OPEN ACCESS

TRANSLATIONAL SCIENCE

Metagenome-wide association study revealed disease-specific landscape of the gut microbiome of systemic lupus erythematosus in Japanese

Yoshihiko Tomofuji,[1] Yuichi Maeda ,[2,3,4] Eri Oguro-Igashira ,[2,3] Toshihiro Kishikawa,[1,5,6] Kenichi Yamamoto,[1,7,8] Kyuto Sonehara,[1,4] Daisuke Motooka,[9] Yuki Matsumoto,[9] Hidetoshi Matsuoka,[10] Maiko Yoshimura,[10] Mayu Yagita,[2,3] Takuro Nii,[2,3] Shiro Ohshima,[10] Shota Nakamura,[4,9] Hidenori Inohara,[5] Kiyoshi Takeda,[3,11] Atsushi Kumanogoh,[2,4,12] Yukinori Okada [1,4,8]

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/annrheumdis-2021-220687).

For numbered affiliations see end of article.

Correspondence to
Dr Yukinori Okada, Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan;
yokada@sg.med.osaka-u.ac.jp

YT, YM and EO-I contributed equally.

Received 29 April 2021
Accepted 20 July 2021
Published Online First 23 August 2021

To cite: Tomofuji Y, Maeda Y, Oguro-Igashira E, *et al*. *Ann Rheum Dis* 2021;**80**:1575–1583.

ABSTRACT

Objective Alteration of the gut microbiome has been linked to the pathogenesis of systemic lupus erythematosus (SLE). However, a comprehensive view of the gut microbiome in SLE and its interaction with the host remains to be revealed. This study aimed to reveal SLE-associated changes in the gut microbiome and its interaction with the host by a comprehensive metagenome-wide association study (MWAS) followed by integrative analysis.

Methods We performed a MWAS of SLE based on shotgun sequencing of the gut microbial DNA from Japanese individuals (N_{case}=47, $N_{control}$=203). We integrated the result of the MWAS with the genome-wide association study (GWAS) data and plasma metabolite data.

Results Via species level phylogenetic analysis, we identified and validated increases of *Streptococcus intermedius* and *Streptococcus anginosus* in the patients with SLE. Microbial gene analysis revealed increases of *Streptococcus*-derived genes including one involved in redox reaction. Additionally, microbial pathways related to sulfur metabolism and flagella assembly were altered in the patients with SLE. We identified an overlap in the enriched biological pathways between the metagenome and the germline genome by comparing the result of the MWAS and the GWAS of SLE (ie, MWAS-GWAS interaction). α-diversity and β-diversity analyses provided evidence of dysbiosis in the metagenome of the patients with SLE. Microbiome-metabolome association analysis identified positive dosage correlation of acylcarnitine with *Streptococcus intermedius*, an SLE-associated taxon.

Conclusion Our MWAS followed by integrative analysis revealed SLE-associated changes in the gut microbiome and its interaction with the host, which contribute to our understanding of the relationship between the microbiome and SLE.

Key messages

What is already known about this subject?

- Alteration of the gut microbiome has been linked to the pathogenesis of systemic lupus erythematosus (SLE), but a comprehensive view of the gut microbiome in SLE and its interaction with the host remains to be revealed.
- Whole metagenome shotgun sequencing technology has many advantages over conventional 16S ribosomal RNA sequencing such as higher taxonomic resolution and applicability for the functional analysis. However, evaluation of the microbiome-disease association based on shotgun sequencing is still incomplete for SLE.

What does this study add?

- Our shotgun sequence-based metagenome-wide association study (MWAS) newly identified two bacterial species (*Streptococcus anginosus* and *Streptococcus intermedius*), eight bacterial genes and seven biological pathways which were significantly different between the healthy controls and the patients with SLE.
- Integrative analysis with the genome-wide association study (GWAS) result and the plasma metabolite data revealed the interactions between the gut microbiome and the host mediated by biological pathways or plasma metabolites.

How might this impact on clinical practice or future developments?

- Our shotgun sequencing-based MWAS and integrative analysis with the GWAS data and plasma metabolite profiles revealed an SLE-specific microbial landscape and its association with the host. Our analysis contributes to our understanding of the relationship between the gut microbiome and SLE.

INTRODUCTION

Systemic lupus erythematosus (SLE) is an autoimmune disease, which is characterised by overactivation of the immune system and involvement of various organs such as kidney and brain. Despite advances in treatment, standardised mortality rates in patients with SLE were three times higher than in the general populations because of poor control of the disease activity or infection due to immunosuppressive treatment.[1] SLE results from a complex interplay of multiple genetic and environmental factors; however, much of the aetiology of SLE

remain to be elucidated. Therefore, extensive efforts have been spent to reveal the pathogenesis of SLE for the development of better clinical care.

Microbiome, which refers to the microbial communities inhabiting the human body, has a remarkable effect on our body by modulating our immune system or taking a part of our metabolic network.[2] The largest community of the human microbiota resides within the gut. Involvement of the gut microbiome is reported for various diseases, such as type 2 diabetes,[3] colorectal cancer,[4] rheumatoid arthritis (RA),[5 6] inflammatory bowel disease (IBD)[7], and multiple sclerosis.[8] Accompanied by the great progresses in high-throughput sequencing technology and success of the treatment such as faecal microbiome transplantation and probiotics, characterisation of the gut microbiome has become a major research area in human diseases.

Recently, relationship between the gut microbiome and SLE was studied to reveal an unexplained part of the SLE aetiology. Faecal microbiome from an SLE model mouse has capacity to induce SLE-like phenotype in a healthy mouse underlying the non-trivial relationship between the gut microbiome and the SLE pathogenesis.[9] Mechanistic insights of association between the gut microbiome and SLE have been obtained by mouse experiments (e.g, activation of the immune system caused by bacterial translocation from gut to liver[10] or microbe derived metabolites[11]). In human, SLE-associated taxa were searched through amplicon sequencing of 16S ribosomal RNA (rRNA) genes.[12–16] Although the findings were not universally consistent, reflecting difference in ethnicities and lifestyles, several SLE-associated clades were identified. However, 16S rRNA sequencing typically provides phylogenetic abundance at up to the genus level, making it suffer from low taxonomic resolution. In addition, 16S rRNA sequencing technology can only be used for phylogenetic analysis; thus, functional aspect of disease-specific microbial environment is overlooked.

In gut microbiome case-control study, metagenome-wide association studies (MWAS) based on whole-genome shotgun sequencing are replacing case-control comparison with 16S rRNA sequencing technology. Shotgun sequencing has a potential to detect the genomic composition of microbes at the species level, achieving higher taxonomic resolution than 16S rRNA sequencing technology. Furthermore, shotgun sequencing can be used for analysing microbial gene and pathway and is therefore useful for surveying the functional aspect of microbial environment. However, shotgun sequencing requires much larger sequencing cost and machine resource than 16S rRNA-sequencing. Additionally, analytic methods applied to shotgun sequencing data are usually complicated. For these reasons, evaluation of microbiome-disease association based on shotgun sequencing is still incomplete for many diseases including SLE.[17] Furthermore, insufficient number of shotgun sequencing studies in non-European population is problematic given the significant impact of ethnicity and lifestyle on the microbial landscape.[18 19]

Even with shotgun sequencing analysis, microbiome-host interaction is hardly evaluated unless performing integrative analysis with other modality data such as metabolic profiles. A large part of microbiome-host interaction is estimated to be mediated by metabolic signals.[2] Multiomics analysis based on microbial and metabolic data was performed in not many but several diseases such as IBD[7] and has revealed the involvement of metabolites in microbiome-disease association. Integrative analysis with the genome-wide association study (GWAS) is also useful for revealing a link between the gut microbiome and the host genome, namely MWAS-GWAS interaction.[5] Nonetheless, microbiome-host interaction in SLE has been never evaluated, hindering us from the comprehensive understanding of the microbiome-associated SLE pathology.

In this study, we carried out shotgun sequencing of faecal samples from 250 Japanese subjects, composed of 47 patients with SLE and 203 healthy controls (HCs). To identify SLE-associated microbes, we performed phylogenetic case-control comparison. We also performed microbial gene analysis followed by pathway analysis for revealing functional differences of the gut metagenome between the HCs and the patients with SLE. To reveal the microbiome-host interaction in SLE, we performed a combined biological pathway analysis of MWAS and GWAS. In addition, we performed an integrative analysis using plasma metabolite profiles obtained through the non-targeted metabolomics approach. The joint study of microbiome and metabolome can identify the functional readouts of disease-specific microbial activity which mediates the microbe-host interaction.

METHODS

Methods are provided in the online supplemental information.

RESULTS

High abundance of *Streptococcus anginosus* and *Streptococcus intermedius* in the SLE gut microbiome

We performed whole-genome shotgun sequencing analysis of a total of 250 faecal DNA samples (47 individuals with SLE and 203 HC subjects) derived from three sequencing groups (online supplemental table 1), which passed stringent quality control (QC) for sequence reads and samples. Procedures of sample QC and definition of the sample sets in each analysis are described in online supplemental figure 1. Then, we obtained phylogenetic relative abundances (online supplemental figure 3). For case-control comparison, we performed clade QC. After clade QC, we had 774 clades for case-control association test, including 12 phyla (L2), 25 classes (L3), 36 orders (L4), 72 families (L5), 178 genera (L6) and 451 species (L7).

