.04E-02	1.54E-01
Enterococcus faecium↑	3.35E-03	2.82E-03	0.00E+00	0.00E+00	1.04E-02	1.54E-01
Akkermansia muciniphila↓	2.13E-03	1.66E-03	3.43E-03	1.33E-03	1.67E-02	1.96E-01
Ruminococcus 39BFAA↑	3.22E-03	1.13E-03	3.62E-04	2.07E-04	1.73E-02	1.96E-01
Eggerthella lenta↑†	1.11E-04	7.23E-05	1.56E-06	1.56E-06	1.82E-02	1.96E-01
Bacteroides salyersiae	7.13E-04	6.58E-04	2.21E-03	1.08E-03	1.93E-02	1.96E-01
Bacteroides faecis†	1.57E-04	1.02E-04	3.79E-03	1.47E-03	1.95E-02	1.96E-01
Streptococcus salivarius↑	7.74E-03	6.31E-03	4.31E-04	1.64E-04	2.15E-02	1.96E-01
Coprobacter fastidiosus↓	1.72E-04	1.23E-04	2.14E-03	1.41E-03	2.24E-02	1.96E-01
Bifidobacterium longum↑†	1.05E-02	5.10E-03	1.01E-03	3.23E-04	2.25E-02	1.96E-01
Actinomyces graevenitzii↑	1.70E-05	9.92E-06	0.00E+00	0.00E+00	2.35E-02	1.96E-01
Lactococcus garvieae↑†	7.46E-05	5.37E-05	0.00E+00	0.00E+00	2.35E-02	1.96E-01
Solobacterium moorei↑	1.88E-05	1.22E-05	0.00E+00	0.00E+00	2.35E-02	1.96E-01
Bacteroides caccae†	3.06E-03	1.03E-03	1.02E-02	3.41E-03	3.30E-02	2.64E-01
Parabacteroides merdae↓	4.55E-03	1.99E-03	7.83E-03	1.79E-03	3.85E-02	2.89E-01
Citrobacter freundii↑	1.70E-03	1.63E-03	1.74E-06	1.74E-06	3.89E-02	2.89E-01

p values produced by Mann-Whitney-Wilcoxon test are displayed. q values are calculated using Benjamini and Hochberg corrections. Only results with $p<0.05$ are displayed here.

↑↓Species that are increased/decreased in RAS4 compared to HC.

†Species that are altered exclusively in RAS4, with $p<0.05$ only observed in comparisons between RAS4 and HC.

RAS4, individuals in the fourth stage of rheumatoid arthritis; HC, healthy controls; SEM, standard error of the mean.

Supplementary table 5 Altered microbial species in OA.

Species	OA (n=19)		HC (n=27)		p	q
	Mean	SEM	Mean	SEM		
Coprococcus catus↑†	2.17E-03	7.19E-04	3.24E-04	1.43E-04	6.27E-05	1.22E-02
Blautia torques↑	8.49E-03	2.79E-03	1.46E-03	3.78E-04	3.72E-04	3.05E-02
Bacteroides cellulosilyticus↓	6.27E-05	6.01E-05	7.73E-03	4.77E-03	4.72E-04	3.05E-02
Enterococcus faecium↑	3.80E-04	1.81E-04	0.00E+00	0.00E+00	7.94E-04	3.85E-02
Coprococcus comes↑	1.39E-03	3.27E-04	2.96E-04	8.87E-05	1.85E-03	6.96E-02
Streptococcus anginosus↑	1.29E-04	5.94E-05	0.00E+00	0.00E+00	2.15E-03	6.96E-02
Holdemania filiformis↑	7.62E-05	3.72E-05	3.23E-06	3.23E-06	4.10E-03	1.14E-01
Blautia obeum↑†	2.91E-03	7.54E-04	1.26E-03	4.87E-04	4.77E-03	1.16E-01
Ruminococcus 39BFAA↑	4.97E-03	1.90E-03	3.62E-04	2.07E-04	8.68E-03	1.75E-01
Parabacteroides distasonis↓	3.03E-03	1.66E-03	4.97E-03	1.37E-03	9.01E-03	1.75E-01
Actinomyces odontolyticus↑	1.55E-05	8.27E-06	0.00E+00	0.00E+00	1.46E-02	1.90E-01
Lactobacillus salivarius↑	8.90E-04	4.79E-04	0.00E+00	0.00E+00	1.46E-02	1.90E-01
Streptococcus vestibularis↑	1.62E-03	1.58E-03	0.00E+00	0.00E+00	1.46E-02	1.90E-01
Solobacterium moorei↑	5.75E-06	3.90E-06	0.00E+00	0.00E+00	1.46E-02	1.90E-01
Bifidobacterium dentium↑	5.74E-05	2.29E-05	3.70E-05	3.70E-05	1.52E-02	1.90E-01
Clostridium citroniae↑	3.42E-05	1.68E-05	4.85E-06	3.50E-06	1.57E-02	1.90E-01
Prevotella copri↑	9.27E-02	3.36E-02	1.04E-02	8.24E-03	1.84E-02	2.10E-01
Streptococcus gordonii↑	6.14E-05	3.05E-05	1.81E-06	1.81E-06	2.22E-02	2.27E-01
Citrobacter freundii↑	4.04E-04	2.63E-04	1.74E-06	1.74E-06	2.22E-02	2.27E-01
Paraprevotella clara↓	6.19E-04	3.20E-04	3.77E-03	1.72E-03	2.66E-02	2.58E-01
Streptococcus parasanguinis↑	9.49E-04	4.41E-04	3.92E-05	2.56E-05	2.81E-02	2.60E-01
Blautia gnavus↑	4.17E-03	2.86E-03	1.22E-03	7.29E-04	3.24E-02	2.86E-01
Escherichia coli↑	8.35E-02	4.00E-02	4.74E-03	1.97E-03	3.58E-02	2.88E-01
Weissella confusa↑	1.07E-04	8.06E-05	0.00E+00	0.00E+00	3.71E-02	2.88E-01
Clostridium perfringens↑†	1.18E-04	7.26E-05	0.00E+00	0.00E+00	3.71E-02	2.88E-01
Dialister invisus↑†	2.22E-03	1.31E-03	8.17E-04	7.28E-04	4.10E-02	3.06E-01

