TRANSLATIONAL SCIENCE

Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population

Toshihiro Kishikawa,[1,2] Yuichi Maeda,[3,4] Takuro Nii,[3,4] Daisuke Motooka,[5] Yuki Matsumoto,[5] Masato Matsushita,[6,7] Hidetoshi Matsuoka,[7] Maiko Yoshimura,[7] Shoji Kawada,[8] Satoru Teshigawara,[7] Eri Oguro,[3,7] Yasutaka Okita,[7] Keisuke Kawamoto,[8] Shinji Higa,[8] Toru Hirano ,[3] Masashi Narazaki,[3] Atsushi Ogata,[8] Yukihiko Saeki ,[7,9] Shota Nakamura,[5] Hidenori Inohara,[2] Atsushi Kumanogoh,[3,10] Kiyoshi Takeda,[4,11] Yukinori Okada [1,12,13]

Handling editor Josef S Smolen

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/annrheumdis-2019-215743).

For numbered affiliations see end of article.

Correspondence to
Dr Yukinori Okada, Department of Statistical Genetics, Osaka University Graduate School of Medicine, 2-2 Yamadaoka, Suita, Japan;
yokada@sg.med.osaka-u.ac.jp

Received 19 May 2019
Accepted 3 October 2019
Published Online First
7 November 2019

To cite: Kishikawa T, Maeda Y, Nii T, *et al*. *Ann Rheum Dis* 2020;**79**:103–111.

ABSTRACT

Objective The causality and pathogenic mechanism of microbiome composition remain elusive in many diseases, including autoimmune diseases such as rheumatoid arthritis (RA). This study aimed to elucidate gut microbiome's role in RA pathology by a comprehensive metagenome-wide association study (MWAS).

Methods We conducted MWAS of the RA gut microbiome in the Japanese population (n_{case}=82, $n_{control}$=42) by using whole-genome shotgun sequencing of high depth (average 13 Gb per sample). Our MWAS consisted of three major bioinformatic analytic pipelines (phylogenetic analysis, functional gene analysis and pathway analysis).

Results Phylogenetic case–control association tests showed high abundance of multiple species belonging to the genus *Prevotella* (e.g., *Prevotella denticola*) in the RA case metagenome. The non-linear machine learning method efficiently deconvoluted the case–control phylogenetic discrepancy. Gene functional assessments showed that the abundance of one redox reaction-related gene (R6FCZ7) was significantly decreased in the RA metagenome compared with controls. A variety of biological pathways including those related to metabolism (e.g., fatty acid biosynthesis and glycosaminoglycan degradation) were enriched in the case–control comparison. A population-specific link between the metagenome and host genome was identified by comparing biological pathway enrichment between the RA metagenome and the RA genome-wide association study results. No apparent discrepancy in alpha or beta diversities of metagenome was found between RA cases and controls.

Conclusion Our shotgun sequencing-based MWAS highlights a novel link among the gut microbiome, host genome and pathology of RA, which contributes to our understanding of the microbiome's role in RA aetiology.

Key messages

What is already known about this subject?

- Rheumatoid arthritis (RA) is one of the diseases for which the microbiome may have an important role in pathology. Gut microbiome has been implied to lead immune abnormality in RA patients such as the activation of immune responses via Th17 cells by *Prevotella copri*.

What does this study add?

- Multiple *Prevotella spp.* other than *P. copri* were related to RA etiology in the gut microbiome of the Japanese population.
- A redox reaction–related gene (R6FCZ7) was abundant in the gut metagenome of the Japanese patients with RA.
- A population-specific biological pathway link between the metagenome and the host genome was identified by comparing the RA metagenome-wide association study (MWAS) and the RA genome-wide association study (GWAS).
- Our study indicated a value of metagenome-wide shotgun sequencing rather than classical amplicon sequencing of 16S ribosomal RNA (rRNA) genes of microbiomes.

How might this impact on clinical practice or future developments?

- We revealed a novel link between the gut microbiome, host genome and pathology of RA. Our study will be a platform model of the microbiome studies to elucidate etiology of rheumatic diseases.

INTRODUCTION

The human microbiome, which refers to the microbial communities inhabiting the human body, has remarkable effects on our immune system and metabolism.[1] Different microbiome compositions have been implicated in the pathogenesis of many diseases, such as type 2 diabetes, cardiometabolic disorders, inflammatory bowel diseases and cancer.[2–5] Rheumatoid arthritis (RA) is a prevalent autoimmune disorder characterised by synovial inflammation and joint damage.[6] In addition to environmental and genetic factors, the microbiome has emerged as a candidate factor responsible for the onset of RA.[7] The microbiomes of various sites within the body such as the gut, oral cavity and lungs have been implicated in RA.[8–14] Several mechanisms leading to immune abnormality in RA have been reported (e.g, the production of citrullinated peptides by *Porphyromonas gingivalis*[15] and the activation of immune responses via Th17 cells by *Prevotella copri*[13]). However, the aetiology of RA linked to the microbiome still remains unknown. Additionally, although ethnicity and dietary habits are major factors leading to differences in microbiome compositions, there have been few population-specific microbiome studies focused on RA in non-European or non-Westernised populations.

The spread of microbiome research in academia and industry has contributed to the development of the theoretical and methodological features of bioinformatic analysis methods. While such methods initially focused on amplicon sequencing of 16S ribosomal RNA (rRNA) genes, today it is recommended to perform metagenome-wide association studies (MWAS) based on whole-genome shotgun sequencing of the microbiome. This method can detect the genomic composition of the microbiome at the species level (bacteria, and also archaea, fungi and viruses) and analyse their biological functional features.[16] Thus, shotgun sequencing has the potential to indicate new targets for drug therapy as well as faecal transplants and probiotics.[17] On the other hand, shotgun sequencing analysis is costly and requires a machine resource that can analyse large data sets as well as complex systematic data processing. As for the gut microbiome in RA, a few studies using shotgun sequencing have been reported,[8 11] and findings are not universally consistent.

As previously suggested, shotgun sequencing has the benefit of providing insights into the functional aspects of microbiome genes. RA is one of the representative diseases of which many associated genomic regions have been discovered by genome-wide association studies (GWAS), and a part of its pathogenesis has been elucidated.[18] However, the link between the host genetic factor and the microbiome, namely MWAS–GWAS interaction, has not been fully assessed to date.

