ONLINE SUPPLEMENTARY INFORMATION

A metagenome-wide association study of gut microbiome revealed novel etiology of rheumatoid arthritis in the Japanese population

Supplementary Information Contents

The Supplementary Information comprises notes about methods and URLs, the three Supplementary Figures, and two Supplementary Tables.

METHODS

Supplementary Figures

Supplementary Figure 1. Bioinformatic pipelines for the metagenome-wide association study.
Supplementary Figure 2. Phylogenetic relative abundances at the genus level (L6).
Supplementary Figure 3. Deconvolution of the taxonomic relative abundance data based on classical linear machine learning.

Supplementary Tables

Supplementary Table 1. Characteristics of the study population.

Supplementary Table 2. Taxonomic clusters detected by DBSCAN.

METHODS

Patient participation

We examined 85 RA patients at Osaka University Hospital, the National Hospital Organization Osaka Minami Medical Center, and Daini Osaka Police Hospital. RA patients were diagnosed according to the American College of Rheumatology/European League Against Rheumatism 2010 criteria for RA[1]. Exclusion criteria for both sequencing groups were as follows: (i) extreme diets (e.g., strict vegetarians), (ii) treatment with antibiotics for at least 3 months prior to sampling, or (iii) known history of malignancy or serious diseases of the heart, liver, or kidney. We also examined 42 healthy controls at the National Hospital Organization Osaka Minami Medical Center and Osaka University Graduate School of Medicine. Healthy controls were age- and sex-matched individuals with no personal history of immune-related diseases. All subjects provided written informed consent before participation. The study protocol was approved by the ethical committees of Osaka University and related medical institutions.

Sample collection and DNA extraction

Fecal samples were collected in tubes containing RNAlater (Ambion). After the weights of the samples were measured, RNAlater was added to make 10-fold dilutions of homogenates. Fecal samples were stored at -30°C within 24 hours after production. Bacterial DNA was extracted according to a previously described method[2,3]. Briefly, 300 µl of sodium dodecyl sulfate–Tris solution, 0.3 g glass beads (diameter 0.1 mm) (BioSpec), and 500 µl EDTA-Tris-saturated phenol were added to the suspension, and the mixture was vortexed vigorously using a FastPrep-24 (MP Biomedicals) at 5.0 power level for 30 seconds. After centrifugation at 20,000g for 5 minutes at 4°C, 400 µl of supernatant was collected. Subsequently, phenol-chloroform extraction was performed, and 250 µl of supernatant was subjected to isopropanol precipitation. Finally, DNAs were suspended in 200 µl EDTA-Tris buffer and stored at -20°C.

Whole-genome shotgun sequencing

A shotgun sequencing library was constructed using the KAPA Hyper Prep Kit (KAPA Biosystems) and 150 bp paired-end reads were generated on a HiSeq 3000. We conducted two runs (group 1 and group 2). Group 1 consisted of 31 samples from 17 RA individuals and 14 controls (average 29 Gb per sample). Group 2 consisted of 96 samples from 68 RA individuals and 28 controls (average 8.1

Gb per sample). The sequence reads were converted to FASTQ format using bcl2fastq (version 2.19).

Quality control of sequencing reads

We followed a series of steps to maximize the quality of the datasets. The main steps in the QC process were: (i) trimming of low-quality bases, (ii) identification and masking of human reads, and (iii) removal of duplicated reads. We trimmed the raw reads to clip Illumina adapters and cut off low-quality bases at both ends using the Trimmomatic (version 0.33, parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:8:true LEADING:20 TRAILING:20 SLIDINGWINDOW:3:15 MINLEN:60). We discarded reads less than 60 bp in length after trimming. Next, we aligned quality-filtered reads to the human reference genome (hg19) using bowtie2 with default parameters (version 2.3.2) and BMTagger (version 3.101). We kept only reads of which both paired ends failed to align in either tool. The average rates of host DNA contamination were 0.12% for fecal samples. As a final QC step, we removed duplicate reads using PRINSEQ-lite (version 0.20.4, parameters: -derep 1). We excluded three samples (two RA samples from group 1 and one RA sample from group 2) due to extremely low ORF numbers (described below).