We performed case-control comparison for each clade and identified that *Streptococcus anginosus* and *Streptococcus intermedius* significantly increased in SLE (effect size=0.617 and $P_{\text{microbe}}=3.7\times10^{-5}$ for *Streptococcus anginosus*, effect size=0.579 and $P_{\text{microbe}}=7.5\times10^{-5}$ for *Streptococcus intermedius*; figure 1A,B and table 1), satisfying an empirically estimated Bonferroni threshold ($\alpha=0.05$; $P_{\text{microbe}}<8.2\times10^{-5}$). As illustrated in a phylogenetic tree indicating the case-control association results of multilayered taxonomic levels (figure 1C), both of the clades with case-control discrepancy were species (L7) level. Since it was difficult to detect the species-level clades using classical 16S rRNA sequencing, our results underlay the strength of MWAS approach with shotgun sequencing to identify disease-associated microbial taxa.

As medication of the patients with SLE and male-female imbalance due to sex biased prevalence could be a confounding factor, we performed subanalysis (online supplemental table 2). Effect sizes were almost similar among subanalyses for the *Streptococcus anginosus* and *Streptococcus intermedius* ($0.487\leq$effect size≤0.647 for *Streptococcus anginosus*, $0.463\leq$effect size≤0.654 for *Streptococcus intermedius* (online supplemental figure 4 and table 2) after removing male subjects or those who took medications such as proton pump inhibitor, antibiotics or therapeutics for SLE. These results suggested that inclusion of the male subjects or those who took medications such as proton pump inhibitor, antibiotics or therapeutics minimally confounded the result but increased statistical power of the MWAS. The abundance of these two clades was not significantly different between the newly onset patients and the other patients, the patients with lupus nephritis (LN) and without

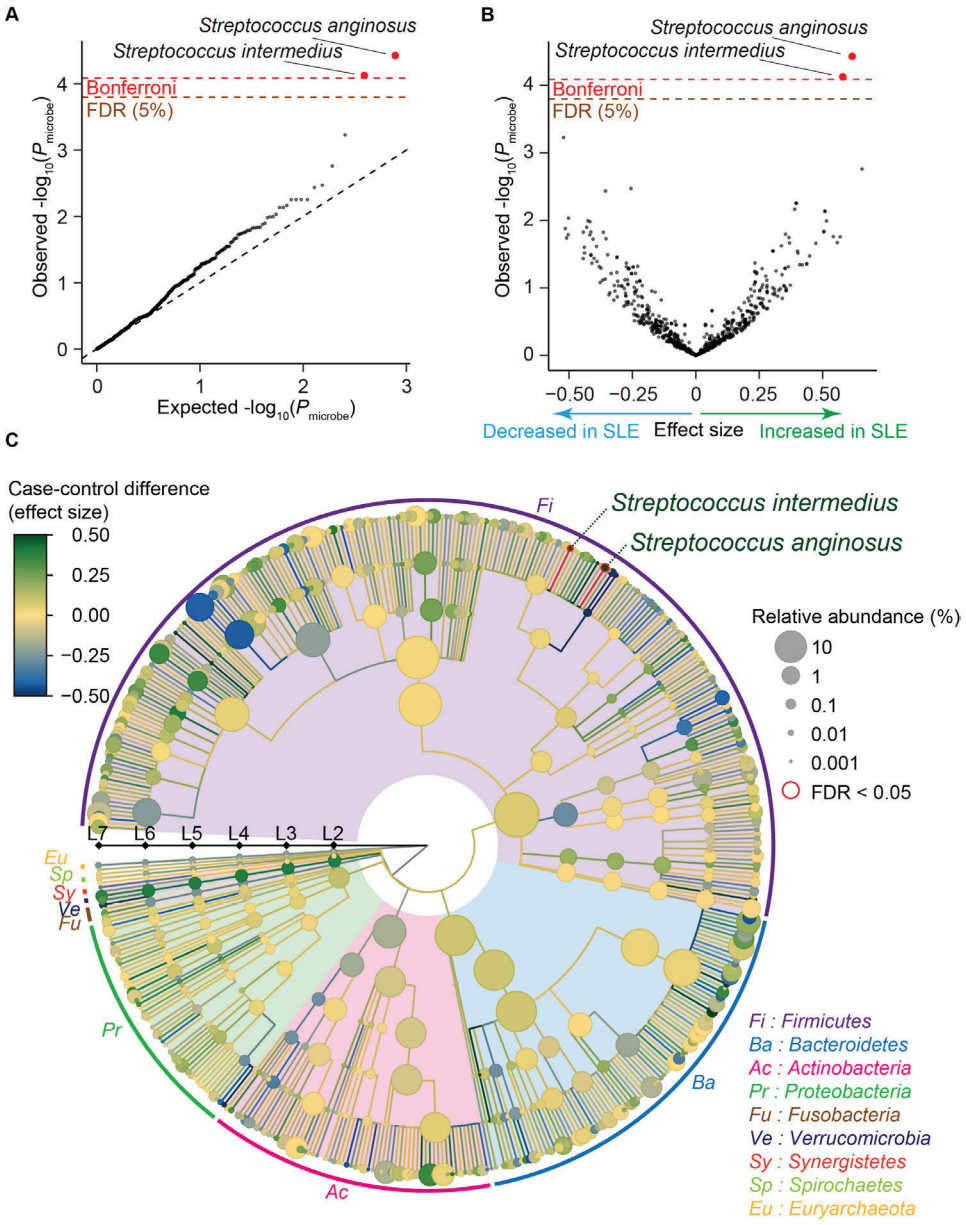


Figure 1 Result of the SLE MWAS based on the phylogenetic abundance data. (A) A quantile-quantile plot of the phylogenetic MWAS p values ($P_{microbe}$) of the clades. The x-axis indicates log-transformed empirically estimated median $P_{microbe}$. The y-axis indicates observed $-\log_{10}(P_{microbe})$. The diagonal dashed 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 threshold (FDR=0.05). Clades with $P_{microbe}$ less than the Bonferroni thresholds are plotted as red dots, and other clades are plotted as black dots. (B) A volcano plot. The x-axis indicates effect sizes in linear regression. The y-axis, horizontal lines and dot colours are the same as in (A). (C) A phylogenetic tree. Levels L2–L7 are from the inside layer to the outside layer. The size and the colour of dots represent relative abundances and effect sizes, respectively. The two clades with significant case-control associations (FDR<0.05) are outlined in red. FDR, false discovery ratio; MWAS, metagenome-wide association study; SLE, systemic lupus erythematosus.

LN or the patients with high SLE Disease Activity Index (SLE-DAI) and low SLE-DAI ($P_{microbe}$ >0.12; online supplemental table 2). We performed replication analysis for *Streptococcus anginosus* and *Streptococcus intermedius* by using a previously published shotgun sequencing study.[17] Association of *Streptococcus anginosus* and *Streptococcus intermedius* was significantly replicated (same effect direction and $P_{microbe}$ <0.05/2=0.025), confirming the associations identified by our SLE MWAS.

Table 1 Clades with significant case-control discrepancy in the SLE MWAS

		This study (Japanese, *N*=250)			Chen *et al* (Chinese, *N*=232)	
Microbe	**Level**	**Effect size**	**SE**	$P_{microbe}$	**Effect direction**	$P_{microbe}$
Streptococcus anginosus	Species (L7)	0.617	0.146	3.7×10^{-5}	Positive	1.9×10^{-6}
Streptococcus intermedius	Species (L7)	0.579	0.143	7.5×10^{-5}	Positive	9.1×10^{-4}

MWAS, metagenome-wide association study; SLE, systemic lupus erythematosus.

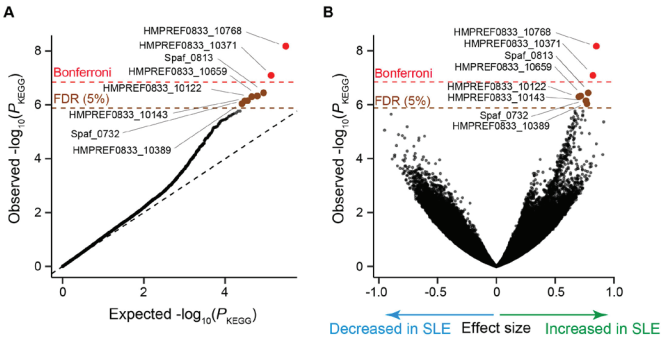


Figure 2 Result of the SLE MWAS based on the microbial gene abundance data. (A) A quantile-quantile plot of the MWAS p values of the genes (P_{KEGG}). The x-axis indicates log-transformed empirically estimated median P_{KEGG}. The y-axis indicates observed $-\log_{10}(P_{KEGG})$. The diagonal dashed 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 threshold (FDR=0.05). Genes with P_{KEGG} less than the Bonferroni thresholds are plotted as red dots. Genes with FDR<0.05 are plotted as brown dots, and other clades are plotted as black dots. (B) A volcano plot. The x-axis indicates effect sizes in linear regression. The y-axis, horizontal lines and dot colours are the same as in (A). FDR, false discovery ratio; KEGG, Kyoto Encyclopedia of Genes and Genomes; MWAS, metagenome-wide association study; SLE, Systemic lupus erythematosus.

