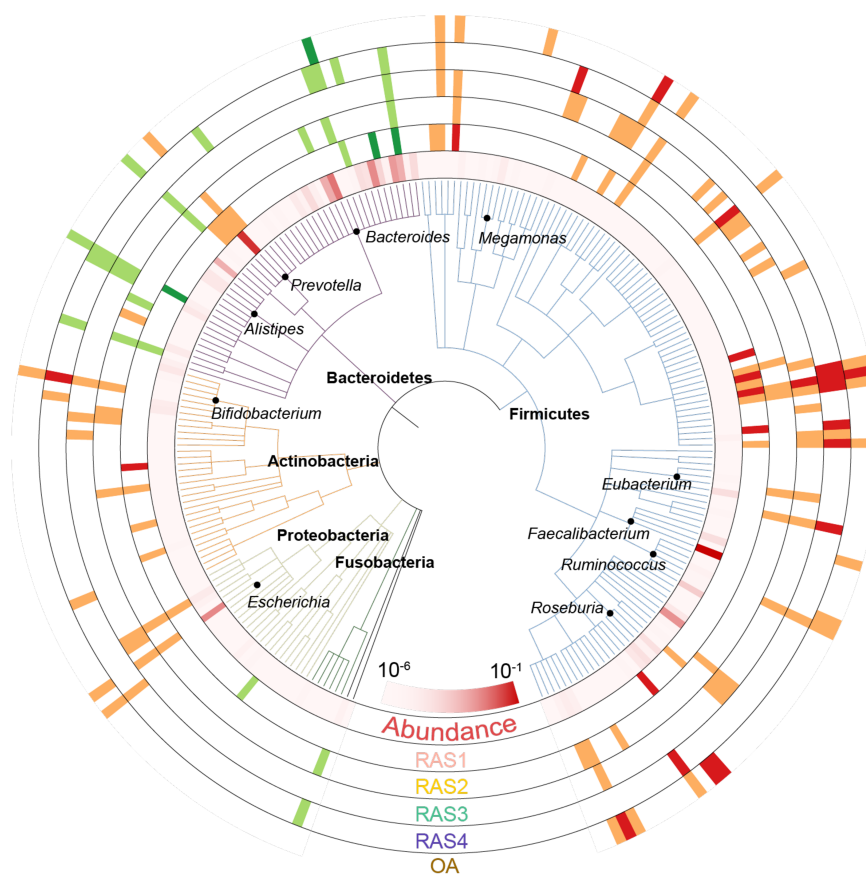
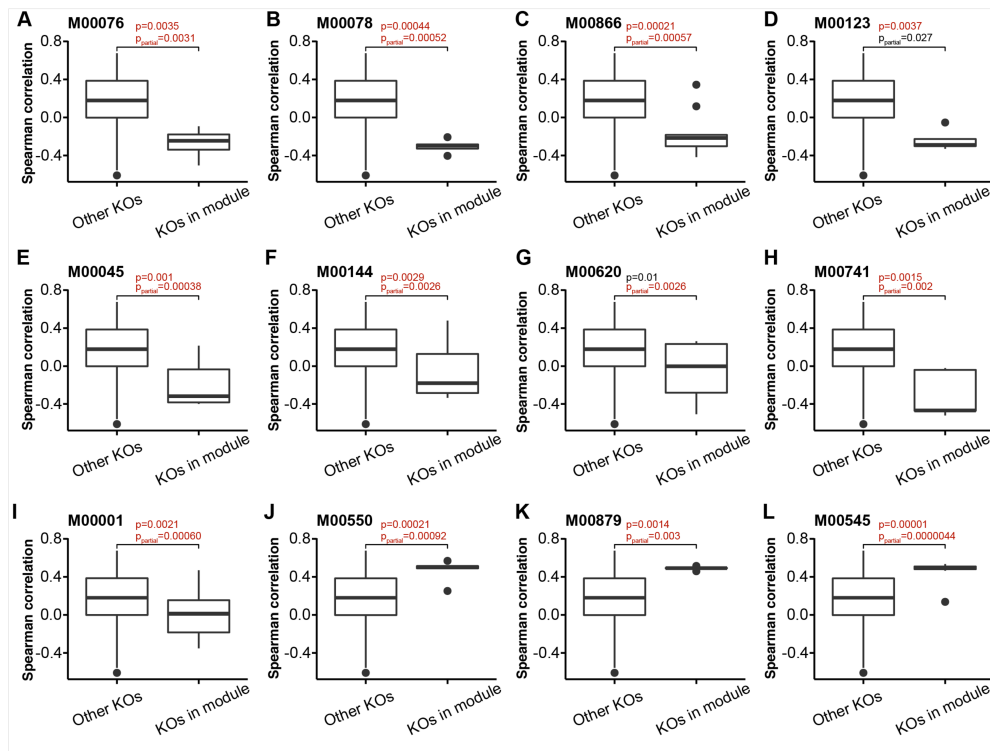


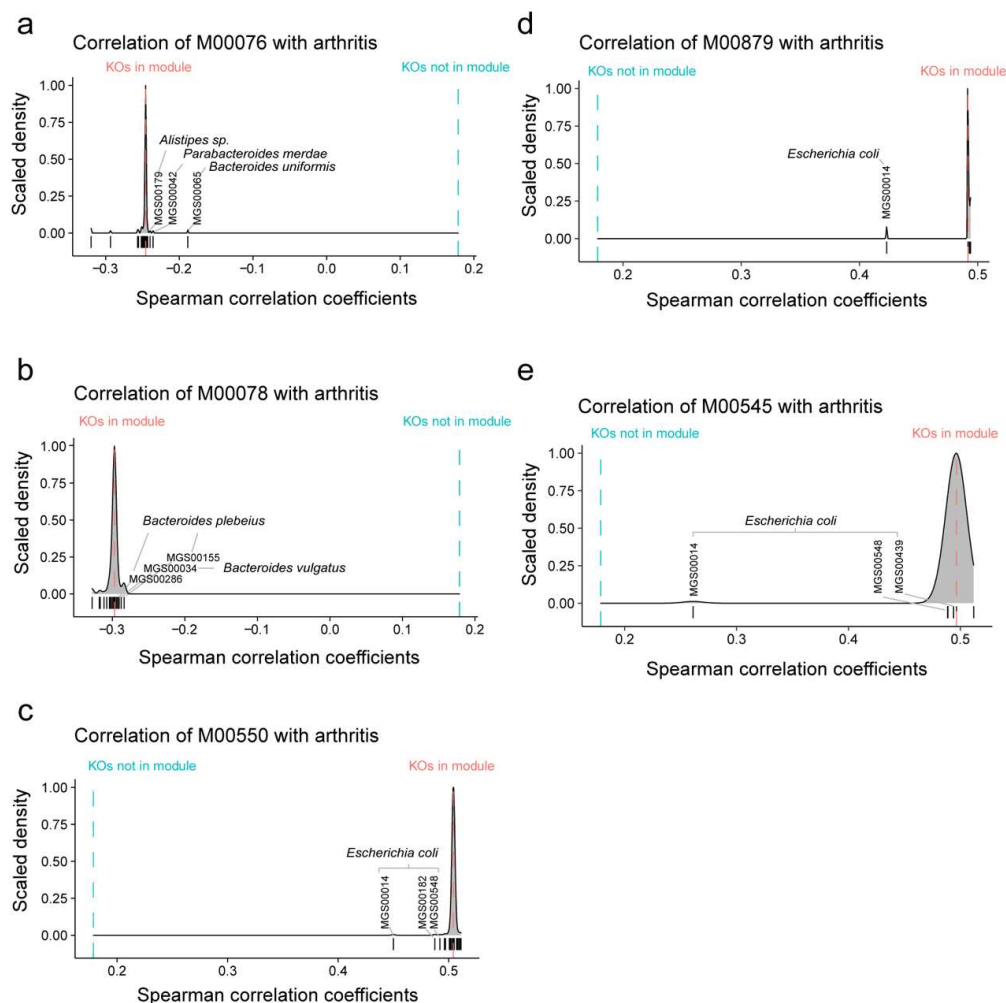
## Supplementary figures



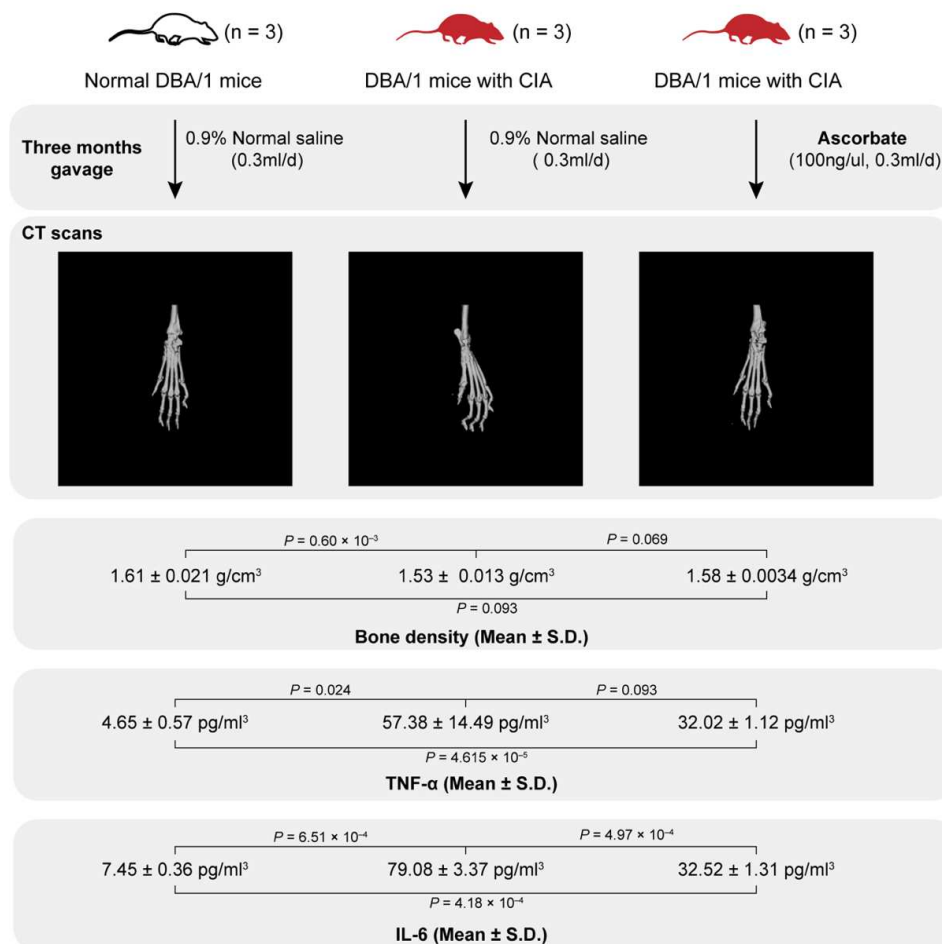
**Supplementary figure 1** Stage-specific microbial taxonomic profiles. In total, 232 classified species whose relative abundance was more than  $1 \times 10^{-6}$  are shown in a phylogenetic tree, grouped in the phyla Fusobacteria, Proteobacteria, Actinobacteria, Bacteroidetes, and Firmicutes. The 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.