Taxonomic annotation of metagenome and abundance quantification

To improve both the efficiency and accuracy of taxonomic assignment, we selected the reference metagenomes of the Japanese population constructed by Nishijima *et al*[4]; 6,139 genomes from the National Center for Biotechnology Information (NCBI) and 10 genomes from in-house complete genome data constructed at Osaka University. Furthermore, we added newly reported genomes from the cultivated human gut bacteria projects[5-7]. After filtration to the genomes annotated to the species with more than 50 reference genomes, the taxonomic reference genome dataset consisted of 7,881 genomes. The filtered paired-end reads were aligned to the reference genome datasets using bowtie2 with default parameters (version 2.3.2). The average mapping rate was 88%. As for multiple-mapped reads, only the best possible alignment was selected by the alignment scores. The number of reads that mapped to each genome was divided by the length of the genome. The value of each genome was summed up by each sample, and the relative abundance of each clade was calculated at six levels (L2: phylum, L3: class, L4: order, L5: family, L6: genus, L7: species).

Functional annotation and abundance calculation

De novo assembly of the filtered paired-end reads into contigs was conducted using MEGAHIT (version 1.1.2, parameters: --min-contig-len 135). We predicted ORFs on the contigs with the *ab initio* gene finder MetageneMark (version 3.38, parameters: -a -k -f G). A non-redundant ORF catalog containing 42,581,555 microbial ORFs was constructed with CD-HIT (version 4.7, parameters: -n 5 -c 0.95 -G 1 -aL 0.9 -aS 0.9 -g 1). As mentioned above, three samples showed extremely low ORF numbers, as well as extremely high rates of duplicated reads during QC. We eventually assessed a total of 124 samples (82 early RA samples: 15 from group 1 and 67 from group 2; 42 control samples: 14 from group 1 and 28 from group 2). Next, we annotated the ORF catalog with two protein databases, UniRef90[8] and KEGG[9]. For the UniRef90 database, we selected prokaryotic, viral, and fungal data. For KEGG genes, we utilized a database of prokaryote KEGG genes and MGENES, a database of KEGG genes from metagenome samples annotated based on orthology, with a bit score > 60. We aligned putative amino acid sequences translated from the ORF catalog against both databases with DIAMOND using BLASTP (v0.9.4.105, parameters: f 6 -b 15.0–k 1 -e 1e-6 --subject-cover 50). We identified 1,738,057 UniRef proteins and 1,295,675 KEGG genes. For quantification of ORF abundance, we mapped the filtered paired-end reads to the assembled contigs using bowtie2 with default parameters (version 2.3.2). To avoid the bias of the gene size, the ORF abundance was defined as the depth of each ORF's region of the ORF catalog according to the mapping result.

Case–control association test for phylogenetic data

We normalized the relative abundance profiles using the Box-Cox transformation function in the car R package (version 3.0.2), including log transformation. We removed clades detected (i) in less than 20% of the samples in either group, (ii) in no sample in either group, or (iii) with an average relative abundance of less than 0.001% of total abundance. After selection, we assessed 803 clades (10 phyla, 23 classes, 34 orders, 72 families, 185 genera, 479 species). Case–control association tests were performed separately for each clade using the generalized linear model function in the R package glm2 (version 1.2.1). We adopted sex, age, sequencing groups, and the top two principal components (PCs) as covariates. We separately obtained PCs from the relative abundance profiles of the groups 1 and 2.

Non-linear unsupervised clustering of metagenome data using UMAP

We used UMAP[10], a recently developed non-linear dimensionality reduction technique, in order to deconvolute the taxonomic relative abundance data at the species level (L7) into two dimensions. To obtain a sufficiently spread distribution for clustering, we adopted UMAP parameters as follows: n_neighbors (number of neighboring points) = 9 and min_dist (tightness of the embedding) = 0.002. After deconvolution, we performed unsupervised clustering of the relative abundance data using the DBSCAN algorithm[11]. We determined the following parameters to optimize the average silhouette width score: minimum number of reachable points = 2, and a reachable ε neighborhood parameter = 0.31. After classifying into clusters, we determined the degree to which each cluster contained the species which showed significant differences in the phylogenic case–control analysis. The definition of "significant" was within the top five percent of p-values. We performed a hypergeometric test with Bonferroni correction for each cluster.