p values produced by Mann-Whitney-Wilcoxon test are displayed. q values are calculated using Benjamini and Hochberg corrections. Only results with $p<0.05$ are displayed here.

↑↓Species that are increased/decreased in OA compared to HC.

†Species that are altered exclusively in OA, with $p<0.05$ only observed in comparisons between the OA and HC.

OA, individuals in the first stage of osteoarthritis; HC, healthy controls; SEM, standard error of the mean.

Supplementary table 6 Correlations of KEGG modules with rheumatoid factor and plasma cytokines.

	M00550		**M00879**		**M00545**	
	p ($p_{partial}$)	**q ($q_{partial}$)**	**p ($p_{partial}$)**	**q ($q_{partial}$)**	**p ($p_{partial}$)**	**q ($q_{partial}$)**
RF, IU/mL	0.217	0.547	0.00135**	0.0472*	0.00013***	0.0118*
	(0.480)	(0.773)	(0.00601**)	(0.0516*)	(0.00319**)	(0.0470*)
IL-1β, pg/mL	0.000544***	0.0370*	0.0126*	0.196	0.0000287***	0.00391*
	(0.00278**)	(0.0432*)	(0.0252*)	(0.294)	(0.000335***)	(0.0325*)
IL-4, pg/mL	0.00420**	0.0967*	0.243	0.55	0.00177**	0.0607*
	(0.0640)	(0.470)	(0.925)	(0.961)	(0.199)	(0.669)
IL-8, pg/mL	0.000745***	0.0707*	0.0201*	0.297	0.000246***	0.0669*
	(0.00400**)	(0.0851*)	(0.0337*)	(0.456)	(0.00347**)	(0.0751*)
IFN-γ, pg/mL	0.00313**	0.106	0.076	0.519	0.000392***	0.0356*
	(0.0366*)	(0.444)	(0.289)	(0.678)	(0.0325*)	(0.0946*)
IL-10, pg/mL	0.00155**	0.105	0.0162*	0.246	0.000041***	0.00557*
	(0.00642**)	(0.218)	(0.0330*)	(0.364)	(0.000373***)	(0.0464*)
IL-12p70, pg/mL	0.00347**	0.0931*	0.172	0.458	0.00261**	0.0888*
	(0.0490*)	(0.478)	(0.609)	(0.837)	(0.176)	(0.543)
IL-13, pg/mL	0.0858	0.421	0.0190*	0.215	0.0000163***	0.00441*
	(0.262)	(0.560)	(0.0139*)	(0.165)	(0.0000639***)	(0.0110*)
IL-17, pg/mL	0.104	0.421	0.829	0.915	0.777	0.895
	(0.600)	(0.820)	(0.493)	(0.810)	(0.111)	(0.451)
IL-2, pg/mL	0.000937***	0.0637*	0.0146*	0.233	0.000074***	0.0101*
	(0.00677**)	(0.0501*)	(0.0264*)	(0.362)	(0.00133**)	(0.0720*)
IL-6, pg/mL	0.00112**	0.0611*	0.125	0.467	0.00612**	0.151
	(0.0103*)	(0.0716*)	(0.358)	(0.738)	(0.204)	(0.587)
TNF-α, pg/mL	0.000659***	0.0504*	0.0318*	0.360	0.0000697***	0.00947*
	(0.00600**)	(0.0632*)	(0.0855)	(0.516)	(0.00182**)	(0.0876*)

*p<0.05 ($p_{partial}$<0.05), **p<0.01 ($p_{partial}$<0.01), ***p<0.001 ($p_{partial}$<0.001), Mann-Whitney-Wilcoxon test; *q<0.1 ($q_{partial}$<0.1), Benjamini and Hochberg corrections. Comparisons of SCCs were used for calculating p and q, while comparisons of partial SCCs with age and gender adjusted were used for calculating $p_{partial}$ and $q_{partial}$.