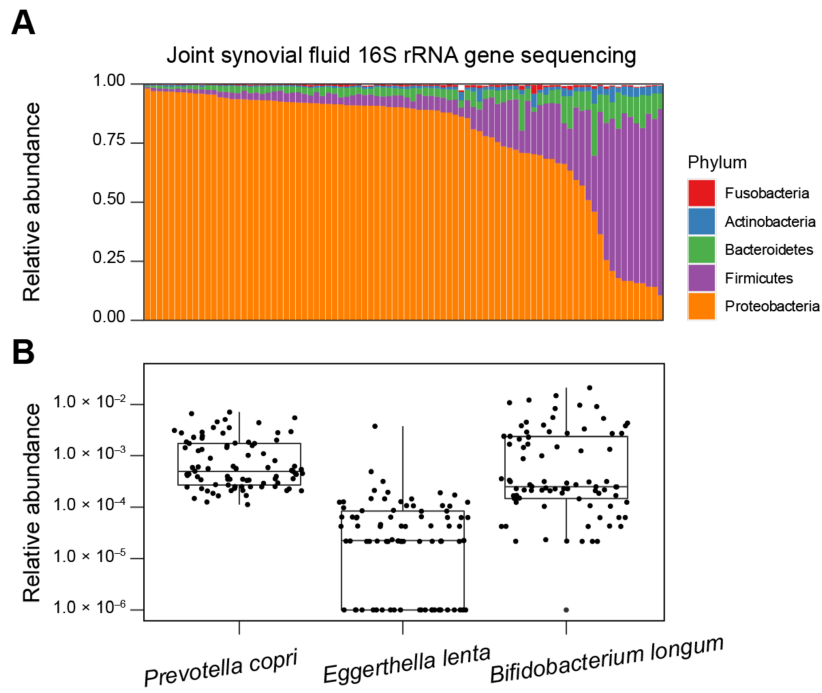
**Supplementary figure 2** KEGG modules that were identified as correlated with arthritis. The correlations between KEGG modules and arthritis (Healthy=0, OA=1, RA=2) were determined by if SCC between the KOs in a given KEGG module and arthritis were different from that between all the other KOs out 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 ( $p_{\text{partial}} < 0.05$  and  $q_{\text{partial}} < 0.1$ , 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  $1.5 \times$  interquartile range from the first and third quartiles, respectively. The p and  $p_{\text{partial}}$  were produced using Mann-Whitney-Wilcoxon test and q and  $q_{\text{partial}}$  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- $\alpha$  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  $1.5 \times$  the interquartile range. RAS4, the fourth stage of rheumatoid arthritis.