Case–control association test for gene abundance data

We converted each ORF abundance to annotated gene abundance for both UniRef90 protein and KEGG gene databases. We performed two steps of normalization. First, we adjusted the gene abundance by the sum of ORF abundance for each sample in order to correct the bias of the amount of sequence reads for each sample. Next, we applied a rank-based inverse normal transformation in order to correct the heterogeneity of each gene's abundance and distribution. We removed genes detected (i) in less than 20% of the samples in each group or (ii) in no sample in either group. After gene selection, we assessed 179,333 genes annotated by the UniRef90 database and 211,315 genes annotated by the KEGG gene database. Case–control association tests were performed using the generalized linear model function in the R package glm2 (version 1.2.1). We adopted sex, age, sequencing groups, and the top five PCs as covariates. We separately obtained PCs from the relative abundance profiles of group 1 and 2.

Metagenome pathway analysis

We performed GSEA using the R package clusterProfiler (version 3.8.1). Gene sets which contained over 30,000 genes or under 50 genes were excluded from the enrichment analysis. For case–control pathway association tests, genes annotated by the UniRef90 database were ranked based on their effect sizes of case–control gene association tests. The UniRef90 gene sets were composed by

GO[12]. Genes annotated by the KEGG gene database were ranked in the same way. The KEGG gene sets were defined according to the KEGG pathway.

Comparison of pathway analysis results between RA metagenome and host GWAS

We assessed whether there were shared biological pathways between the gut metagenome and the human germline genome; we compared the pathway enrichment data of the metagenome with that of the host GWAS in RA. For the host GWAS in RA, we used Pascal with summary statistics from RA GWAS in the East Asian and European population in order to determine KEGG pathway enrichment of the germline in RA (22,515 East Asians and 58,284 Europeans)[13]. We compared the p-values of KEGG pathways shared between the GWAS data and metagenome data. We evaluated the overlap of the pathway enrichment, by classifying the pathways based on the significance threshold of $P < 0.05$ or $P \geq 0.05$ and using Fisher's exact test.

Empirical estimation of metagenome-wide significance threshold

We empirically estimated the statistical significance threshold for both phylogenetic and gene case–control analyses, performing a phenotype permutation procedure[14]. We randomly simulated case–control phenotypes ($\times$50,000 iterations) and calculated empirical null distributions of the minimum p-values (= P_{min}) in each iteration. We defined an empirical Bonferroni significance threshold, $-\log_{10}(P_{sig})$, as the 95th percentile of $-\log_{10}(P_{min})$ at a significance level of 0.05. We calculated $-\log_{10}(P_{sig})$ using the Harrell–Davis distribution-free quantile estimator[15] and calculated a 95% confidence interval for $-\log_{10}(P_{sig})$ by a bootstrapping method in the R package Hmisc (version 4.1.1). To estimate the null distribution of the test statistics, we applied the same process used for minimum p-values to all the ranked p-values. We defined an empirical FDR threshold of 0.05 as the 95th percentile of $-\log_{10}$ p-values of each rank at a significance level of 0.05.

Case–control difference between alpha-diversity and beta-diversity of the metagenome

Alpha-diversity (within-sample diversity) was calculated based on gene abundance and six levels of phylogenetic relative abundance (L2–L7) for each sample according to the Shannon index. Statistical comparisons of Shannon index between RA cases and controls were assessed by Welch's t-test. To quantify beta-diversity, MDS on the Bray–Curtis dissimilarity was performed. For evaluating case–control differences in the dissimilarity, we performed permutational multivariate analysis of variance

(PERMANOVA)[16] with 100,000 permutations using the R package vegan (version 2.5.4).

URLs

The URLs for data presented herein are as follows:

Trimmomatic, http://www.usadellab.org/cms/?page=trimmomatic

bowtie2, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

BMTagger, ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/

PRINSEQ, http://prinseq.sourceforge.net/

MEGAHIT, https://github.com/voutcn/megahit

MetaGeneMark, http://exon.gatech.edu/meta_gmhmmp.cgi

CT-HIT, http://weizhongli-lab.org/cd-hit/

KEGG genes, https://www.kegg.jp/kegg/genes.html

MGENES, ftp://ftp.genome.jp/pub/db/mgenes

DIAMOND, https://ab.inf.uni-tuebingen.de/software/diamond/

R car package, https://cran.r-project.org/web/packages/car/index.html

R glm2 package, https://cran.r-project.org/web/packages/glm2/index.html

R clusterProfiler package, http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html

KEGG pathway, http://www.genome.jp/kegg/pathway.html

Pascal, https://www2.unil.ch/cbg/index.php?title=Pascal

R Hmisc package, https://cran.r-project.org/web/packages/Hmisc/index.html

R vegan package, https://cran.r-project.org/web/packages/vegan/index.html

REFERENCES

1 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 Rheum* 2010; 62: 2569-81.