High abundance of *Streptococcus*-derived genes in the gut metagenome of patients with SLE

We next performed a gene level MWAS. We obtained microbial gene abundance data based on Kyoto Encyclopedia of Genes and Genomes (KEGG) database.[20] After gene level QC, we retained 240,544 genes for case-control comparison. We performed case-control comparison for each gene and identified eight genes which significantly increased in SLE (empirically estimated false discovery ratio (FDR)=0.05; figure 2A,B, table 2). As conducted in the phylogenetic analysis, we performed subanalysis (online supplemental table 3). For the eight-genes increased in SLE, effect sizes were almost similar among subanalyses (online supplemental figure 5). These results suggested that inclusion of the male subjects or those who took medications such as proton pump inhibitor, antibiotics or therapeutics did not confound the result. The abundance of these eight genes was not significantly different between the newly onset patients and the other patients, the patients with LN and without LN or the patients with high SLE-DAI and low SLE-DAI (online supplemental table 3).

All of these eight genes were registered as *Streptococcus parasanguinis ATCC 15912* or *Streptococcus parasanguinis FW213* derived. In our metagenome data, although major derivation of these genes were *Streptococcus parasanguinis*, reference genomes of other *Streptococcus* such as *Streptococcus sanguinis* or unclassified *Streptococcus* were linked to these genes. *Streptococcus parasanguinis* was not significantly increased in the gut metagenome of the patients with SLE in our phylogenetic analysis (effect size=0.122, $P_{microbe}$=0.35), indicating the possibility of collective enrichment of the multiple species of *Streptococcus* which had the several genes in common or difference in the composition of genes among *Streptococcus parasanguinis* strains. Among the eight genes which significantly increased in the patients with SLE, Spaf_0732 was a glutaredoxin-like protein. Some glutaredoxin-like protein was involved in reactive oxygen metabolism.[21] As previously described, gut redox environment has substantial effect on the host's immune system,[22] and its alteration in the gut microbiome of autoimmune diseases such as RA was reported.[5 6]

Identification of metagenomic biological pathways altered in the patients with SLE

Using the result of the gene level MWAS, we performed a gene set enrichment analysis to evaluate the case-control discrepancy of the gut metagenome at pathway level. We evaluated 126 QC-passed pathways registered in KEGG database. We found that genes differentially abundant between case and control were significantly enriched on seven pathways (FDR<0.05; figure 3A,B, online supplemental table 4). One of the significant pathways was sulfur metabolism and sulfur was associated with redox reaction,[6] suggesting that altered redox reaction was associated with the pathology of SLE. Enrichment of flagellar assembly might result from bacteria-host interaction mediated by strong immune reaction to bacterial flagellar.[23]

SLE-specific biological pathways shared between metagenome and human genome

We integrated the result of the current SLE MWAS data and the previously published SLE GWAS data (4,943 cases and 8,483 controls)[24] for assessing the sharing of biological pathways between the gut microbiome and the host. We used PASCAL[25] for pathway analysis of the GWAS summary statistics. A total of 94 pathways registered in KEGG database were commonly evaluated for MWAS and GWAS. We compared the p values of the each KEGG pathway ($P_{pathway}$) between the SLE MWAS and the SLE GWAS. We found a significant overlap between the pathways enriched both in the SLE MWAS ($P_{pathway}$ for metagenome <0.05) and in the SLE GWAS ($P_{pathway}$ for SLE GWAS<0.05; P_{Fisher}=0.041; figure 3C). To check whether the overlap of the enriched pathways between the metagenome and the host genome truly reflected the SLE-specific changes in the biological pathways, we performed the same experiment

Table 2 Genes with significant case-control discrepancy in the SLE MWAS.

KEGG gene	Effect size	SE	P_{KEGG}	Gene name, definition	Organism
HMPREF0833_10768	0.850	0.141	6.7×10^{-9}	udk; uridine kinase	*Streptococcus parasanguinis ATCC 15912*
HMPREF0833_10371	0.821	0.148	8.1×10^{-8}	Hypothetical protein	*Streptococcus parasanguinis ATCC 15912*
Spaf_0813	0.781	0.149	3.6×10^{-7}	Hypothetical protein	*Streptococcus parasanguinis FW213*
HMPREF0833_10659	0.716	0.138	4.7×10^{-7}	Methyltransferase small domain protein	*Streptococcus parasanguinis ATCC 15912*
HMPREF0833_10122	0.701	0.135	5.0×10^{-7}	Membrane protein	*Streptococcus parasanguinis ATCC 15912*
HMPREF0833_10143	0.762	0.149	7.0×10^{-7}	Hydrolase	*Streptococcus parasanguinis ATCC 15912*
Spaf_0732	0.760	0.149	7.1×10^{-7}	nrdH; Glutaredoxin-like protein	*Streptococcus parasanguinis FW213*
HMPREF0833_10389	0.770	0.152	9.1×10^{-7}	Hypothetical protein	*Streptococcus parasanguinis ATCC 15912*

KEGG, Kyoto Encyclopedia of Genes and Genomes.;

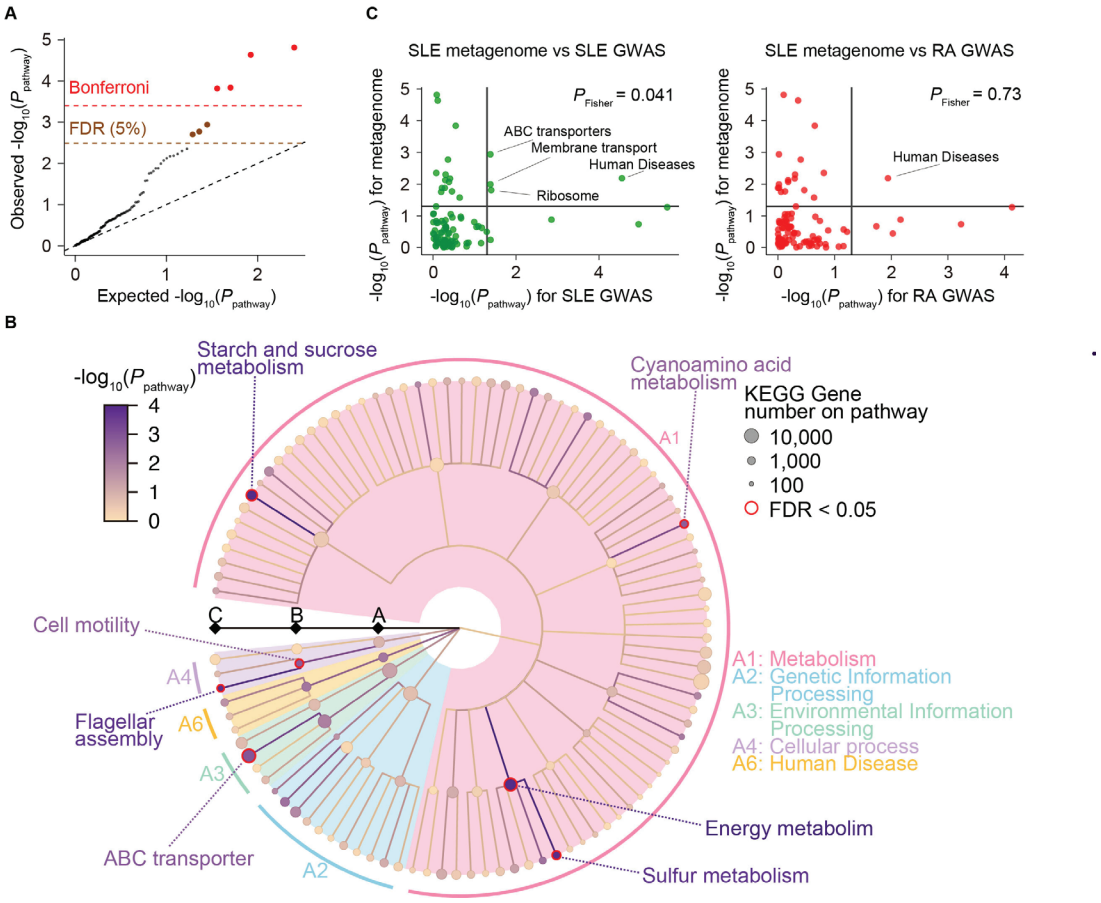


Figure 3 MWAS results of the SLE case-control pathway association tests. (A) A quantile-quantile plot of the MWAS p values of pathways based on KEGG pathways ($P_{pathway}$). The x-axis indicates log-transformed empirically estimated median $P_{pathway}$. The y-axis indicates observed $-\log_{10}(P_{pathway})$. The diagonal dashed line represents y=x, which corresponds to the null hypothesis. The horizontal red dashed line indicates the Bonferroni-corrected threshold (α=0.05), and the brown dashed line indicates the FDR threshold (FDR=0.05) calculated with Benjamini-Hochberg method. Pathways with p values less than the Bonferroni thresholds are plotted as red dots. Pathways with FDR<0.05 are plotted as brown dots, and other pathways are plotted as black dots. (B) 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 the colour of dots represent set sizes and $P_{pathway}$, respectively. The seven pathways with significant enrichment (FDR<0.05) are outlined in red. (C) Comparison of $P_{pathway}$ between the SLE MWAS and GWAS data. The x-axis indicates the $P_{pathway}$ of the GWAS data (left, SLE GWAS; right, RA GWAS). The y-axis indicates the $P_{pathway}$ of the SLE MWAS. The horizontal and vertical black lines indicate $P_{pathway}$ of 0.05. The overlap of the pathway enrichment was evaluated by classifying the pathways based on the significance threshold of $P_{pathway}$<0.05 or $P_{pathway}$≥0.05 and using Fisher's exact test. FDR, false discovery rate; GWAS, genome-wide association study; KEGG, Kyoto Encyclopedia of Genes and Genomes; MWAS, metagenome-wide association study; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus.

with the SLE MWAS data and RA GWAS data (14,361 cases and 43,923 controls).[26] When using the RA GWAS data, the overlap of the enriched pathways between the metagenome and the host genome was not significant (P_{Fisher}=0.73; figure 3C). Therefore, our results suggested that there was a commonality in the enriched biological pathways between human genome and metagenome in SLE, namely MWAS-GWAS interaction.