KEGG, Kyoto Encyclopedia of Genes and Genomes; RF, rheumatoid factor; IL, interleukin; TNF, tumour necrosis factor; IFN, Interferon; M00550, ascorbate degradation; M00879, arginine succinyltransferase pathway; M00545, trans-cinnamate degradation.

Supplementary table 7 The overlap of genera from gut metagenomic data and synovial fluid 16S data.

	Gut metagenome		Synovial fluid 16S data	
	Mean	**S.E.M.**	**Mean**	**S.E.M.**
Abiotrophia	6.45E-07	5.62E-07	5.46E-05	5.61E-06
Acidaminococcus	3.76E-05	2.32E-05	3.23E-06	8.82E-07
Acinetobacter	2.63E-06	2.54E-06	0.001319	0.00025738
Actinobacillus	1.22E-06	7.32E-07	0.00039202	7.36E-05
Actinomyces	4.10E-05	1.70E-05	0.00197494	0.0001988
Adlercreutzia	0.00094741	0.0003535	6.59E-06	1.23E-06
Aerococcus	7.06E-07	4.66E-07	2.93E-05	1.04E-05
Aggregatibacter	8.31E-06	5.13E-06	0.00108706	0.00014183
Akkermansia	0.00268821	0.00066918	0.00043613	0.00010519
Anaerococcus	8.54E-07	7.75E-07	1.58E-05	3.75E-06
Anaerostipes	0.00026032	6.53E-05	0.00034683	6.36E-05
Anaerotruncus	4.99E-05	1.80E-05	5.39E-06	2.67E-06
Atopobium	5.53E-06	2.23E-06	0.00042272	3.19E-05
Bacillus	6.64E-07	6.36E-07	1.84E-05	3.17E-06
Bacteroides	0.29013943	0.02109775	0.00840018	0.0013154
Bifidobacterium	0.01829813	0.0065187	0.00315388	0.00064278
Bilophila	0.00113221	0.0001444	0.00034651	6.52E-05
Blautia	0.01098432	0.001956	0.00526647	0.00099596
Brachybacterium	5.13E-08	5.13E-08	3.21E-05	5.29E-06
Brevundimonas	1.44E-07	1.44E-07	1.59E-06	4.31E-07
Bulleidia	6.85E-07	5.30E-07	0.00020587	1.40E-05
Burkholderia	1.36E-06	1.36E-06	3.73E-05	5.89E-06
Butyricimonas	7.75E-07	4.08E-07	1.29E-05	3.68E-06
Butyrivibrio	0.00666776	0.00289482	2.12E-05	2.49E-06
Campylobacter	6.30E-06	2.99E-06	0.00028281	5.46E-05
Catenibacterium	0.0002607	0.00013758	8.73E-06	2.88E-06
Caulobacter	8.14E-06	4.54E-06	4.71E-06	3.38E-06
Cetobacterium	4.80E-06	4.47E-06	7.04E-07	3.60E-07
Citrobacter	0.00078944	0.00046747	3.70E-05	3.02E-06
Clostridium	0.00343966	0.00085761	0.00035388	6.61E-05
Collinsella	0.00203712	0.00056893	0.00100336	0.00025988
Coprobacillus	0.00029756	0.00010014	8.17E-05	3.65E-05
Coprococcus	0.01041403	0.0024852	0.00244928	0.00040413
Corynebacterium	6.73E-06	4.83E-06	0.00059499	0.00010074
Deinococcus	6.97E-07	6.97E-07	4.93E-05	6.36E-06
Delftia	5.40E-08	5.40E-08	7.72E-05	9.51E-06
Desulfovibrio	0.00033105	0.00011009	0.00061664	0.00017221
Dialister	0.00190872	0.00069492	0.00139218	0.00023702
Dorea	0.00293626	0.00042856	0.0004349	5.84E-05