2 Maeda Y, Kurakawa T, Umemoto E *et al*. Dysbiosis Contributes to Arthritis Development via Activation of Autoreactive T Cells in the Intestine. *Arthritis Rheumatol* 2016; 68: 2646-61.

3 Okumura R, Kurakawa T, Nakano T *et al*. Lypd8 promotes the segregation of flagellated microbiota and colonic epithelia. *Nature* 2016; 532: 117.

4 Nishijima S, Suda W, Oshima K *et al*. The gut microbiome of healthy Japanese and its microbial and functional uniqueness. *DNA Res* 2016; 23: 125-33.

5 Zou Y, Xue W, Luo G *et al*. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* 2019; 37: 179-85.

6 Forster SC, Kumar N, Anonye BO *et al*. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol* 2019; 37: 186-92.

7 Almeida A, Mitchell AL, Boland M *et al*. A new genomic blueprint of the human gut microbiota. *Nature* 2019; 568: 499-504.

8 Suzek BE, Wang Y, Huang H *et al*. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015; 31: 926-32.

9 Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000; 28: 27-30.

10 McInnes L, Healy J, Saul N *et al*. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw* 2018; 3: 861.

11 Ester M, Kriegel H-P, Sander J *et al*. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. Second Int. Conf. Knowl. Discov. Data Min* : 1996: 226-31.

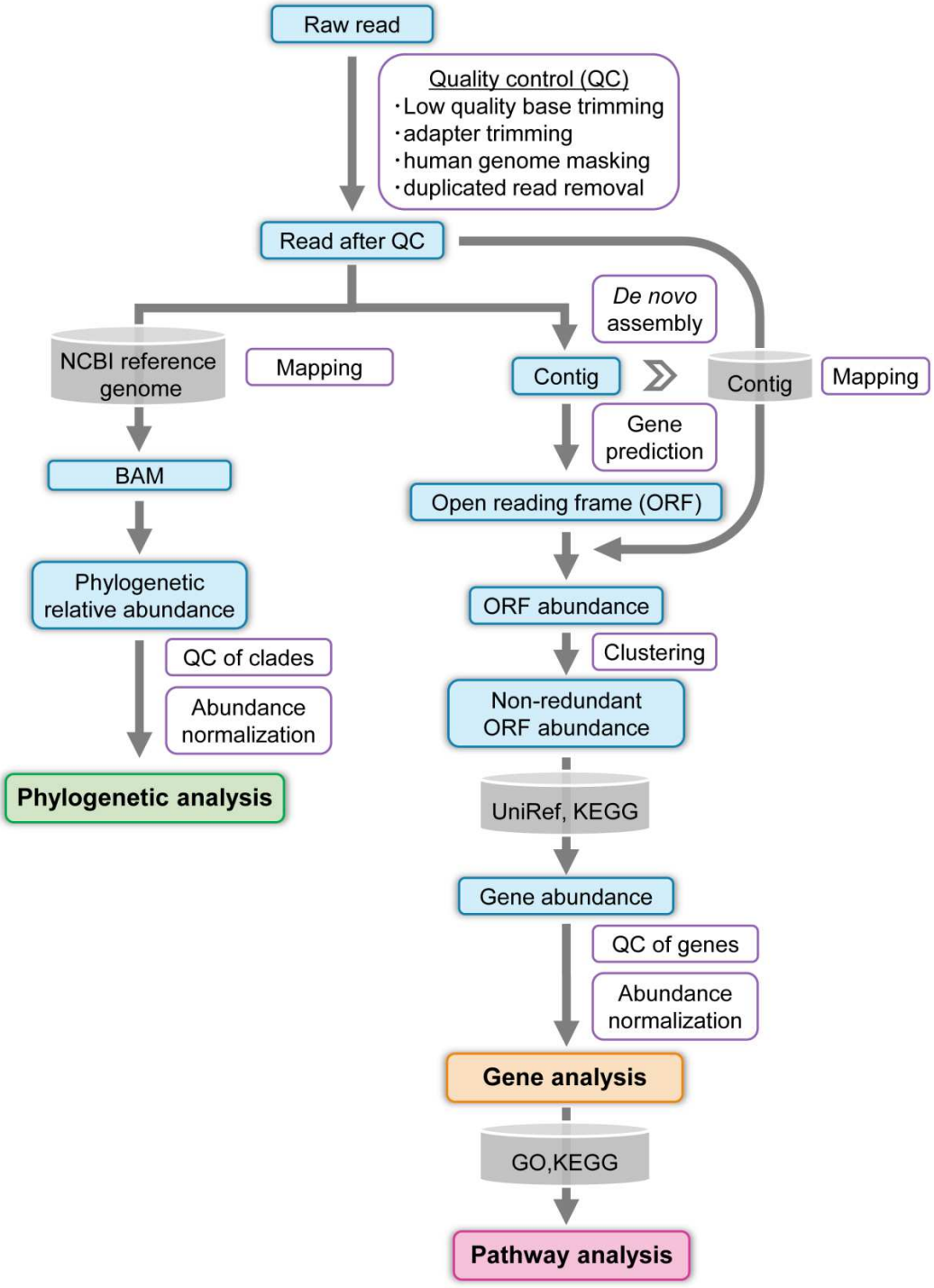
12 Harris MA, Clark J, Ireland A *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004; 32: D258-61.