Unsupervised clustering analyses of taxa or genes have been performed to grasp the overall phylogenetic or functional picture of the metagenome. Machine learning (ML) is one of today's most rapidly growing technologies. Various clustering approaches that incorporate an ML method have been used, such as principal component analysis (PCA) and the k-means method,[19 20] but these classical methods mostly assume linearity of the microbiome data. Recently, ML methods for non-linear dimensionality reduction such as uniform manifold approximation and projection (UMAP)[21] have been successfully adopted in diverse analyses.[22] Application of such non-linear ML methods should contribute to our advanced understanding of the metagenome.

Here, we report a comprehensive MWAS of the gut microbiome in an RA case–control cohort of the Japanese population. We carried out whole-genome shotgun sequencing of 124 faecal samples (82 individuals with RA and 42 healthy controls). Our MWAS consisted of three major bioinformatic analytic techniques (phylogenetic analysis, functional gene analysis and pathway analysis), which allowed us to comprehensively grasp case–control disparity in the gut metagenome. We newly introduced an unsupervised ML-based clustering approach to depict the RA metagenome landscape. To evaluate the link between the gut metagenome and the human germline genome, we compared the pathway enrichment of the gut microbiome MWAS and that of the host GWAS in RA.

METHODS

Methods are provided in the online supplementary information.

RESULTS

High abundance of multiple species belonging to the genus *Prevotella* in the RA gut microbiome

We performed whole-genome shotgun sequencing of a total of 124 faecal DNA samples (sequencing group 1: 29 samples from 15 individuals with RA and 14 controls; sequencing group 2: 95 samples from 67 individuals with RA and 28 controls; see online supplementary table 1), which passed stringent quality control (QC) for sequence reads and samples (online supplementary figure 1).

After QC, we assigned sequence reads to taxonomic references and quantified phylogenetic relative abundances separately for each taxonomic level (online supplementary figure 2). We selected a total of 803 clades with an average relative abundance of $>1.0\times10^{-5}$ and detected in more than 20% of the samples in both sequencing groups, including 10 phyla (L2), 23 classes (L3), 34 orders (L4), 72 families (L5), 185 genera (L6) and 479 species (L7). Case–control phylogenetic association tests using a generalised linear regression model identified significant associations of the nine clades (empirically estimated false discovery rate (FDR) $q<0.05$; figure 1A,B, table 1). Notably, all the associated clades increased in RA samples compared with controls. As illustrated in a phylogenetic tree indicating the case–control association results of multilayered taxonomic levels (figure 1C), clades with more specific taxonomic levels (such as genus (L6) or species (L7)) tended to show greater differences in case–control factors (i.e., effect size). Among the nine clades with significant case–control associations, eight clades belonged to species (L7, the most specific level). Because it was difficult to detect these species level clades using classical 16S rRNA sequence analysis, these results clearly demonstrate the value of metagenome shotgun sequencing in identifying disease-associated taxa. We note that our shotgun sequencing study newly identified that the majority of RA-associated clades belonged to the genus *Prevotella* (*P. denticola*, *P. marshii*, *P. disiens*, *P. corporis* and *P. amnii*). While *P. copri* has been reported to be abundant in the gut microbiome of patients with RA,[13] our results revealed that multiple other species in the genus *Prevotella* are also related to RA aetiology.

Identification of significant taxa cluster using a non-linear ML method

We performed ML-based deconvolution of the metagenome data to comprehensively assess the case–control discrepancy at the species level (L7). First, we adopted UMAP, a non-linear dimensionality reduction technique, for deconvolution of the complex taxonomic relative abundance data into two-dimensional data (figure 1D). We subsequently performed unsupervised clustering of the results of UMAP with the density-based spatial clustering of applications with noise (DBSCAN) algorithm,[23] which classified the species into 14 clusters (online supplementary table 2). Next, we assessed the degree to which each cluster contained species with significant differences in the case–control phylogenetic association tests. One cluster was enriched with species with case–control discrepancies, as shown by hypergeometric tests ($p=1.9\times10^{-8}$, OR=27; figure 1E). Among the eight species which showed q-values <0.05 in the case–control phylogenetic association tests, half belonged to this cluster ($n=4$). The sum of the species belonging to this cluster also showed a significant difference in the case–control association test ($p=0.0057$,