Eggerthella	0.00095208	0.00053581	6.71E-05	1.96E-05
Enterobacter	0.00249825	0.0010198	0.00426978	0.00017481
Enterococcus	0.00178069	0.00130997	0.00045165	0.0001509
Escherichia	0.05092025	0.0097853	5.49E-07	3.96E-07
Faecalibacterium	0.11490269	0.00978819	0.00025102	2.25E-05
Finegoldia	5.34E-06	4.89E-06	8.70E-06	1.76E-06
Fusobacterium	0.00147211	0.00122664	0.00252811	0.00027304
Gardnerella	2.31E-06	1.51E-06	1.48E-06	1.10E-06
Gemella	2.54E-05	1.42E-05	8.18E-05	6.69E-06
Granulicatella	2.74E-05	9.11E-06	0.00242879	0.00014535
Haemophilus	0.00230483	0.00049083	0.01195307	0.00111646
Holdemania	8.44E-05	2.01E-05	2.97E-05	7.72E-06
Klebsiella	0.00652499	0.00277918	7.14E-07	2.66E-07
Kocuria	6.89E-07	6.89E-07	7.25E-05	2.36E-05
Lactobacillus	0.00222834	0.0014829	0.00025986	3.40E-05
Lactococcus	3.20E-05	2.02E-05	1.07E-05	2.59E-06
Lautropia	7.20E-08	4.14E-08	0.00165815	0.00012318
Leptotrichia	9.00E-08	9.00E-08	0.00202201	0.00012525
Leuconostoc	3.13E-05	1.76E-05	3.41E-05	1.51E-05
Macrococcus	2.78E-06	2.78E-06	1.38E-06	3.84E-07
Megamonas	0.01585212	0.00474073	0.00037145	5.20E-05
Megasphaera	0.00093587	0.00037737	0.00054735	6.90E-05
Micrococleus	4.48E-07	4.48E-07	1.09E-06	1.09E-06
Mitsuokella	0.00139377	0.00092944	1.09E-06	3.51E-07
Morganella	6.11E-07	4.31E-07	3.08E-06	1.75E-06
Mycobacterium	1.45E-06	1.45E-06	3.06E-05	1.13E-05
Mycoplasma	4.85E-07	4.85E-07	1.96E-07	1.38E-07
Neisseria	1.12E-06	6.94E-07	0.02508268	0.00219087
Odoribacter	0.00362157	0.00058102	1.90E-05	2.53E-06
Oribacterium	1.65E-06	1.58E-06	0.00026302	1.99E-05
Oxalobacter	1.01E-05	6.04E-06	1.94E-06	6.16E-07
Parabacteroides	0.01438242	0.00176937	8.45E-05	1.22E-05
Paracoccus	7.80E-08	7.80E-08	8.29E-05	3.01E-05
Paraprevotella	0.00385868	0.00083775	2.84E-05	5.10E-06
Parvimonas	8.92E-06	6.01E-06	6.81E-05	5.29E-06
Peptoniphilus	3.14E-07	2.22E-07	1.91E-05	4.48E-06
Peptostreptococcus	2.23E-05	1.10E-05	7.40E-05	1.47E-05
Phascolarctobacterium	0.0048501	0.00152055	0.0064638	0.00108653
Porphyromonas	5.60E-05	3.65E-05	0.00725885	0.00051105
Prevotella	0.17470022	0.0238075	0.01687759	0.00100036
Propionibacterium	1.12E-05	1.03E-05	0.00076495	0.00015864
Proteus	8.41E-06	6.20E-06	5.44E-07	2.81E-07
Providencia	3.81E-07	3.81E-07	1.63E-06	6.24E-07

Pseudomonas	6.75E-06	3.10E-06	0.00115194	0.00016507
Pyramidobacter	5.20E-05	3.63E-05	4.35E-05	9.16E-06
Roseburia	0.02886248	0.0042608	0.00020924	5.07E-05
Rothia	6.59E-05	1.62E-05	0.00228264	0.00017808
Ruminococcus	0.02661598	0.00409609	0.00275781	0.00042481
Salmonella	0.00036895	0.00036895	1.88E-06	4.73E-07
Scardovia	2.02E-05	8.95E-06	1.04E-05	2.48E-06
Selenomonas	2.51E-05	1.96E-05	0.00057192	6.05E-05
Shuttleworthia	7.11E-07	4.44E-07	1.30E-05	1.39E-06
Staphylococcus	1.04E-05	6.94E-06	8.93E-05	7.56E-06
Stenotrophomonas	1.13E-06	1.13E-06	2.81E-05	4.76E-06
Streptococcus	0.0040409	0.00140639	0.01874711	0.00127545
Sutterella	0.00180087	0.0006182	0.00180496	0.00035735
Turicibacter	5.52E-05	2.04E-05	2.27E-05	4.44E-06
Veillonella	0.00345126	0.00078786	0.00672211	0.00040338
Weissella	0.00044631	0.00025918	5.45E-06	1.40E-06

The mean relative abundances and SEM of genera across all the samples from two datasets were calculated. SEM, standard error of the mean.

Supplementary table 8 16S rRNA gene sequencing of the isolated single colony from the joint synovial fluid of patients in the fourth stage of rheumatoid arthritis.