13 Okada Y, Wu D, Trynka G *et al*. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014; 506: 376-81.

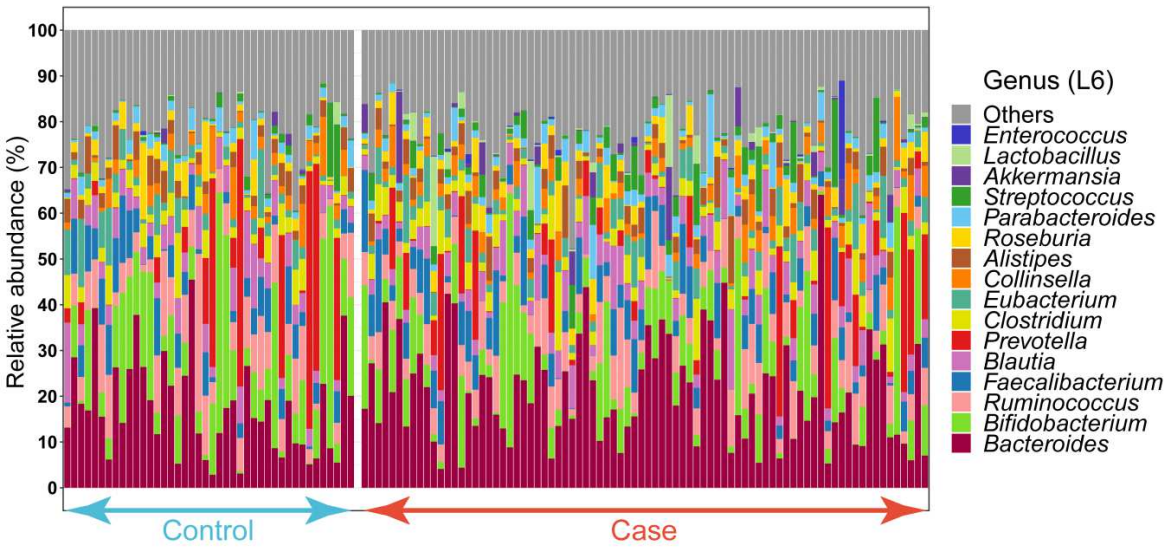
14 Kanai M, Tanaka T, Okada Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J Hum Genet* 2016; 61: 861-66.