Dysbiosis in the gut microbiome of the patients with SLE

Dysbiosis refers to an unbalanced microbiota, which is usually harmful for us. Decrease in the α-diversity (ie, within individual diversity) of microbiome was one of the most constant findings of the gut dysbiosis[27] and reported in many disease conditions including IBD.[28] As for SLE, decrease in α-diversity of the gut microbiome was still controversial.[12–17] Therefore, we performed case-control comparison of α-diversity in the phylogenetic data (L2–L7) and the gene abundance data based on KEGG database. Significant decreases of α-diversity in the low taxonomic level phylogenetic data (L5–L7; $P_{\alpha\text{-diversity}}$<0.05/6=0.0083) and the gene abundance data were observed ($P_{\alpha\text{-diversity}}=7.9\times10^{-5}$; figure 4A,B, online supplemental tables 5 and 6). In subanalysis, significant decreases of α-diversity in the phylogenetic data at L5 and L6 levels and the gene abundance data were still observed ($P_{\alpha\text{-diversity}}$<0.05/6=0.0083 for phylogenetic data and $P_{\alpha\text{-diversity}}$<0.05 for gene abundance data; online supplemental tables 5 and 6). Although decrease of α-diversity in the phylogenetic data at L7 level was not significant when removing patients with antibiotics usage ($P_{\alpha\text{-diversity}}$=0.052), direction of the effect size was consistent. Microbial α-diversity was not significantly different between the newly onset patients and the other patients, the patients with and without LN or the patients with high SLE-DAI and low SLE-DAI ($P_{\alpha\text{-diversity}}$ > 0.05/6 in phylogenetic analysis and $P_{\alpha\text{-diversity}}$ >0.05 in functional analysis; online supplemental tables 5 and 6).

Next, we performed a β-diversity analysis for checking whether SLE affected the overall microbial composition or

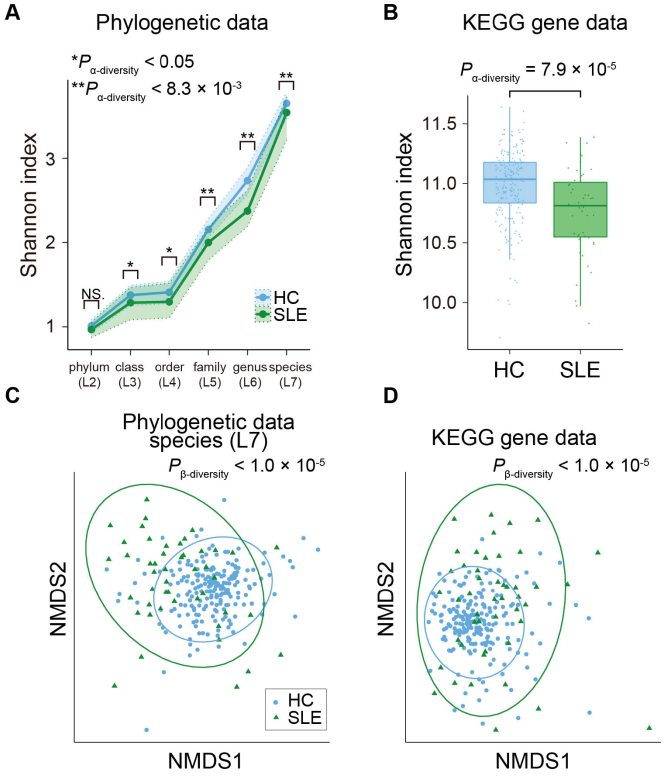


Figure 4 Case-control comparison of the microbial diversities in SLE. (A) α-diversities of the phylogenetic relative abundance data for the six taxonomic levels. Blue and green dots represent the median Shannon index of the HC and SLE subjects. Upper and lower dashed lines indicate the first and third quantile of Shannon index for the HC and SLE subjects. (B) α-diversities of the gene abundance based on KEGG gene databases. Boxplots indicate the median values (centre lines) and the IQRs (box edges), with whiskers extending to the most extreme points within the range between (lower quantile − [1.5×IQR]) and (upper quantile + [1.5×IQR]). (C) β-diversities of the phylogenetic relative abundance data at the species level. Result of NMDS based on Bray-Curtis distance is represented. Blue and green dots represent the HC and SLE subjects. (D) β-diversities of the gene abundance based on KEGG gene database. Result of NMDS based on Bray-Curtis distance is represented. Blue and green dots represent the HC and SLE subjects. *$P_{\alpha\text{-diversity}}$<0.05; **$P_{\alpha\text{-diversity}}$<0.0083. HC, healthy control; KEGG, Kyoto Encyclopedia of Genes and Genomes; NMDS, non-metric multidimensional scaling; SLE, systemic lupus erythematosus.

not. We performed PERMANOVA,[29] based on Bray-Curtis distance calculated from the phylogenetic data (L2–L7) and the gene abundance data. Significant differences were detected in the phylogenetic data (L2–L7) and the gene abundance data with consistency in subanalysis ($P_{\beta\text{-diversity}}$<0.05/6=0.0083 for phylogenetic data and $P_{\beta\text{-diversity}}$<0.05 for gene abundance data; figure 4C,D, online supplemental figure 6, tables 7 and 8). There was no significant difference in the overall microbial composition between the newly onset patients and the other patients, the patients with and without LN or the patients with high SLE-DAI

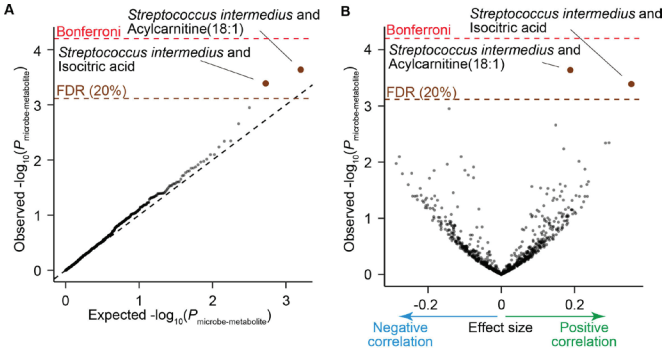


Figure 5 Result of the microbe-metabolite association analysis. (A) A quantile-quantile plot of the p values from the microbe-metabolite association analysis ($P_{microbe\text{-}metabolite}$). The x-axis indicates log-transformed empirically estimated median $P_{microbe\text{-}metabolite}$. The y-axis indicates observed -$\log_{10}$($P_{microbe\text{-}metabolite}$). The diagonal dashed line represents y=x, which corresponds to the null hypothesis. The horizontal red dashed line indicates the Bonferroni-corrected threshold (α=0.05), and the brown dashed line indicates the FDR threshold (FDR=0.20) calculated with Benjamini-Hochberg method. The microbe-metabolite pairs with FDR<0.20 are plotted as brown dots, and the other microbe-metabolite pairs are plotted as black dots. (B) A volcano plot. The x-axis indicates effect sizes in linear regression. The y-axis, horizontal dashed lines and dot colours are the same as in (A). FDR, false discovery rate; HC, healthy control; SLE, systemic lupus erythematosus.

and low SLE-DAI ($P_{\beta\text{-diversity}}$>0.05/6 in phylogenetic analysis and $P_{\beta\text{-diversity}}$>0.05 in functional analysis; online supplemental tables 7 and 8). Collectively, diversity analysis provided evidence of the dysbiosis observed in the gut microbiome of the patients with SLE.

Association between plasma metabolite and the gut microbiome of the patients with SLE

Gut microbiome can affect our body by changing the profiles of circulating metabolites.[30–32] To assess the association between the SLE-associated taxa and plasma metabolites, we integrated the phylogenetic data and plasma metabolite profiles which were previously obtained from a part of the participants of this study.[33] As the focus of this analysis is not the case-control discrepancy but the microbe-metabolite association, we combined 94 HC subjects and 9 patients with SLE, which resulted in 103 participants. Abundance of the two metabolites, acylcarnitine(18:1) and isocitric acid were significantly positively correlated with the abundance of *Streptococcus intermedius* ($P_{microbe\text{-}metabolite}$ <4.1×10^{-4}; FDR<0.20; figure 5A,B, table 3). We performed a replication analysis using another dataset composed of 75 HC subjects.[8] Positive correlation between acylcarnitine(18:1) and *Streptococcus intermedius* was successfully replicated ($P_{microbe\text{-}metabolite}$=0.0080). Acylcarnitine is formed by carnitine and acyl-coenzyme A (CoAs) derived from fatty acids. Acylcarnitine was reported to be one of the main components of faecal bacteria-metabolite network and associated with the numerous dysbiosis associated species.[7] Acylcarnitine can work as an inflammatory

Table 3 Microbe-metabolite pairs with significant association

		Discovery (*N*=103)				Replication (*N*=75)		
Microbe	Metabolite	Effect size	SE	$P_{microbe\text{-}metabolite}$	*q*	Effect size	SE	$P_{microbe\text{-}metabolite}$
Streptococcus intermedius	Acylcarnitine(18:1)	0.188	0.049	2.3×10^{-4}	0.16	0.166	0.061	0.0080
Streptococcus intermedius	Isocitric acid	0.355	0.097	4.1×10^{-4}	0.16	0.155	0.119	0.19

signal,[34] suggesting that the gut microbiome of the patients with SLE is associated with the overactivation of the immune system in SLE via acylcarnitine.