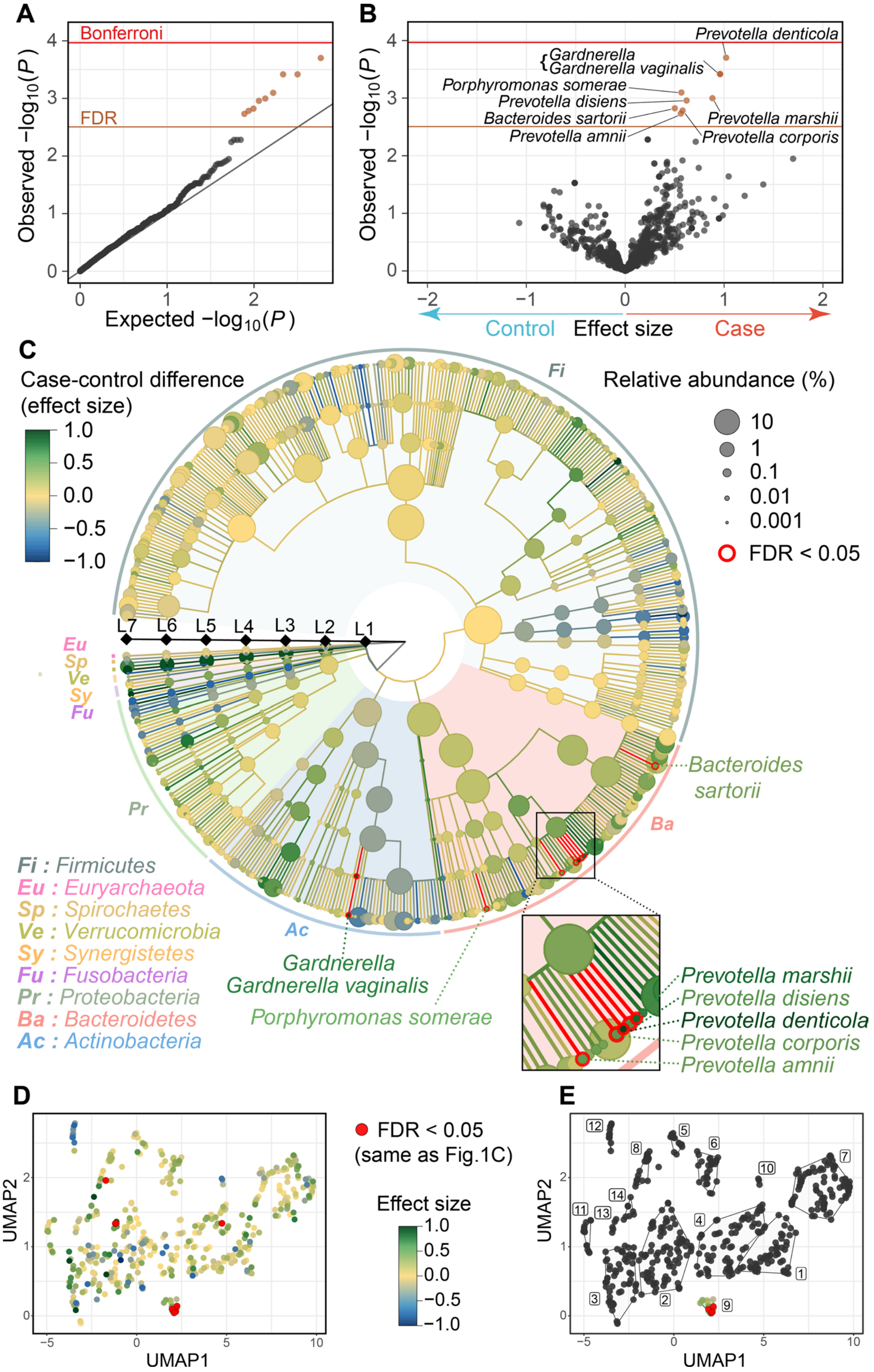

Figure 1 MWAS results of RA case–control phylogenetic association tests. (A) A quantile-quantile plot of the MWAS p values of the clades. The *x*-axis indicates empirically estimated median $-\log_{10}$ p values. The *y*-axis indicates observed $-\log_{10}$ p values. The diagonal grey line represents $y = x$, which corresponds to the null hypothesis. The horizontal red line indicates the empirical Bonferroni-corrected threshold (α=0.05), and the brown line indicates the empirically estimated FDR (q=0.05). Clades with p values less than the Bonferroni thresholds are plotted as red dots. Clades with q<0.05 are plotted as brown dots, and other clades as black dots. (B) A volcano plot. The *x*-axis indicates effect sizes of generalised linear model. The *y*-axis, horizontal lines and dot colours are the same as in (A). (C) Phylogenetic tree. Levels L2–L7 are from the inside layer to the outside layer. The size and colour of dots represent relative abundance and effect sizes, respectively. The nine clades with significant case–control associations (q<0.05) are outlined in red. (D) A dimension-reduced plot of the 479 species using UMAP. 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. (E) Unsupervised clustering results according to the density-based spatial clustering of applications with noise (DBSCAN) algorithm. Seven clusters are illustrated as polygons. Cluster 9, which shows the significant enrichment of species with significant case–control discrepancies, is coloured according to effect sizes (as in figure 1D), while other clusters are shown in black. FDR, false discovery rate; MWAS, metagenome-wide association study; RA, rheumatoid arthritis; UMAP, uniform manifold approximation and projection.

effect size=0.48). While we parallelly applied classical linear ML methods such as PCA and non-metric multidimensional scaling in the same manner, cluster deconvolution was not clear and no cluster showed a significant case–control discrepancy (online supplementary figure 3). These results indicate that the non-linear ML method could efficiently provide novel knowledge in metagenome analyses.

Table 1 Clades with significant case–control discrepancy

Microbe	Level	Effect size	*p*-value
Prevotella denticola	Species (L7)	1.021	2.0×10^{-4}
Gardnerella	Genus (L6)	0.959	3.8×10^{-4}
Gardnerella vaginalis	Species (L7)	0.959	3.8×10^{-4}
Porphyromonas somerae	Species (L7)	0.568	8.0×10^{-4}
Prevotella marshii	Species (L7)	0.882	1.0×10^{-3}
Prevotella disiens	Species (L7)	0.621	1.1×10^{-3}
Bacteroides sartorii	Species (L7)	0.502	1.5×10^{-3}
Prevotella corporis	Species (L7)	0.580	1.6×10^{-3}
Prevotella amnii	Species (L7)	0.563	1.9×10^{-3}

A redox reaction-related gene of the gut metagenome decreased in RA samples

To evaluate functional aspects of the RA gut metagenome, we performed metagenomic shotgun sequencing according to the following procedures: (1) de novo assembly, (2) prediction of open reading frames (ORFs), (3) clustering and annotation of ORFs, and (4) read mapping to assembled contigs. We selected 179 333 and 211 315 genes annotated by the UniRef90[24] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[25] databases, respectively. Case–control gene association tests using a generalised linear regression model found significant associations for a total of 12 genes (1 and 11 for UniRef90 and KEGG, respectively; empirically estimated FDR: $q<0.05$; figure 2, table 2). Of these, nine showed increased abundance in RA