Sample	Sequences	Taxonomy
SF1, SF3, SF4, SF8, SF9	GCTGGCTCCTTACGGTTACCTCACGGACTTCGGGTGTTACCAGCTCTCATGGTGTGACGGGCGGTGTGTACAAGGCCCG GGAACGTATTCACCGCGACATTCTGATTCGCGATTACTAGCAACTCCAGCTTCATGTAGGCGAGTTGCAGCCTACAATCCG AACTGAGATAGGTTTTATAAGTTTTGCTCCACCTCACGGTCTTGCGTCTTATTGTACCTACCATTGTAGCACGTGTGTAGCC CTGGACATAAGGGGCATGATGATTTGACGTCATCCCCACCTTCCTCCTGGTTACCCAGGCAGTCTCATTAGAGTGCTCAAC TTAATGGTAGCAACTAATAACAAGGGTTGCGCTCGTTGCAGGACTTAACCTAACATCTCACGACACGAGCTGACGACAACC ATGCACCACCTGTCTCCTTGCCCCGAAGGGCTTCACCTATCTCTAGGCTATGCAAGGGATGTCAAGTCCAGGTAAGGTTC TTCGCGTTGCTTCGAATTAAACCACATGCTCCGCTGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTTAATCTTGCGAC CGTACTCCCCAGGCGGGATACTTATTGTGTTAACTGCGGCACAGGGGGAGTTGATACCCCCTACACCTAGTATCCATCGT TTACGGCGTGGACTACCAGGGTATCTAATCCTGTTTGCTACCCACGCTTTCGTGCCTCAGCGTCAGTTACAGTCCAGAAAG CCGCCTTCGCCACTGGTGTTCTTCCTAATCTCTACGCATTTCACCGCTACACTAGGAATTCCGCTTTCCTCTCCTGCACT CTAGATATCCAGTTTGGAATGCAGCCCCCAGGTTAAGCCCGGGGATTTCACATCCCACTTAAACATCCGCCTACGCACCC TTTACGCCCAGTAAATCCGGACAACGCTCGCCACCTACGTATTACCGCGGCTGCTGGCACGTAGTTAGCCGTGGCTTCCT CCTCTGGTACCGTCATTATCGTCCCAGAAAACAGGGCTTTACAATCCGAAGACCTTCATCACCCACGCGGCGTTGCTGCG TCAGGGTTTCCCCCATTGCGCAATATTCCCCACTGCTGCCTCCCGTAGGAGTCTGGACCGTGTCTCAGTTCCAATGTGGC CGATCACCCTCTCAGGTCGGCTACGCATCGTTGCCTTGGTAAGCCGTTACCTTACCAACTAGCTAATGCGCCGCGGGTCC ATCTCAAAGCAATAAATCTTTGATAAGAAAATCATGCGATTCTCTTATGTTATGCGGTATTAATCTTCCTTTCGGAAGGCTAT CCCCCACTTTGAGGCAGGTTACCCACGTGTTACTCACCCGTCCGCCGCTAATCCACTTCCCGAAGGAAGCTTCATCGCTC GACTTGCATGTGTTAAGCACGCCGCCAGCGTTCGTCCTGAGCC	1. Clostridium sporogenes strain JCM 1416 (NR_113245.1); 2. Clostridium sporogenes strain McClung 2004 (NR_029231.1)
SF2, SF6, SF7	GGGGCCTGCGCAGCCTATACATGCAAGTCGAACGCTTTTTCTTTCACCGGAGCTTGCTCCACCGAAAGAAAAAGAGTGGC GAACGGGTGAGTAACACGTGGGTAACCTGCCCATCAGAAGGGGATAACACTTGGAAACAGGTGCTAATACCGTATAACACT ATTTTCCGCATGGAAGAAAGTTGAAAGGCGCTTTTGCGTCACTGATGGATGGACCCGCGGTGCATTAGCTAGTTGGTGAGG TAACGGCTCACCAAGGCCACGATGCATAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGAC TCCTACGGGAGGCAGCAGTAGGGAATCTTCGGCAATGGACGAAAGTCTGACCGAGCAACGCCGCGTGAGTGAAGAAGGTT TTCGGATCGTAAAACTCTGTTGTTAGAGAAGAACAAGGATGAGAGTAGAACGTTCATCCCTTGACGGTATCTAACCAGAAAG CCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGC GCAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCCCGGCTCAACCGGGGAGGGTCATTGGAAACTGGGAGACTTGAGTGCA GAAGAGGAGAGTGGAATTCCATGTGTAGCGGTGAAATGCGTAGATATATGGAGGAACACCAGTGGCGAAGGCGGCTCTCTG GTCTGTAACTGACGCTGAGGCTCGAAAGCGTGGGGAGCGAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATG AGTGCTAAGTGTTGGAGGGTTTCCGCCCTTCAGTGCTGCAGCAAACGCATTAAGCACTCCGCCTGGGGAGTACGACCGCA AGGTTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCT TACCAGGTCTTGACATCCTTTGACCACTCTAGAGATAGAGCTTCCCCTTCGGGGGCAAAGTGACAGGTGGTGCATGGTTGT CGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATTGTTAGTTGCCATCATTTAGTTGGG CACTCTAGCGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAATCATCATGCCCCTTATGACCTGGGCT ACACACGTGCTACAATGGGAAGTACAACGAGTTGCGAAGTCGCGAGGCTAAGCTAATCTCTTAAAGCTTCTCTCAGTTCGG ATTGTAGGCTGCAACTCGCCTACATGAAGCCGGAATCGCTAGTAATCGCGGATCAGCACGCCGCGGTGAATACGTTCCCG GGCCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGTCGGTGAGGTAACCTTTTGGAGCCAGCCGC CTAAGTGGAAAAGGAAGT	1. Enterococcus gallinarum strain NBRC 100675 (NR_113924.1); 2. Enterococcus gallinarum strain LMG 13129 (NR_104559.2)
SF5	GGGGCTGCGCGGCTACACTGCAGTCGAACGCTTCTTTCCTCCAGCTTGCTTGCTGTCTGACGAGTGGCGGATGGTGAGT AGTGTATGGGAAACTGCCCGATGGAAGGGGATACCTACTGGACACGGTAGCTAATACCGCATAACGTCTTGGACCATTATG GGGGACCTTCGGGCCTCTTGCCATCGGATGCGTGTCATGGGATTATGGACCCGGTGGGGTAACGGCTCACCTAGGCGAC GATCCCTAGATGGTCTGAATGATGACCAGCCATGCTGGGGCTATCACACCCACCAGACTCCAGCGGGAGGCAGCACTGG GGAATATGGCACGATGGGGGCAAGCCTGATGCAGCACTGCCGGGTGTATGAAGAAGGCCTTCTGGTTGAAAAGTGTTTTCA GCGAGGATGAAGGTGTTGTGGTTAATAACCAAGGAAATTGACGTTACTCGCTCCCCTGACACCGGCTAACTCCGTGCAACG GCCGCTACATACGGAGGGTGCAAGCGTTAATCGGAATTACTGCGCGTAAAGGACAATTGGGGGTCTGTCAAGTCGGATGT GAAATTCCCGGGGATACCTGGGAACTGCATCCGAAACTGGAGGGCTTTGTCTTGTGGAGGACGGTAGAATTCCAGGTGTAA CGGTAATTTGCGGAGAGATCTGGAGGAATACCGATGGCGAAGGAGCCCCCCTGGACAAAGACTGACCCCTCAGGTGCTAA CTCATGCTGAGCCCCAGGATTAGATACCCTGATAGTACACGCCGTACACGATGTCCACTTGCAGGTTGAGCACTTGAAGC GTGGCTACGGGAGCTAACGCGTTAGTCGACGACTGGAGAGTACCGCCGCGAGGGTAAAACTCAATGAATGACGGGGTCCA TACAAGCGGCGGAGCATGTGTTACATTTCGATGCATCGCGTAGAATCTACCTACTCCTGACATCCCCATCAGGTTCTCAGA ATTCCTTGGCGGCCTGCGGGAACTCCTATACAGTGCGGGACAGATTGACGAGGGTCCATGTGTGACTTAAGAGGCAGTCA GCCAACCCTGGAACTGAGACCTGTCAGACTCCTACGAGCAGCAGTGCGATCATCGCCATGACGACAGTCTGACTGCAGCC ATGACGCGTGATTGCAGAGGCATCAGAGTTGTAAAGTACGTCATCGAGAGACGTGATCGTCGATTATTAGCAGCGTGCCAC TGACGTACTCAGCAGAAGAAGCACCGGCTAACTACCGTGCCAGCAGCCGCGGTAATACGGAGGGGTGCAAGCGTGTAATC GGAATTATTGGGCGTAAAGCGACACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCCGGGCTCAACCGTGGGAGGGT GCATCTGAAACTGGCAGACTTAAAGTCTGATGGAGGAGAGATGGGAATTCCATGTGTAGCGGGTGAAAATGCGTAGATAATA TGGAAGGAACACCAGTGGGCGAAGGGCGGGTCTCTTGGTACATAGACTGACGCTTCGGCTCGGAAAGCGTGGGGGAAGC AAACACGGATTAGATACCCCTGTTATTCCCGCCCGTAACGGATGTCGACATGGAGTGTGGTGGGCCTTCACCTGTGGCTT CCGGAGCTAACGCATTAAGTCACTCCGCCTGGGGGAGTACGGCCGCAAGGTTAAAACTCAAAGGAATTGACGGGGGCCC GCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCACTCTTGACATCCAGAGACCTAGCAGAG ATAGAGCTTTCCTTCGGGAACTCAAGGACAGGGTGGGCATGGCGGTTTCTCATCGTGTTGTGAAATGTTGTATTATGTCTC GCACCGAGAGCACCCCTTATCCTTAGTGCCATCATTCGGTCGGCAACTCAAAGGAGACTCCCAGGGCTAAACTGGAGGAA GGTGGGGATGACGTCAAGTCATCATGGCCCTTACGAGTAGGGCTCCACACGTGTTCCAATGGAAGATACAAAGAGAAGCAA CCCCGGGAGTCCAAGCAATCTTCATAAGCTTCTTTCAATTCCGGATGCAGGTTGCAAATTGCCTTCATGAAGCTGGAAATC GTTAGTAATGGGGATCGCACGCCCCCGGTGAATAGTTTCCCGGCCTTTACCCCCCCCCCGTCACCACCGAGAATTGGAA CCCAAGAAGTTGGGGAGGTAACCGTAAGGAGCCAGCTCCTAAGTGAACGTTGT	1. Citrobacter freundii strain LMG 3246 (NR_117752.1); 2. Citrobacter freundii strain JCM 1657 (NR_113340.1)