DISCUSSION

In this study, we conducted a MWAS of Japanese patients with SLE using whole-genome shotgun sequencing. Our study revealed following biological features associated with the SLE gut metagenome (online supplemental figure 7): (1) *Streptococcus anginosus* and *Streptococcus intermedius* were increased in the SLE metagenome; (2) eight genes derived from *Streptococcus* including a gene related to redox reaction increased in the SLE metagenome; (3) various biological pathways, including those related to sulfur metabolism and flagella assembly were enriched among genes differentially abundant between case and control; (4) there existed an SLE-specific link between biological pathway of the gut microbiome and the host genome, namely MWAS-GWAS interaction; (5) the features of dysbiosis, decreases in α-diversity and changes in the overall composition of the gut microbiome, were observed among the patients with SLE; (6) plasma acylcarnitine(18:1) level was positively associated with the abundance of *Streptococcus intermedius*.

One of the principal findings of our study was increase of *Streptococcus anginosus* and *Streptococcus intermedius* in the SLE metagenome. Because it was difficult to detect these species level clades using classical 16S rRNA sequence analysis, these results demonstrate the value of metagenome shotgun sequencing in identifying disease-associated taxa. Considering the clinical features of SLE such as female-biased prevalence and frequent prophylactic antibiotics usage before renal biopsy or during immunosuppressive treatment, we evaluate the effect of these factors on the result of MWAS. Although the findings of the MWAS were often not replicated across studies due to the difference in ethnicity and lifestyle, our results were validated in the independent SLE metagenome dataset from the Chinese cohort,[17] suggesting that our statistical analysis robustly identified the taxa specifically abundant in the SLE metagenome. *Streptococcus anginosus* and *Streptococcus intermedius* belong to *Streptococcus anginosus* group and are parts of normal flora of the oral cavity and gastrointestinal tracts. Involvement of oral-gut interaction mediated by microbes was reported in several diseases,[5 35] suggesting the possibility of association between the oral-gut axis and the SLE pathology. Liu *et al* reported that *Streptococcus intermedius* produced and secreted a histone-like DNA binding protein, which induced proinflammatory cytokine production of macrophage derived cell line.[36] Therefore, interaction between *Streptococcus intermedius* and the host immune system could be related to the pathology of SLE.

Although the *Streptococcus anginosus* and *Streptococcus intermedius* consistently increased in the existing two shotgun sequencing studies (Chen *et al*[17] and our study), some previous findings in Chen *et al* were not replicated in this study. Although Chen *et al* reported that *Ruminococcus gnavus* increased in the patients with SLE, especially those with LN, it did not increase in our study (effect size=−0.001, SE=0.155, P_{microbe}=1.00 in case-control comparison and effect size=0.195, SE=0.355, P_{microbe}=0.59 in comparison between patients with SLE with LN or without LN). Additionally, among the 74 species which increased in Chen *et al* and evaluated in our study, two species, *Streptococcus anginosus* and *Streptococcus intermedius*, were replicated with $P_{\text{microbe}} < 0.05/74 = 6.8 \times 10^{-4}$. Among the 74 species, 38 species have the same directional effects between the studies. This heterogeneity might reflect the effects of geography and lifestyle on the gut microbiome, while differences in the analytic methods could not be rejected. Thus, further studies in the different countries or global meta-analysis will be warranted to further clarify the difference in the gut microbiome of the patients with SLE among different countries.

Our gene analysis revealed that eight genes including a glutaredoxin-like protein gene were increased in the SLE gut metagenome. Since all of these genes were derived from *Streptococcus,* there might be a possibility that some of these genes were enriched simply because of coabundance with *Streptococcus* genes truly relevant to the SLE condition. Further functional validation of each gene would be warranted. Subsequent pathway analysis based on the result of the gene analysis revealed an alteration of various biological pathways including sulfur metabolism and flagella assembly. Sulfur metabolism is reported to be altered in the metagenomes of other diseases such as RA,[6] and it is related to redox environment. Together with the result of the gene level analysis, our results suggested that alteration of redox environment was associated with the pathology of SLE. Flagellar is known to elicit strong immune response. Zeevi *et al* reported that gut bacteria which had structural variant in flagellar protein had higher growth ratio, implying the loss-of-function adaptation to the host's immune system.[23] There was an interaction between the gut bacteria and the immune system via flagellar, and the alteration of flagellar-related pathways in the SLE gut metagenome could be associated with changes in the host's immune system. Through the MWAS-GWAS integrative analysis, we showed that there was a biological pathway level commonality between the host genome and the metagenome in among the patients with SLE. Although biological pathway level commonality between host genome and metagenome was evaluated in other autoimmune diseases,[5 8] disease specificity of the commonality had not been evaluated. In this study, we showed that there was no pathway level commonality between the result of the SLE MWAS and the RA GWAS. Pathway level microbe-host interaction detected from the SLE MWAS and the SLE GWAS should reflect an SLE-specific disease mechanism.

Decrease in α-diversity of the gut microbiome, which is one of the major characteristics of dysbiosis, in patients with SLE had been controversial; some reported significant decreases in α-diversity in the SLE gut microbiome,[13 14 17] and others showed no differences.[12 15 16] This might be due to the difference in study design, such as sample number, country, medication and treatment of the patients with SLE. In our analysis, α-diversity of the gut microbiome in the patients with SLE significantly decreased. Additionally, we certified the robustness of our result by subanalysis. Diversity of the gut microbiome is considered as important for the homeostasis of the host's immune system, and decrease in α-diversity is reported in autoimmune diseases such as type 1 diabetes and IBD.[27] Observed decrease of α-diversity in the gut microbiome of SLE could be associated with abnormal activation of the immune system in the patients with SLE. Through β-diversity analysis, we found that SLE condition significantly affected the overall microbial composition. As recently suggested by Ma *et al*,[37] heterogeneity of the human microbiome among individuals tended to increase in the disease condition (Anna Karenina principle; AKP), and there is a possibility that AKP is also applicable to the case of SLE.

In this study, we identified a positive correlation between *Streptococcus intermedius* and acylcarnitine(18:1) followed by replication in another dataset. Acylcarnitine is known to form complex network with various microbes,[7] association between *Streptococcus intermedius* and acylcarnitine(18:1) could be mediated by both direct interaction and indirect interaction mediated by other microbes. Rutkowsky *et al* reported that acylcarnitine induced inflammation

in macrophage derived cell line by cyclooxygenase-2 dependent manner.[34] Our microbe-metabolome association analysis revealed a functional readout from the SLE gut microbiome, which could be associated with the pathology of SLE. Further analysis including case-control comparison with greater number of cases and in vivo validation would be warranted.

Our SLE MWAS had a few limitations. First, our study had only a moderate sample size compared with other studies on more common diseases such as colorectal cancer[38 39] and type 2 diabetes[3] due to the relatively rare prevalence of SLE. Although we robustly detected SLE-associated taxa and genes, there may exist other taxa and genes with smaller effect size. Thus, future large-scale studies such as cross-cohort meta-analysis are needed to detect the evidence of such weaker associations (online supplemental figure 8), where our study will contribute. Second, some of the patients in out cohort were under treatment or antibiotics, which could be potential confounding factors. However, the stable consistency of the effect sizes among the subanalyses indicated that these factors might not confound the result of the MWAS. Replication by independent cohort further supported the robustness of the result. Third, it is still challenging to reveal mechanistic insights into disease biology from MWAS. Pathway and bacteria-metabolite analysis in our study provided potential causal mechanisms as well as those suggested previously.[9–11 40] However, biological overview is still elusive due to the low throughput of mice experiment and technical and ethical difficulty in intervention to the human subjects. Future studies involving the latest technologies such as organoids and organs-on-chips technology,[41] would be promising for studying the mechanistic insights into the relationship between the gut microbiome and SLE.

In conclusion, our shotgun sequencing-based MWAS and integrative analysis with GWAS and plasma metabolite profiles revealed the altered gut microbiome in SLE and its association with the host. Our analysis contributes to the understanding of the relationship between the gut microbiome and SLE and provides useful resources for future research such as in vivo functional investigation or large-scale meta-analysis.

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

Acknowledgements We would like to thank all the participants involved in this study. We thank Ms. Miho Kaneda for supporting the study.

Contributors YT, TK and YO designed the study and conducted the data analysis. YT and YO wrote the manuscript. YT, YMae, EO-I, TK, DM, YMat, TN and SN conducted the experiments. YT, YMae, EO-I, TK, KY, KS, HM, MYo, MYa, TN and SO collected the samples. HI, KT, AK and YO supervised the study. All authors contributed to the article and approved the submitted version.

Funding This research was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (19H01021 and 20K21834), the Japan Agency for Medical Research and Development (AMED; JP21km0405211, JP21ek0109413, JP21gm4010006, JP21km0405217 and JP21ek0410075), JST Moonshot R&D Grant Number JPMJMS2021, Takeda Science Foundation, Bioinformatics Initiative of Osaka University Graduate School of Medicine, Clinical Investigator's Research Project of the Osaka University Graduate School of Medicine, Grant Programme for Next Generation Principal Investigators at Immunology Frontier Research Centre (WPI-IFReC), Osaka University.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This study was approved by the ethical committee of Osaka University (ID: 734–12).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data are available on reasonable request. 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.