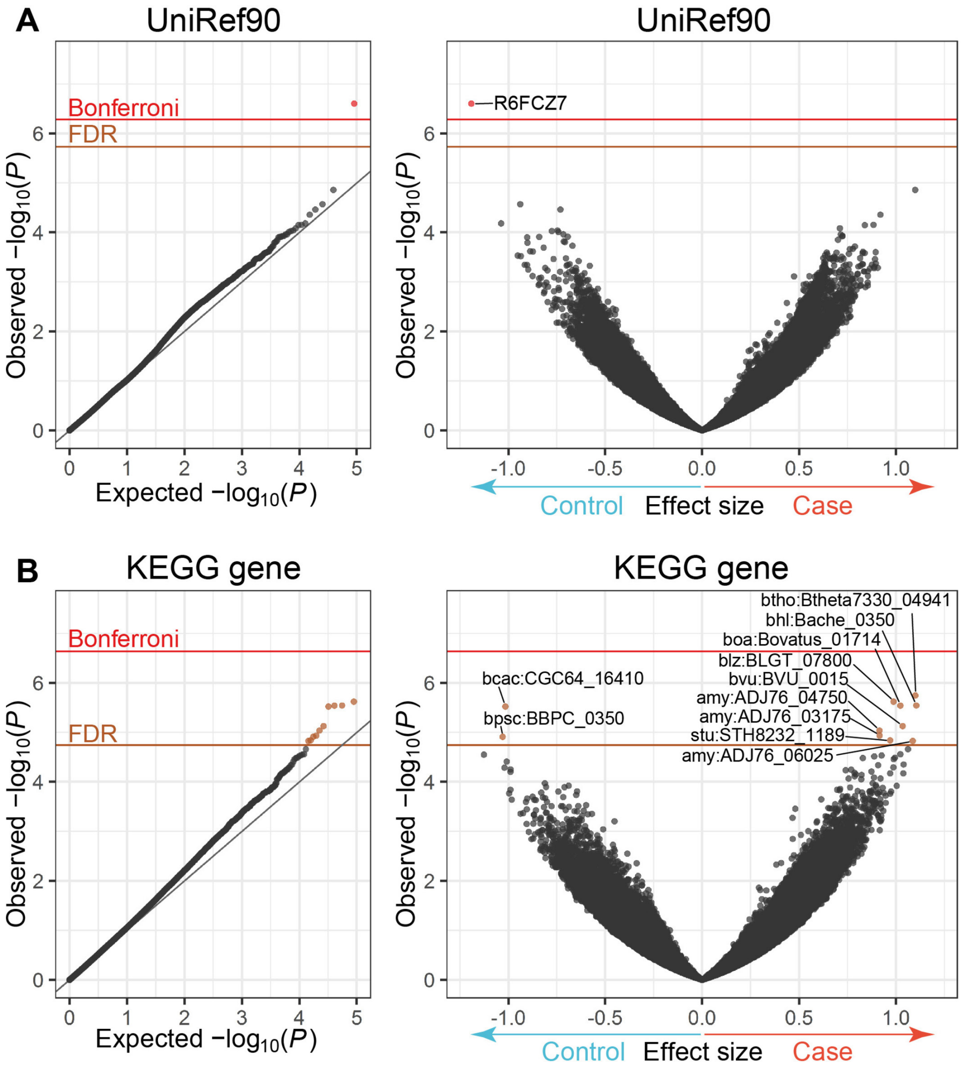

Figure 2 MWAS results of RA case–control gene association tests. (A) A QQ plot (left) and volcano plot (right) of the MWAS p values of genes based on the UniRef90 protein database. (B) A QQ plot (left) and a volcano plot (right) of genes based on the KEGG gene database. In the QQ plots, the *x*-axis indicates empirically estimated median $-\log_{10}$ p values. In the volcano plot, the *x*-axis indicates beta of generalised linear model as effect sizes. The *y*-axis in both plots indicates observed $-\log_{10}$ p values. The diagonal grey line represents $y=x$, which corresponds to the null hypothesis. The horizontal red line indicates the empirical Bonferroni-corrected threshold (α=0.05), and the brown line indicates the empirically estimated FDR (q=0.05). Genes with p values less than Bonferroni thresholds are plotted as red dots. Clades with $q<0.05$ are plotted as brown dots, and other clades as black dots. MWAS, KEGG, Kyoto Encyclopedia of Genes and Genomes; MWAS, metagenome-wide association study; QQ, quantile-quantile; RA, rheumatoid arthritis.

Table 2 Genes with significant case–control discrepancy

UniRef ID	Effect size	*p*-value	Description	Organism
R6FCZ7	−1.19	2.5×10^{-7}	FeS assembly sulphur utilization factor system protein	*Bacteroides sp. CAG:633*

KEGG gene	Effect size	*p*-value	Gene name, definition	Organism
btho:Btheta7330_04941	1.10	1.8×10^{-6}	acpP; Acyl carrier protein	*Bacteroides thetaiotaomicron 7330*
blz:BLGT_07800	0.988	2.4×10^{-6}	ABC transporter permease	*Bifidobacterium longum subsp. longum GT15*
bhl:Bache_0350	1.11	2.9×10^{-6}	Protein of unknown function DUF59	*Bacteroides helcogenes*
boa:Bovatus_01714	1.02	2.9×10^{-6}	gpmA; 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	*Bacteroides ovatus*
bcac:CGC64_16410	−1.02	3.0×10^{-6}	gpmA; 2,3-diphosphoglycerate-dependent phosphoglycerate mutase	*Bacteroides caccae*
bvu:BVU_0015	1.04	7.5×10^{-6}	Two-component system response regulator	*Bacteroides vulgatus*
amy:ADJ76_04750	0.915	9.2×10^{-6}	ABC transporter substrate-binding protein	*Schaalia meyeri*
amy:ADJ76_03175	0.916	1.2×10^{-5}	Hypothetical protein	*Schaalia meyeri*
bpsc:BBPC_0350	−1.03	1.2×10^{-5}	30S ribosomal protein S14	*Bifidobacterium pseudocatenulatum*
amy:ADJ76_06025	0.970	1.5×10^{-5}	Preprotein translocase subunit SecE	*Schaalia meyeri*
stu:STH8232_1189	1.09	1.5×10^{-5}	Hypothetical protein SMU.1171c	*Streptococcus thermophilus JIM 8232*

KEGG, Kyoto Encyclopedia of Genes and Genomes.

samples compared with controls. One gene demonstrated a highly significant association that satisfied the empirical Bonferroni's correlation ($p=2.5\times10^{-7}$, effect size=−1.19, R6FCZ7). R6FCZ7 is registered as a FeS assembly sulphur utilisation factor system protein in the UniRef database; it has not yet been given an official gene name by nomenclature committees. This kind of protein performs a wide range of bacterial functions, such as electron transfer, redox catalysis and gene regulation.[26] Several previous studies have reported that reactive oxygen species play an important role in the pathogenesis of RA,[27] and Zhang *et al.*[8] also reported alteration of the redox environment in the RA gut microbiome. In taxonomic assignment, the source strain of R6FCZ7 was registered as *Bacteroides sp. CAG:633*. In our metagenome data, the R6FCZ7 sequences were further linked to the taxonomic reference genomes of *Bacteroides uniformis*, *B. rodentium* and *B. fragilis*, as well as *Bacteroides sp.* This implies that several species belonging to the genus *Bacteroides* functionally possess this gene. Overall, our results suggest that the redox function of the microbiome, especially the genus *Bacteroides*, may have an important role in the pathology of RA.