15 Harrell FE, Davis CE. A new distribution-free quantile estimator. *Biometrika* 1982; 69: 635-40.

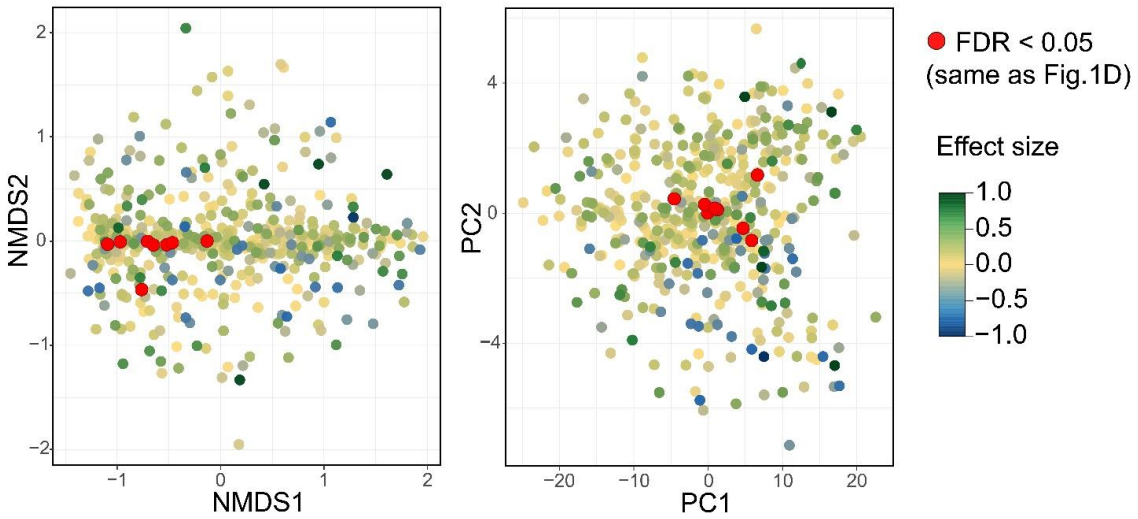
16 McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 2001; 82: 290-7.



Supplementary Figure 1. Bioinformatic pipelines for the metagenome-wide association study. Whole-genome shotgun sequencing reads of the gut microbiome were processed following this pipeline. It consisted of three major bioinformatic analytic techniques (phylogenetic analysis, functional gene analysis, and pathway analysis).



Supplementary Figure 2. Phylogenetic relative abundances at the genus level (L6). The relative abundance profiles were constructed utilizing whole genome shotgun sequencing ($n_{control}$ = 42, n_{case} = 82).



Supplementary Figure 3. Deconvolution of the taxonomic relative abundance data based on classical linear machine learning. Dimension-reduced plots of the 479 species using non-metric multidimensional scaling (MDS; left) and principal component analysis (PCA; right). The eight species with $q < 0.05$ are indicated in red, while the others are shown according to the effect sizes in phylogenetic case–control association tests.

Supplementary Table 1. Characteristics of the study population.

	RA (n = 82)	Control (n = 42)
Age, years, mean (median)	55.7 (54)	53.0 (52)
Sequencing group1	15 (18.3%)	14 (33.3%)
Female	66 (80%)	38 (90%)
Disease duration < 1 year	60 (73%)	
Treatment		
non-treatment	58 (71%)	
csDMARDs use *1	23 (28%)	
bDMARDs use *2	2 (2 %)	
ACPA *3 > 4.5 U/mL	54 (66%)	
RF *4 > 15 IU/mL	61 (74%)	
Stage *5 I / II / III / IV	58 / 18 / 3 / 3	
DAS28-CRP *6, mean (range)	4.04 (1.08 – 6.67)	

*1 Conventional synthetic disease modifying anti-rheumatic drugs
*2 Biological DMARDs
*3 Anti-citrullinated peptide antibody
*4 Rheumatoid factor
*5 Stage indicates Steinbrocker classification of the joint X-rays
*6 Disease Activity Score 28-joint count C-reactive protein

Supplementary Table 2. Taxonomic clusters detected by DBSCAN.

Cluster label	*P*-value of the hypergeometric test	Species number in the cluster
0 (noise)	1	1
1	0.947	56
2	0.224	87
3	0.952	58
4	0.649	67
5	1	15
6	1	27
7	0.941	86
8	0.652	21
9	1.91×10^{-8}	14
10	1	4
11	1	16
12	1	13
13	1	5
14	0.36	9