ORCID iDs
Yuichi Maeda http://orcid.org/0000-0002-6831-8205
Eri Oguro-Igashira http://orcid.org/0000-0001-5863-0775
Yukinori Okada http://orcid.org/0000-0002-0311-8472

REFERENCES

1 Stojan G, Petri M. Epidemiology of systemic lupus erythematosus: an update. Curr Opin Rheumatol 2018;30:144–50.
2 Holmes E, Li JV, Marchesi JR, *et al*. Gut microbiota composition and activity in relation to host metabolic phenotype and disease risk. Cell Metab 2012;16:559–64.
3 Qin J, Li Y, Cai Z, *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490:55–60.
4 Yachida S, Mizutani S, Shiroma H, *et al*. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 2019;25:968–76.
5 Kishikawa T, Maeda Y, Nii T, *et al*. Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population. Ann Rheum *Dis* 2020;79:103–11.
6 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.
7 Lloyd-Price J, Arze C, Ananthakrishnan AN, *et al*. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 2019;569:655–62.
8 Kishikawa T, Ogawa K, Motooka D, *et al*. A Metagenome-Wide association study of gut microbiome in patients with multiple sclerosis revealed novel disease pathology. Front Cell Infect Microbiol 2020;10:585973.
9 Ma Y, Xu X, Li M, *et al*. Gut microbiota promote the inflammatory response in the pathogenesis of systemic lupus erythematosus. Mol Med 2019;25:35.
10 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.
11 Choi S-C, Brown J, Gong M, *et al*. Gut microbiota dysbiosis and altered tryptophan catabolism contribute to autoimmunity in lupus-susceptible mice. Sci Transl Med 2020;12:eaax2220.
12 Hevia A, Milani C, López P, *et al*. Intestinal dysbiosis associated with systemic lupus erythematosus. mBio 2014;5:e01548:14.

13 Azzouz D, Omarbekova A, Heguy A, *et al*. Lupus nephritis is linked to disease-activity associated expansions and immunity to a gut commensal. Ann Rheum Dis 2019;78:947–56.
14 Li Y, Wang H-F, Li X, *et al*. Disordered intestinal microbes are associated with the activity of systemic lupus erythematosus. Clin Sci 2019;133:821–38.
15 Wei F, Xu H, Yan C, *et al*. Changes of intestinal flora in patients with systemic lupus erythematosus in northeast China. PLoS One 2019;14:e0213063.
16 van der Meulen TA, Harmsen HJM, Vila AV, *et al*. Shared gut, but distinct oral microbiota composition in primary Sjögren's syndrome and systemic lupus erythematosus. J Autoimmun 2019;97:77–87.
17 Chen B-di, Jia X-M, Xu J-Y, *et al*. An Autoimmunogenic and proinflammatory profile defined by the gut microbiota of patients with untreated systemic lupus erythematosus. Arthritis Rheumatol 2021;73:232–43.
18 Deschasaux M, Bouter KE, Prodan A, *et al*. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nat Med 2018;24:1526–31.
19 He Y, Wu W, Zheng H-M, *et al*. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 2018;24:1532–5.
20 Kanehisa M. Kyoto encyclopedia of genes and genomes. , 2000: 28, 27–30.
21 Si M-R, Zhang L, Yang Z-F, *et al*. Nrdh Redoxin enhances resistance to multiple oxidative stresses by acting as a peroxidase cofactor in Corynebacterium glutamicum. Appl Environ Microbiol 2014;80:1750–62.
22 Campbell EL, Colgan SP. Control and dysregulation of redox signalling in the gastrointestinal tract. Nat Rev Gastroenterol Hepatol 2019;16:106–20.
23 Zeevi D, Korem T, Godneva A, *et al*. Structural variation in the gut microbiome associates with host health. Nature 2019;568:43–8.
24 JuliA, López-Longo FJ, Pérez Venegas JJ. Genome-wide association study meta-analysis identifies five new loci for systemic lupus erythematosus 2018;20:100.
25 Lamparter D, Marbach D, Rueedi R, *et al*. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput Biol 2016;12:e1004714.
26 Okada Y, Wu D, Trynka G, *et al*. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 2014;506:376–81.
27 Mosca A, Leclerc M, Hugot JP. Gut microbiota diversity and human diseases: should we Reintroduce key predators in our ecosystem? Front Microbiol 2016;7:455.
28 Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology 2014;146:1489–99.
29 Anderson MJ. A new method for non-parametric multivariate analysis of variance: non-parametric MANOVA for ecology. Austral Ecology;26:32–46.
30 Visconti A, Le Roy CI, Rosa F, *et al*. Interplay between the human gut microbiome and host metabolism. Nat Commun 2019;10:4505.
31 Vojinovic D, Radjabzadeh D, Kurilshikov A, *et al*. Relationship between gut microbiota and circulating metabolites in population-based cohorts. Nat Commun 2019;10:5813.
32 Wilmanski T, Rappaport N, Earls JC, *et al*. Blood metabolome predicts gut microbiome α-diversity in humans. Nat Biotechnol 2019;37:1217–28.
33 Kishikawa T, Maeda Y, Nii T, *et al*. Increased levels of plasma nucleotides in patients with rheumatoid arthritis. Int Immunol 2021;33:119–24.
34 Rutkowsky JM, Knotts TA, Ono-Moore KD, *et al*. Acylcarnitines activate proinflammatory signaling pathways. Am J Physiol Endocrinol Metab 2014;306:E1378–87.
35 Kitamoto S, Nagao-Kitamoto H, Jiao Y, *et al*. The Intermucosal connection between the mouth and gut in commensal Pathobiont-Driven colitis. Cell 2020;182:447–62.
36 Liu D, Yumoto H, Hirota K, *et al*. Histone-like DNA binding protein of Streptococcus intermedius induces the expression of pro-inflammatory cytokines in human monocytes via activation of ERK1/2 and JNK pathways. Cell Microbiol;10:070921225744001.
37 Ma ZS. Testing the Anna Karenina principle in human Microbiome-Associated diseases. iScience 2020;23:101007.
38 Thomas AM, Manghi P, Asnicar F, *et al*. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25:667–78.
39 Wirbel J, Pyl PT, Kartal E, *et al*. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 2019;25:679–89.
40 Ogunrinde E, Zhou Z, Luo Z, *et al*. A link between plasma microbial translocation, microbiome, and autoantibody development in first-degree relatives of systemic lupus erythematosus patients. Arthritis Rheumatol 2019;71:1858–68.
41 Puschhof J, Pleguezuelos-Manzano C, Clevers H. Organoids and organs-on-chips: insights into human gut-microbe interactions. Cell Host Microbe 2021;29:867–78.

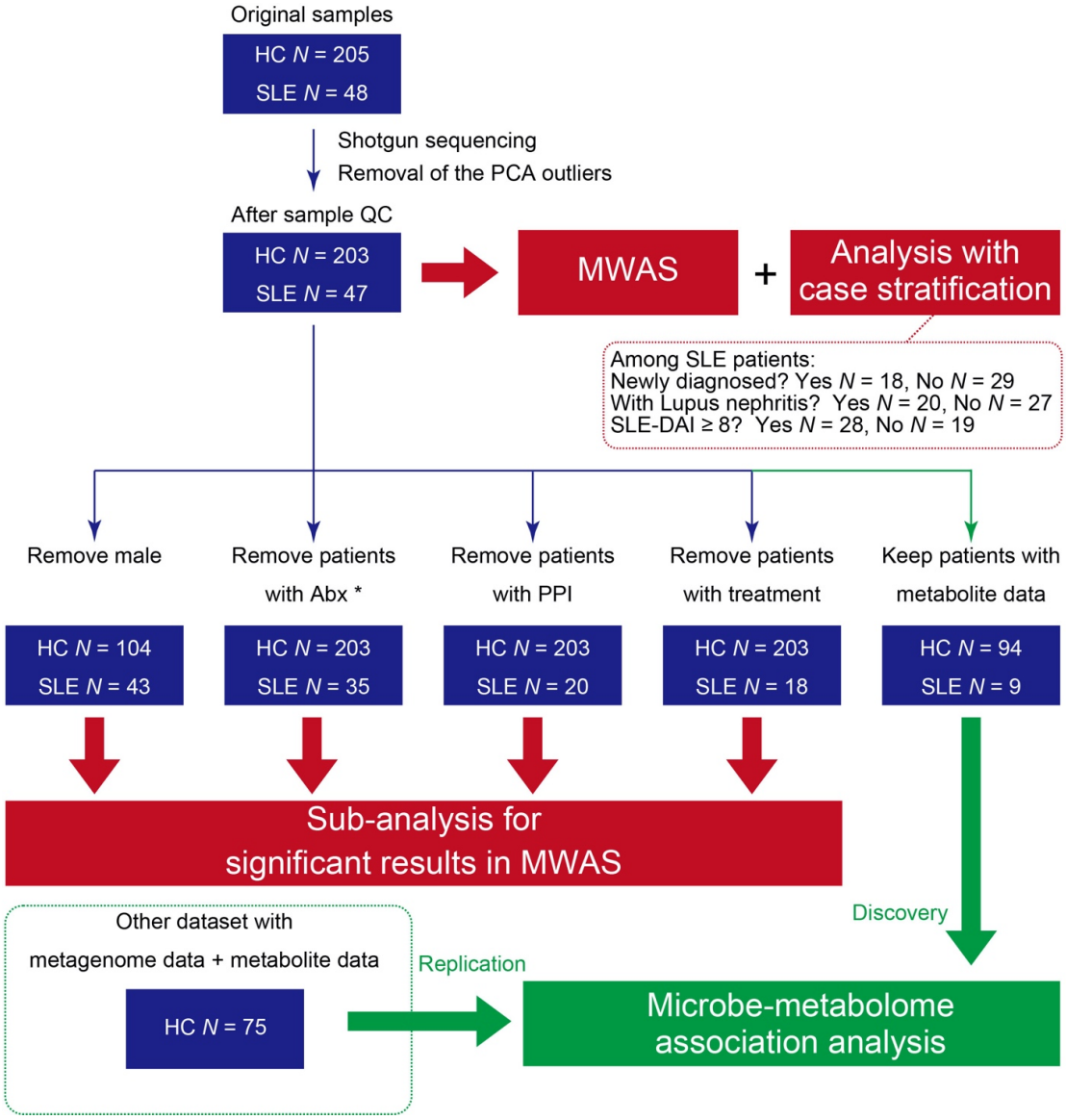
Supplementary information

A metagenome-wide association study revealed disease-specific landscape of the gut microbiome of systemic lupus erythematosus in Japanese