Identification of metagenomic biological pathways contributing to the pathophysiology of RA

Using the results of the gene analysis part of our MWAS, we performed gene set enrichment analysis to conduct case–control pathway association tests. We found significant associations for 19 Gene Ontology (GO) terms that satisfied the empirical Bonferroni's correlation (figure 3A, table 3). One of the significant GO terms was metal ion binding ($p=2.0\times10^{-5}$, GO:0046872), which implies that the interaction between reactive oxygen species and metal ions is associated with the pathology of RA, which is similar to the result of the gene association tests. The eight KEGG pathways showed significant associations ($q<0.05$; figure 3B). A part of the substances related to these pathways are involved in the pathophysiology of RA as follows: (1) Fatty acids have been found to be related to inflammation and several free fatty acids in serum are reported to be upregulated in patients with RA.[28] (2) Terpenoids suppress nuclear factor-kB signalling, the major regulator in the pathogenesis of inflammatory diseases.[29] They are the main component of *Tripterygium wilfordii*, a plant used in traditional Chinese medicine which has been shown to be non-inferior to methotrexate in the treatment of RA.[30] (3) Adipocytokines have been reported to contribute to the proinflammatory state of RA, increasing in the synovial fluid of patients with RA.[31] (4) Glycosaminoglycans (GAGs) are major components of joint cartilage and other soft connective tissues. Modification of GAG metabolism may play a crucial role in RA pathogenesis.[32] These results suggest that the aforementioned substances in the gut could influence the pathology of RA. The system diagram of the KEGG pathways with the results of the case–control pathway association tests clearly illustrates that a variety of pathways, including those related to metabolism, showed significant case–control differences (figure 3C).

Population-specific shared biological pathways between the metagenome and host genome in RA

We assessed whether biological pathways of the gut metagenome and the host germline genome were shared in RA. In addition to our KEGG pathway enrichment of the RA gut metagenome of the Japanese population, we estimated pathway enrichment in the host genome of RA samples, using previously conducted RA GWAS data in East Asian and European populations ($n=22\,515$ and $n=58\,284$, respectively).[18] We note that the majority of subjects in the East Asian RA GWAS were of Japanese ancestry (93.1%). We compared the p values of the KEGG pathways shared between the RA MWAS data (Japanese) and the RA GWAS data (separately for East Asians and Europeans; figure 3D). Several pathways showed significant enrichment between the metagenomes and host genomes ($p<0.05$ in both; e.g., FoxO signalling pathway, Th17 cell differentiation and interleukin-17 signalling pathway). We observed a significant correlation between pathway p values in East Asians ($p_{\text{Fisher}}=0.045$), but not in Europeans ($p_{\text{Fisher}}=0.41$). Considering that MWAS–GWAS pathway correlation was specifically observed when the metagenome and the host genome data in similar populations were compared, there should be a population-specific link between the germline genome and metagenome in RA pathology.

No apparent discrepancy in metagenome diversity between RA cases and controls

The microbial diversity of the RA gut microbiome is still controversial. We thus evaluated alpha and beta diversity in the phylogenetic data (phylogenetic relative abundance of six