Supplementary table 9 QC metrics of metagenomic sequences.

Filename	Read count	base count	Q20 (%)	Q30 (%)	GC content (%)
HF16	45426814	6814022100	94.53	86.89	42.7
HF17	42464234	6369635100	96.79	91.31	44.75
HF18	41901214	6285182100	97.05	91.95	44.22
HF20	43980332	6597049800	94.36	86.63	45.49
HF21	44239462	6635919300	95.31	88.27	43.22
HF22	42648660	6397299000	96.68	91.15	45.89
HF25	41728154	6259223100	96.86	91.55	46.14
HF28	42442994	6366449100	97.02	91.77	43.74
HF31	48715360	7307304000	95.06	87.98	43.67
HF32	42228228	6334234200	95	87.81	46.25
HF33	43057824	6458673600	94.84	87.53	46.94
HF36	45431934	6814790100	97.71	93.31	44
HF37	45420844	6813126600	95.15	88.03	45.22
HF38	44381380	6657207000	94.3	86.56	45.06
HF39	41799946	6269991900	95.06	87.91	45.36
HF3	42678234	6401735100	95.19	88.11	44.44
HF41	46132006	6919800900	94.57	86.98	45.28
HF43	41788168	6268225200	94.44	86.81	48.96
HF44	45101768	6765265200	95.25	88.22	44.22
HF46	44819996	6722999400	95.03	87.8	45.08
HF49	42261648	6339247200	96.6	90.94	43.89
HF4	46476802	6971520300	94	85.75	45.38
HF51	41250820	6187623000	96.8	91.38	46.22
HF5	39758590	5963788500	94.5	86.82	42.44
HF7	49212934	7381940100	94.45	86.79	47.59
HF8	40649274	6097391100	94.66	87.15	44.74
HF9	42372194	6355829100	95.95	89.4	44.47
OAF33	44103910	6615586500	96.8	91.37	45.05
OAF39	41028814	6154322100	94.72	87.25	44.02
OAF40	42711768	6406765200	97.15	92.08	45.99
OAF41	41925700	6288855000	93.81	86.2	42.91
OAF43	41294354	6194153100	95.66	89.14	45.54
OAF44	39774718	5966207700	95.3	88.37	46.97
OAF45	40516468	6077470200	97	92.2	50.55
OAF46	42165434	6324815100	94.98	87.84	43.3
OAF49	43398242	6509736300	97.93	93.94	50.6
OAF53	45107442	6766116300	95.37	88.45	44.84
OAF54	43498328	6524749200	97.48	92.92	49.78
OAF63	44173628	6626044200	97.1	91.96	43.53
OAF66	42944028	6441604200	95.2	88.13	45.17
OAF68	45564888	6834733200	94.92	87.69	44.96
OAF69	45345760	6801864000	97.3	92.39	43.26