Tomofuji Y et al.

Corresponding to Yukinori Okada (yokada@sg.med.osaka-u.ac.jp)

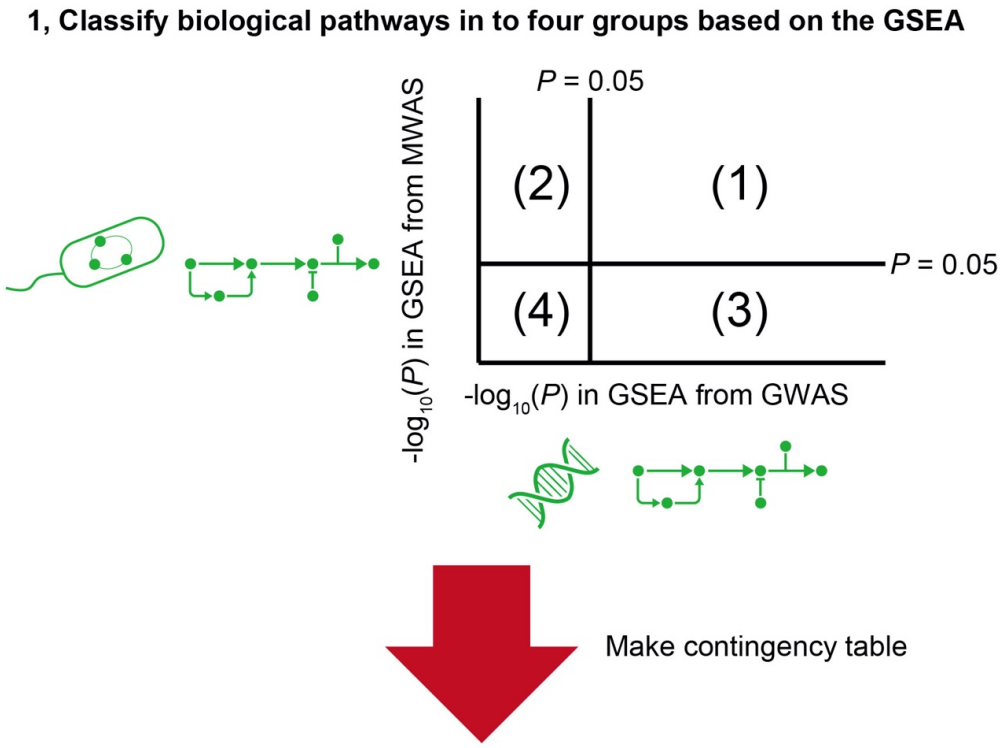
Table of contents



Supplementary Figure 1. A flowchart of the study.

A schematic illustration of the study design. A total of 250 samples (N_{HC} = 203, N_{SLE} = 47) passed the sample QC and were used for the MWAS and analysis with the stratification of the cases. We further extracted subsets of samples for sub-analysis and microbe-metabolite association analysis as described in the figure. Abx, antibiotics; HC, healthy control; MWAS, metagenome-wide association study. PPI, proton pump inhibitor; QC, quality control; SLE, systemic lupus erythematosus. *An SLE patient whose information about antibiotics usage could not be obtained was removed in the sub-analysis.

Tomofuji Y, *et al. Ann Rheum Dis* 2021;0:1–9. doi: 10.1136/annrheumdis-2021-220687



2, Evaluate the enrichment of the biological pathways classified into group (1)

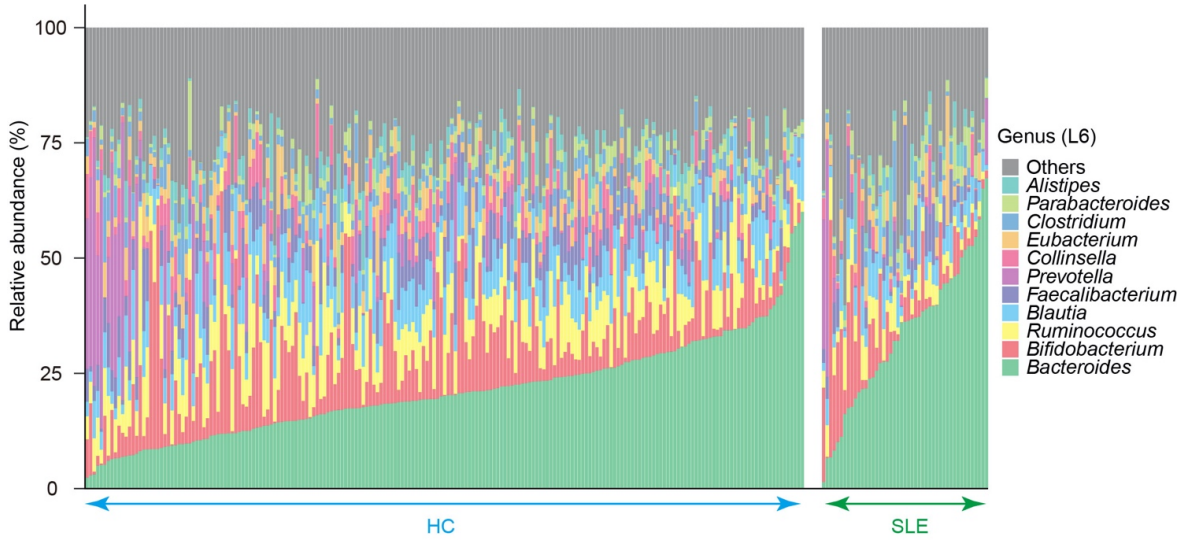
		$P < 0.05$ in GSEA of GWAS	
		Yes	No
$P < 0.05$ in GSEA of MWAS	Yes	(1)	(2)
	No	(3)	(4)

Supplementary Figure 2. A workflow of the MWAS-GWAS interaction analysis.

Firstly, based on the biological pathway enrichment analysis from MWAS and GWAS, we classified biological pathways into the following four groups; (1) Biological pathways enriched in MWAS and GWAS results ($P_{pathway} < 0.05$ in MWAS and $P_{pathway} < 0.05$ in GWAS) (2) Biological pathways enriched only in MWAS result ($P_{pathway} < 0.05$ in MWAS and $P_{pathway} \geq 0.05$ in GWAS). (3) Biological pathways enriched only in GWAS result ($P_{pathway} \geq 0.05$ in MWAS and $P_{pathway} < 0.05$ in GWAS). (4) Biological pathways not enriched in MWAS and GWAS results ($P_{pathway} \geq 0.05$ in MWAS and $P_{pathway} \geq 0.05$ in GWAS). Then,

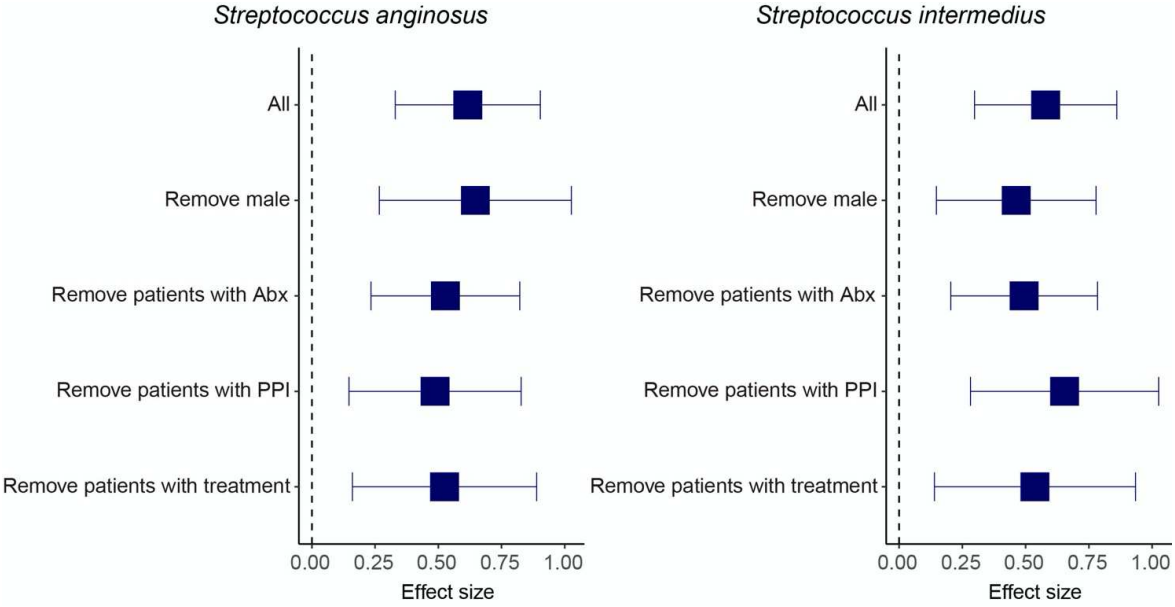
we evaluated the enrichment of the number of biological pathways classified as the group (1) compared to the group (2~4) by one-tailed Fisher's exact test. GSEA, Gene set enrichment analysis; GWAS, Genome-wide association study; MWAS, Metagenome-wide association study.