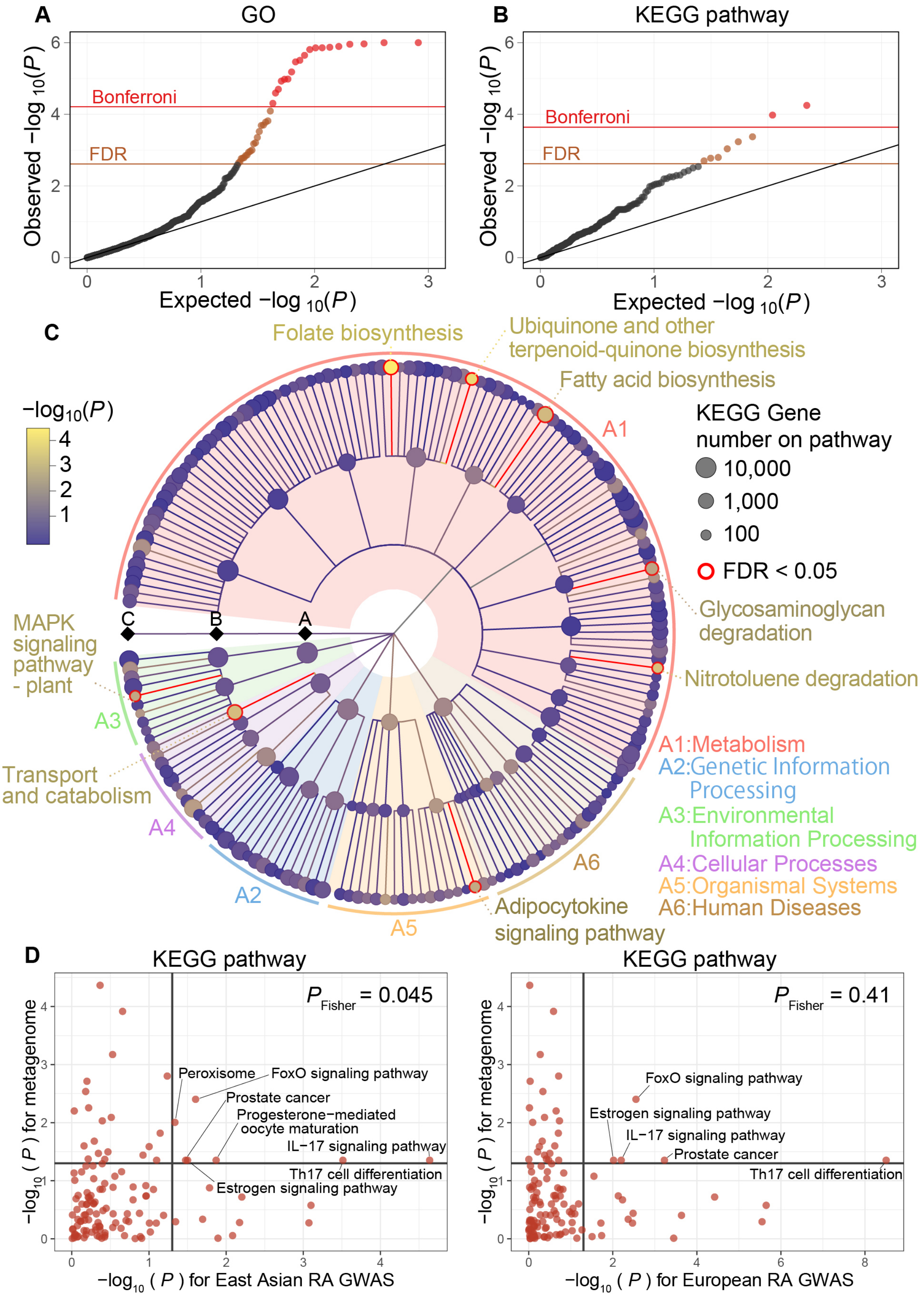

Figure 3 MWAS results of RA case–control pathway association tests. (A) QQ plot of the MWAS p values of pathways based on GO terms. (B) A QQ plot of the MWAS p values of pathways based on GO terms. KEGG pathways. (C) System diagram of KEGG pathways. The three levels are defined as A, B and C, and described from the inside layer out. The size and colour of dots represent set sizes and p values, respectively. The eight pathways with significant enrichment ($q<0.05$) are outlined in red. (D) Comparison of p values of KEGG pathways between the RA MWAS and GWAS data. The *x*-axis indicates the p values of the GWAS data (left, East Asians; right, Europeans). The *y*-axis indicates the p values of the MWAS of Japanese. The horizontal and vertical black lines indicate *p* value of 0.05. The overlap of the pathway enrichment was evaluated by classifying the pathways based on the significance threshold of $p<0.05$ or $p\geq0.05$ and using Fisher's exact test. FDR, false discovery rate; GO, Gene Ontology; GWAS, genome-wide association study; KEGG, Kyoto Encyclopedia of Genes and Genomes; MAPK, mitogen-activated protein kinase; MWAS, metagenome-wide association study; QQ, quantile-quantile; RA, rheumatoid arthritis.

Table 3 Pathways with significant case–control discrepancy

GO term	Set size	*p*-value	*q*-value	Name
GO:0005886	7320	1.0×10^{-6}	1.3×10^{-4}	Plasma membrane
GO:0006810	4484	1.0×10^{-6}	1.3×10^{-4}	Transport
GO:0004872	1062	1.1×10^{-6}	1.3×10^{-4}	Receptor activity
GO:0009279	867	1.1×10^{-6}	1.3×10^{-4}	Cell outer membrane
GO:0005829	305	1.3×10^{-6}	1.3×10^{-4}	Cytosol
GO:0030288	206	1.4×10^{-6}	1.3×10^{-4}	Outer membrane-bounded periplasmic space
GO:0042597	162	1.4×10^{-6}	1.3×10^{-4}	Periplasmic space
GO:0071973	154	1.4×10^{-6}	1.3×10^{-4}	Bacterial-type flagellum-dependent cell motility
GO:0042802	59	1.6×10^{-6}	1.3×10^{-4}	Identical protein binding
GO:0030246	754	2.3×10^{-6}	1.7×10^{-4}	Carbohydrate binding
GO:0016485	59	3.1×10^{-6}	2.1×10^{-4}	Protein processing
GO:0016757	658	3.4×10^{-6}	2.2×10^{-4}	Transferase activity, transferring glycosyl groups
GO:0016020	959	6.6×10^{-6}	3.8×10^{-4}	Membrane
GO:0048038	93	1.0×10^{-5}	5.3×10^{-4}	Quinone binding
GO:0015031	242	1.1×10^{-5}	5.3×10^{-4}	Protein transport
GO:0005887	526	1.2×10^{-5}	5.6×10^{-4}	Integral component of plasma membrane
GO:0046872	6747	2.0×10^{-5}	8.9×10^{-4}	Metal ion binding
GO:0008152	2314	2.5×10^{-5}	0.0011	Metabolic process
GO:0006518	64	5.0×10^{-5}	0.0020	Peptide metabolic process

KEGG pathway	Set size	p-value	q-value	Definition
ko00790	740	4.3×10^{-5}	0.0080	Folate biosynthesis
ko00130	175	1.2×10^{-4}	0.011	Ubiquinone and other terpenoid-quinone biosynthesis
ko00633	108	4.1×10^{-4}	0.025	Nitrotoluene degradation
B21	580	5.5×10^{-4}	0.025	Transport and catabolism
ko00061	834	6.7×10^{-4}	0.025	Fatty acid biosynthesis
ko04920	118	0.0016	0.045	Adipocytokine signalling pathway
ko04016	87	0.0017	0.045	MAPK signalling pathway-plant
ko00531	257	0.0020	0.045	Glycosaminoglycan degradation

GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

levels (L2–L7)) and functional data (gene abundance based on UniRef90 protein and KEGG gene databases). No significant difference was found in any case–control comparisons of alpha diversity based on the Shannon index ($p>0.05$; figure 4A and B). As for beta diversity, no level of phylogenetic data showed significant case–control differences in either sequencing group ($p>0.05$; figure 4C). The gene abundance data also did not show significant differences ($p>0.05$; only UniRef90 data shown in figure 4D). Overall, there was no apparent discrepancy in metagenome diversity between RA cases and controls.

DISCUSSION

Whole-genome shotgun sequencing of the metagenome has been increasing in importance as more attention is paid to the association between diseases and microbiomes. In our study, we conducted a comprehensive MWAS by using whole-genome shotgun sequencing. We found that the RA gut metagenome of the studied Japanese population has the following interesting characteristics: (1) Multiple species belonging to the *Prevotella* genus increased in the RA gut metagenome. (2) Non-linear ML methods such as UMAP could be used to successfully deconvolute case–control phylogenetic discrepancies. (3) The abundance of one gene related to redox reaction was decreased in the RA metagenome. (4) Various biological pathways, including those related to metabolism, were enriched in a case–control comparison. (5) A population-specific pathway link between the metagenome (MWAS) and the host genome (GWAS) was identified. (6) No apparent discrepancy in microbial diversity was found between RA cases and controls. Our study greatly exploited the benefits of shotgun sequencing, since it would be challenging to obtain such novel findings using the classical method of 16S rRNA.

Our study is the first to robustly demonstrate that multiple *Prevotella spp.* other than *P. copri* increased in the RA gut microbiome by using shotgun sequencing. Recently, Alpizar-Rodriguez *et al.* reported that *Prevotella spp.* other than *P. copri* were enriched in preclinical RA stages, although their assessment was limited to operational taxonomic units assigned at the species level because they used the classical 16S rRNA method and did not conduct statistical analysis.[33] Increased levels of microbes of oral origin, including *Prevotella*, have been reported to be correlated with several diseases.[34 35] Atarashi *et al.*[36] reported that strains of *Klebsiella spp.* isolated from the salivary microbiota were strong inducers of T helper 1 cells when they colonised the gut.[36] In our study, most species of *Prevotella* showing significant case–control differences were taxa that have been identified in the oral cavity. All species included in the taxa cluster with case–control discrepancies identified by unsupervised ML methods