OAF70	41783728	6267559200	95.09	87.96	49.05
OAF74	40480268	6072040200	96.83	91.43	45.73
OAF77	43064918	6459737700	96.63	91.04	44.71
OAF79	40788874	6118331100	95.95	89.4	44.3
RAF101	43034368	6455155200	94.58	87.17	46.85
RAF102	43178154	6476723100	97.11	92.14	52.33
RAF104	41053400	6158010000	94.89	87.69	44.34
RAF106	44211600	6631740000	96.72	91.26	46.36
RAF107	45474716	6821207400	95.69	89.16	48.41
RAF111	41272668	6190900200	96.8	91.38	45.73
RAF112	44410128	6661519200	96.59	91.06	49.55
RAF113	41958710	6293806500	94.96	87.7	45.23
RAF116	41167720	6175158000	96.51	90.84	45.02
RAF117	40297500	6044625000	94.44	86.77	42.93
RAF118	41616114	6242417100	94.98	87.65	42.66
RAF120	46528830	6979324500	92.37	83.39	44.37
RAF122	42132620	6319893000	96.55	91.12	50.61
RAF123	43624680	6543702000	97.13	91.93	41.14
RAF124	44983948	6747592200	96.78	91.36	46.38
RAF125	40346302	6051945300	95.24	88.14	41.47
RAF126	42991048	6448657200	96.87	91.56	46.75
RAF127	40315134	6047270100	94.66	87.16	43.78
RAF128	44285618	6642842700	94.86	87.62	44.91
RAF129	40564390	6084658500	97.19	92.25	46.88
RAF131	41919100	6287865000	96.95	91.86	52.05
RAF132	41020208	6153031200	96.77	91.34	46.2
RAF135	40542788	6081418200	95.38	88.65	48.34
RAF137	43476654	6521498100	95.41	88.6	45.38
RAF138	43774724	6566208600	97.05	91.92	46.46
RAF139	41291580	6193737000	95.31	88.38	43.58
RAF140	45914110	6887116500	95.3	88.54	51.92
RAF141	47544378	7131656700	94.28	86.53	43.34
RAF142	45594532	6839179800	94.63	87.12	44.94
RAF143	42483748	6372562200	94.87	87.55	42.98
RAF146	47249506	7087425900	94.98	87.74	44.97
RAF147	40224094	6033614100	97.17	92.18	47.36
RAF149	45841894	6876284100	97.61	93.08	43.89
RAF150	40316734	6047510100	97.12	92.07	45.59
RAF151	45314032	6797104800	94.75	87.32	44.55
RAF153	40635770	6095365500	95.55	88.88	45.11
RAF158	40476348	6071452200	96.66	91.15	47.72
RAF159	40627908	6094186200	96.5	90.82	47.55
RAF160	42544754	6381713100	95.23	88.27	45.51
RAF161	41397638	6209645700	96.42	90.7	49.47

RAF166	41874280	6281142000	97.42	92.59	45.65
RAF167	41896240	6284436000	97.13	92.03	44.47
RAF170	43880640	6582096000	96.65	91.07	45.8
RAF174	43765520	6564828000	96.74	91.28	47.36
RAF177	42902240	6435336000	97.5	92.83	43.69
RAF179	41915000	6287250000	97.18	92.15	45.03
RAF180	44576348	6686452200	97.07	91.93	44.21
RAF184	44365800	6654870000	97.25	92.34	46.2
RAF185	43686968	6553045200	97.25	92.31	45.5
RAF186	44277220	6641583000	97.63	93.16	45.45
RAF187	40748474	6112271100	96.78	91.38	45.67
RAF25	43019142	6452871300	97.05	91.86	45.33
RAF44	43568814	6535322100	97.2	92.27	48.01
RAF58	41975628	6296344200	96.45	90.84	45.49
RAF61	42669114	6400367100	94.89	87.63	44.57
RAF63	40123452	6018517800	96.81	91.39	44.33
RAF66	44098600	6614790000	96.66	91.14	47.41
RAF68	44615320	6692298000	97.13	91.94	44.86
RAF70	43380268	6507040200	94.63	87.09	44.51
RAF71	44516994	6677549100	97.5	92.91	46.35
RAF74	41052480	6157872000	96.97	91.8	49.8
RAF75	44768028	6715204200	97.35	92.5	43.84
RAF77	42776628	6416494200	97.42	92.88	46.89
RAF78	40979382	6146907300	95	87.75	45.33
RAF79	43642800	6546420000	95.55	88.85	43.52
RAF80	45535296	6830294400	97.05	91.81	42.02
RAF81	46449788	6967468200	93.88	85.57	46.78
RAF84	43913220	6586983000	97.4	92.69	47.39
RAF86	41103940	6165591000	96.89	91.62	50.15
RAF87	42285100	6342765000	96.09	90.61	47.07
RAF88	47400690	7110103500	94.73	87.45	45.66
RAF89	40617160	6092574000	94.8	87.4	44.51
RAF90	47154162	7073124300	95.16	88.03	45.3
RAF92	40483234	6072485100	96.84	91.4	45.42
RAF95	42641890	6396283500	94.38	86.66	45.22
RAF96	44305868	6645880200	97.12	92.02	43.39

QC, quality control

Q20, percentages of bases whose correct base recognition rates are greater than 99% in total bases.

Q30, percentages of bases whose correct base recognition rates are greater than 99.9% in total bases.

GC content: (G & C base count) / (Total base count)