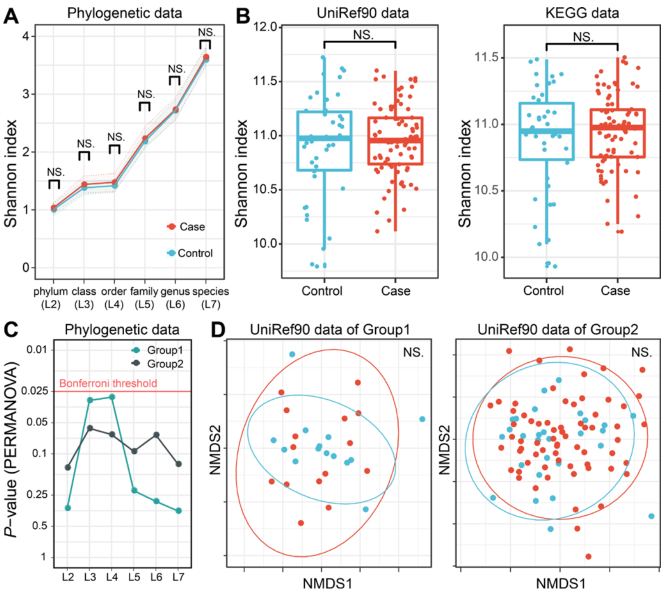

Figure 4 Case–control comparison of microbial diversities in RA. (A) Alpha diversities of the phylogenetic relative abundance data for six levels. Welch's t-test of Shannon index between RA cases and controls showed no significant difference at any level. (B) Alpha diversities of the gene abundance data of the UniRef90 protein and KEGG gene databases. No significant case–control difference was found. (C) Beta diversities of phylogenetic relative abundance data at six levels. PERMANOVA based on Bray–Curtis dissimilarities found no significant differences among levels for either sequencing group with Bonferroni correction. (D) Beta diversities of the gene abundance of the UniRef90 protein database. No significant case–control difference was found. KEGG, Kyoto Encyclopedia of Genes and Genomes; NMDS, non metric multidimensional scaling; PERMANOVA; permutational multivariate analysis of variance; RA, rheumatoid arthritis.

were registered in the expanded Human Oral Microbiome Database.[37] The colonisation of the intestine by oral bacteria could be related to the pathogenesis of RA as well as other diseases.

The gene and pathway elements of our MWAS analysis successfully demonstrated the novel functional aspects of the RA gut metagenome. A representative finding was the decrement of a redox reaction-related gene of the genus *Bacteroides* in RA. It has previously been reported that *Prevotella* and *Bacteroides* in the gut in RA showed inverse relationships in terms of their proportional quantity.[11] The increase in several species of the genus *Prevotella* in RA samples further suggests that the balance between these two major taxa in genetic function and composition reflects the disease-specific features of the RA gut metagenome. There could be a possibility that R6FCZ7 and the genus *Prevotella* were inversely associated via the relationship with the genus *Bacteroides*, while further functional investigation was required.

The overall association between the metagenome and host genome has mostly been investigated in healthy people[38–40]; there are few studies focusing on specific diseases. Our MWAS–GWAS interaction analysis demonstrated the population-specific pathway link between the germline genome and metagenome in RA pathology. Further studies investigating biological roles of the detected pathways as well as validating these findings in other independent populations, or in other diseases, are warranted. Furthermore, considering the substantial roles of the human leukocyte antigen (HLA) phenotypes in RA aetiology, the study to investigate the link between HLA phenotypes and metagenome diversity would be also warranted.

To date, discussions on the microbial diversity of the RA gut microbiome have been controversial; some reports have indicated significant differences,[10 11] while others have shown no differences.[8 41] Recently, a comprehensive analysis of the microbial diversity in many diseases (not including RA) indicated that there were no significant differences in most diseases relative to controls.[42] Our results indicate that the same is true for RA, and that the gut microbiome dysbiosis caused by RA would not be sufficient to affect overall microbial diversity. Nevertheless, it would be important how to treat multimapped reads for accurately calculating relative abundances, which is a challenge for further study.

In conclusion, our shotgun sequencing-based comprehensive MWAS in the Japanese RA population revealed a novel link between the gut microbiome, host genome and pathology of RA. In addition to providing novel insights into the RA gut microbiome, our study will provide useful resources for future functional investigations to further elucidate details of the microbiome's role in RA aetiology.

Author affiliations
[1]Department of Statistical Genetics, Osaka University School of Medicine Graduate School of Medicine, Suita, Japan
[2]Department of Otorhinolaryngology - Head and Neck Surgery, Osaka University Graduate School of Medicine, Suita, Japan
[3]Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine, Suita, Japan
[4]Laboratory of Immune Regulation, Department of Microbiology and Immunology, Osaka University Graduate School of Medicine, Suita, Japan
[5]Department of Infection Metagenomics, Research Institute for Microbial Diseases, Osaka University, Suita, Japan
[6]Department of Rheumatology and Allergology, Saiseikai Senri Hospital, Suita, Japan
[7]Rheumatology and Allergology, NHO Osaka Minami Medical Center, Kawachinagano, Japan
[8]Division of Rheumatology, Department of Internal Medicine, Daini Osaka Police Hospital, Tennoji-ku, Japan
[9]Clinical Research, NHO Osaka Minami Medical Center, Kawachinagano, Japan
[10]Department of Immunopathology, Immunology Frontier Research Center, Osaka University, Suita, Japan
[11]WPI Immunology Frontier Research Center, Osaka University, Suita, Japan
[12]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan
[13]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan

Acknowledgements We thank Ms Miho Kaneda for supporting the study.

Contributors TK and YuO designed the study, conducted the data analysis and wrote the manuscript. YMae, TN, DM, YMat and SN conducted the experiments. YMae, TN, MM, HM, MY, SK, ST, EO, YaO, KK, SH, TH, MN, AO and YS collected the samples. HI, AK, KT and YuO supervised the study.

Funding This research was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (15H05911, 19H01021), the Japan Agency for Medical Research and Development (AMED; 19gm6010001h0004, 19ek0410041h0003, 19ek0109413h0001, 19km0405211h0001), Takeda Science Foundation, Bioinformatics Initiative of Osaka University Graduate School of Medicine, and Grant Program for Next Generation Principal Investigators at Immunology Frontier Research Center (WPI-IFReC), Osaka University.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The whole-genome shotgun sequencing data are deposited in National Bioscience Database Center (NBDC) Human Database (http://humandbs.biosciencedbc.jp/) with the accession number of hum0197. The data are available upon reasonable request.

ORCID iDs
Toru Hirano http://orcid.org/0000-0001-8467-3154
Yukihiko Saeki http://orcid.org/0000-0003-3870-0275
Yukinori Okada http://orcid.org/0000-0002-0311-8472

REFERENCES

1 Levy M, Kolodziejczyk AA, Thaiss CA, *et al*. Dysbiosis and the immune system. *Nat Rev Immunol* 2017;17:219–32.
2 Forslund K, Hildebrand F, Nielsen T, *et al*. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 2015;528:262–6.
3 Franzosa EA, Sirota-Madi A, Avila-Pacheco J, *et al*. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* 2019;4:293–305.
4 Tilg H, Adolph TE, Gerner RR, *et al*. The intestinal microbiota in colorectal cancer. *Cancer Cell* 2018;33:954–64.
5 Tang WHW, Wang Z, Levison BS, *et al*. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N Engl J Med* 2013;368:1575–84.
6 Okada Y, Eyre S, Suzuki A, *et al*. Genetics of rheumatoid arthritis: 2018 status. *Ann Rheum Dis* 2019;78:446–53.
7 Scher JU, Abramson SB. The microbiome and rheumatoid arthritis. *Nat Rev Rheumatol* 2011;7:569–78.
8 Zhang X, Zhang D, Jia H, *et al*. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* 2015;21:895–905.
9 Scher JU, Joshua V, Artacho A, *et al*. The lung microbiota in early rheumatoid arthritis and autoimmunity. *Microbiome* 2016;4.
10 Chen J, Wright K, Davis JM, *et al*. An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med* 2016;8:43.
11 Scher JU, Sczesnak A, Longman RS, *et al*. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *eLife* 2013;2:e01202.
12 Lopez-Oliva I, Paropkari AD, Saraswat S, *et al*. Dysbiotic subgingival microbial communities in periodontally healthy patients with rheumatoid arthritis. *Arthritis Rheumatol* 2018;70:1008–13.
13 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.
14 Scher JU, Ubeda C, Equinda M, *et al*. Periodontal disease and the oral microbiota in new-onset rheumatoid arthritis. *Arthritis Rheum* 2012;64:3083–94.
15 Wegner N, Wait R, Sroka A, *et al*. Peptidylarginine deiminase from Porphyromonas gingivalis citrullinates human fibrinogen and α-enolase: implications for autoimmunity in rheumatoid arthritis. *Arthritis Rheum* 2010;62:2662–72.
16 Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci* 2014;5:209.

17 Ranjan R, Rani A, Metwally A, *et al*. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun* 2016;469:967–77.

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

19 Karlsson FH, Tremaroli V, Nookaew I, *et al*. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 2013;498:99–103.

20 Thaiss CA, Itav S, Rothschild D, *et al*. Persistent microbiome alterations modulate the rate of post-dieting weight regain. *Nature* 2016;540:544–51.

21 McInnes L, Healy J, Saul N, *et al*. UMAP: uniform manifold approximation and projection. *JOSS* 2018;3.

22 Hirata J, Hosomichi K, Sakaue S, *et al*. Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat Genet* 2019;51:470–80.

23 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.

24 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.

25 Kanehisa M, Goto S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30.

26 Blanc B, Gerez C, Ollagnier de Choudens S. Assembly of Fe/S proteins in bacterial systems: biochemistry of the bacterial ISC system. *Biochim Biophys Acta* 1853;2015:1436–47.

27 Filippin LI, Vercelino R, Marroni NP, *et al*. Redox signalling and the inflammatory response in rheumatoid arthritis. *Clin Exp Immunol* 2008;152:415–22.

28 Zhou J, Chen J, Hu C, *et al*. Exploration of the serum metabolite signature in patients with rheumatoid arthritis using gas chromatography–mass spectrometry. *J Pharm Biomed Anal* 2016;127:60–7.

29 Salminen A, Lehtonen M, Suuronen T, *et al*. Terpenoids: natural inhibitors of NF-κB signaling with anti-inflammatory and anticancer potential. *Cell. Mol. Life Sci.* 2008;65:2979–99.

30 Lv Q-W, Zhang W, Shi Q, *et al*. Comparison of Tripterygium wilfordii hook F with methotrexate in the treatment of active rheumatoid arthritis (TRIFRA): a randomised, controlled clinical trial. *Ann Rheum Dis* 2015;74:1078–86.

31 Ruscitti P, Di Benedetto P, Berardicurti O, *et al*. Adipocytokines in rheumatoid arthritis: the hidden link between inflammation and cardiometabolic comorbidities. *Journal of Immunology Research* 2018;2018:1–10.

32 Szeremeta A, Jura-Półtorak A, Koźma EM, *et al*. Effects of a 15-month anti-TNF-α treatment on plasma levels of glycosaminoglycans in women with rheumatoid arthritis. *Arthritis Res Ther* 2018;20.

33 Alpizar-Rodriguez D, Lesker TR, Gronow A, *et al*. *Prevotella copri* in individuals at risk for rheumatoid arthritis. *Ann Rheum Dis* 2019;78:590–3.

34 Qin N, Yang F, Li A, *et al*. Alterations of the human gut microbiome in liver cirrhosis. *Nature* 2014;513:59–64.

35 Sears CL, Garrett WS, Microbes GWS. Microbes, microbiota, and colon cancer. *Cell Host Microbe* 2014;15:317–28.

36 Atarashi K, Suda W, Luo C, *et al*. Ectopic colonization of oral bacteria in the intestine drives T $_{H}$ 1 cell induction and inflammation. *Science* 2017;358:359–65.

37 Escapa IF, Chen T, Huang Y, *et al*. New insights into human Nostril microbiome from the expanded human oral microbiome database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *mSystems* 2018;3.

38 Bonder MJ, Kurilshikov A, Tigchelaar EF, *et al*. The effect of host genetics on the gut microbiome. *Nat Genet* 2016;48:1407–12.

39 Turpin W, Espin-Garcia O, Xu W, *et al*. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat Genet* 2016;48:1413–7.

40 Blekhman R, Goodrich JK, Huang K, *et al*. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* 2015;16:191.

41 Picchianti-Diamanti A, Panebianco C, Salemi S, *et al*. Analysis of gut microbiota in rheumatoid arthritis patients: disease-related dysbiosis and modifications induced by etanercept. *Int J Mol Sci* 2018;19:2938.

42 Ma Z, Li L, Gotelli NJ. Diversity-disease relationships and shared species analyses for human microbiome-associated diseases. *Isme J* 2019;13:1911–9.