Tomofuji Y, *et al. Ann Rheum Dis* 2021;0:1–9. doi: 10.1136/annrheumdis-2021-220687



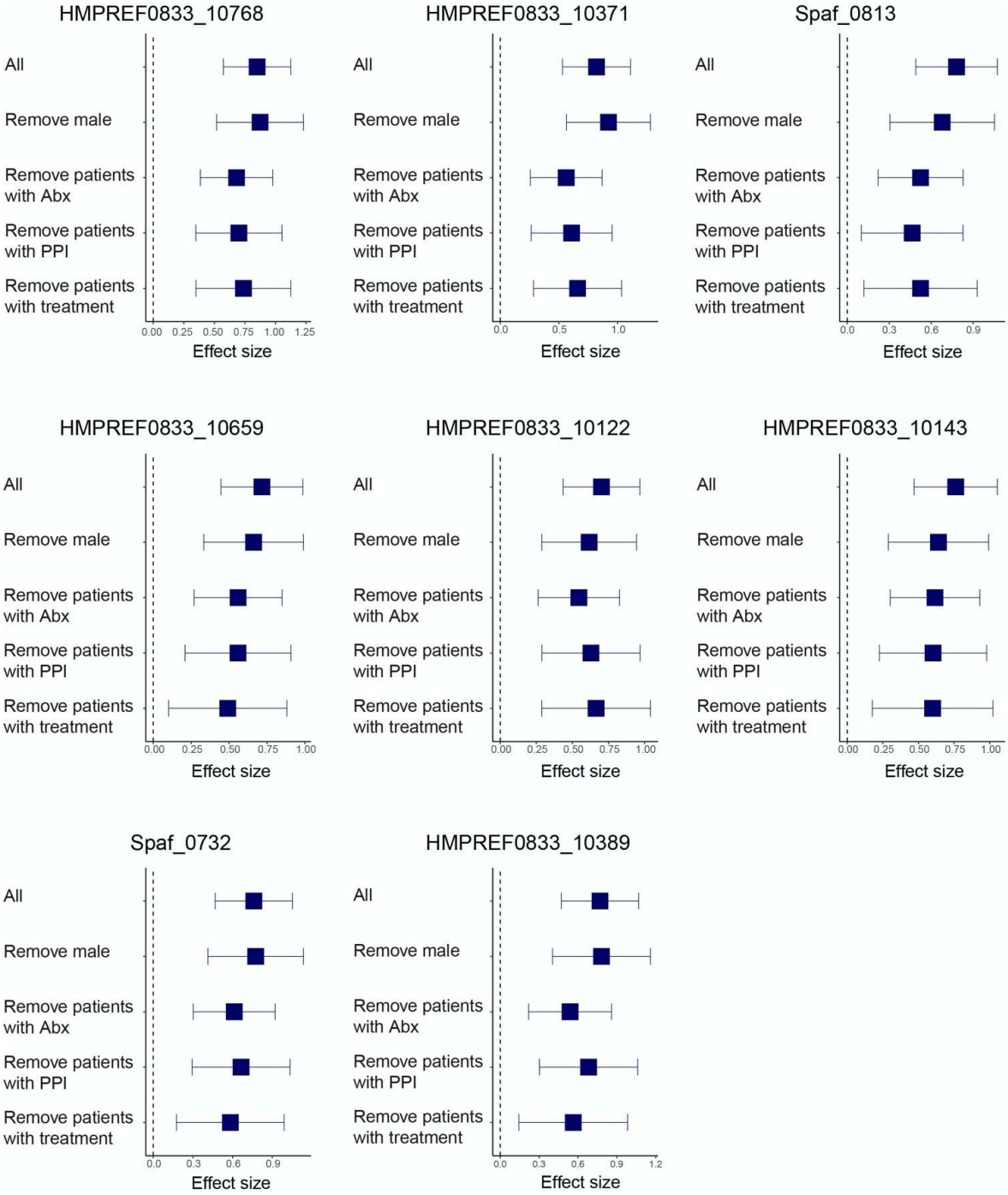
Supplementary Figure 3. Phylogenetic relative abundance at the genus level (L6).

The relative abundance profiles were constructed utilizing whole genome shotgun sequencing (N_{HC} = 203, N_{SLE} = 47). HC, healthy control; SLE, systemic lupus erythematosus.



Supplementary Figure 4. Forest plots from the result of the sub-analysis for the clades with significant SLE case-control discrepancy.

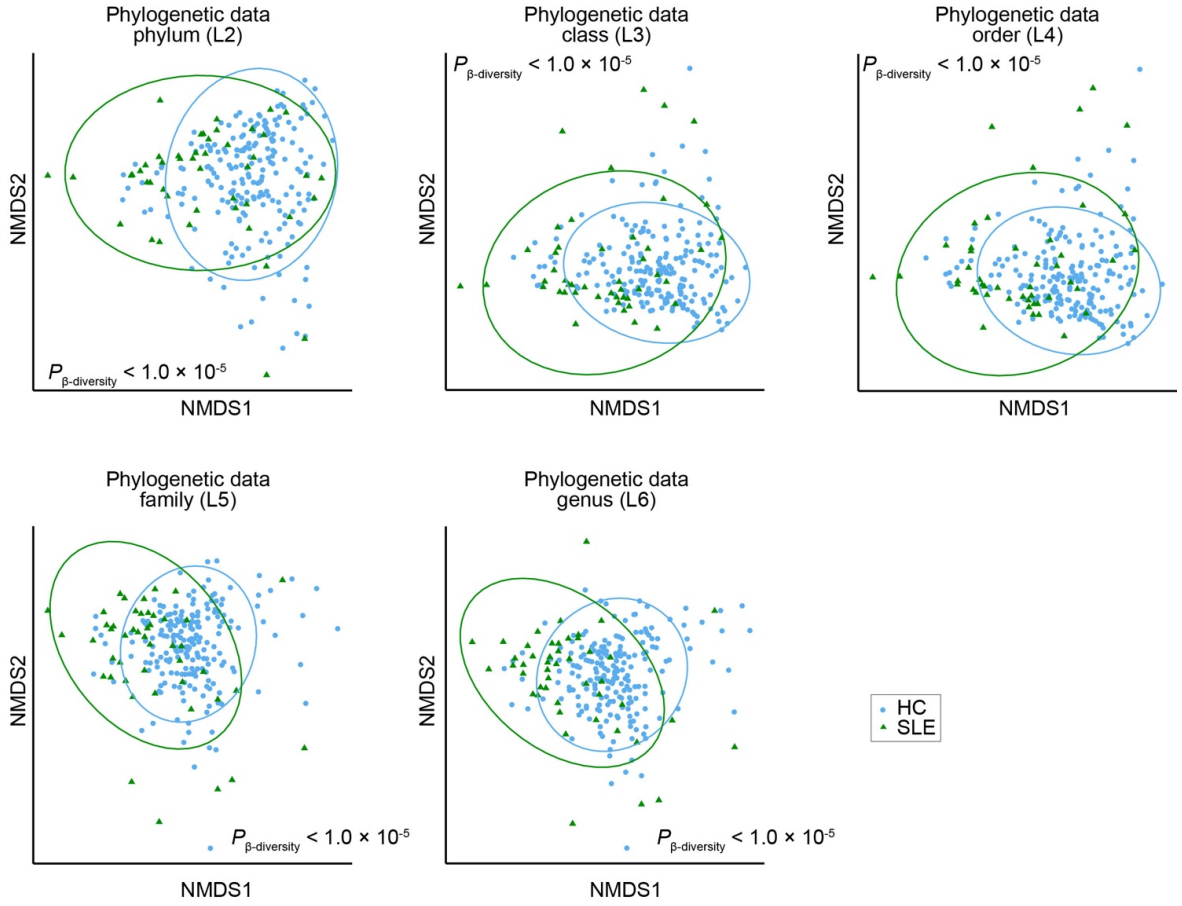
The effect size of the sub-analysis for the clades with significant SLE case-control discrepancy. The boxes indicate the point estimates, and the error bars indicate the 95% confidence interval. Abx, antibiotics; PPI, proton pump inhibitor.



Supplementary Figure 5. Forest plots from the result of the sub-analysis for the genes with significant SLE case-control discrepancy.

The effect size of the sub-analysis for the genes with significant SLE case-control discrepancy. The boxes indicate the point estimates, and the error bars indicate the 95% confidence interval. Abx, antibiotics; PPI, proton pump inhibitor.

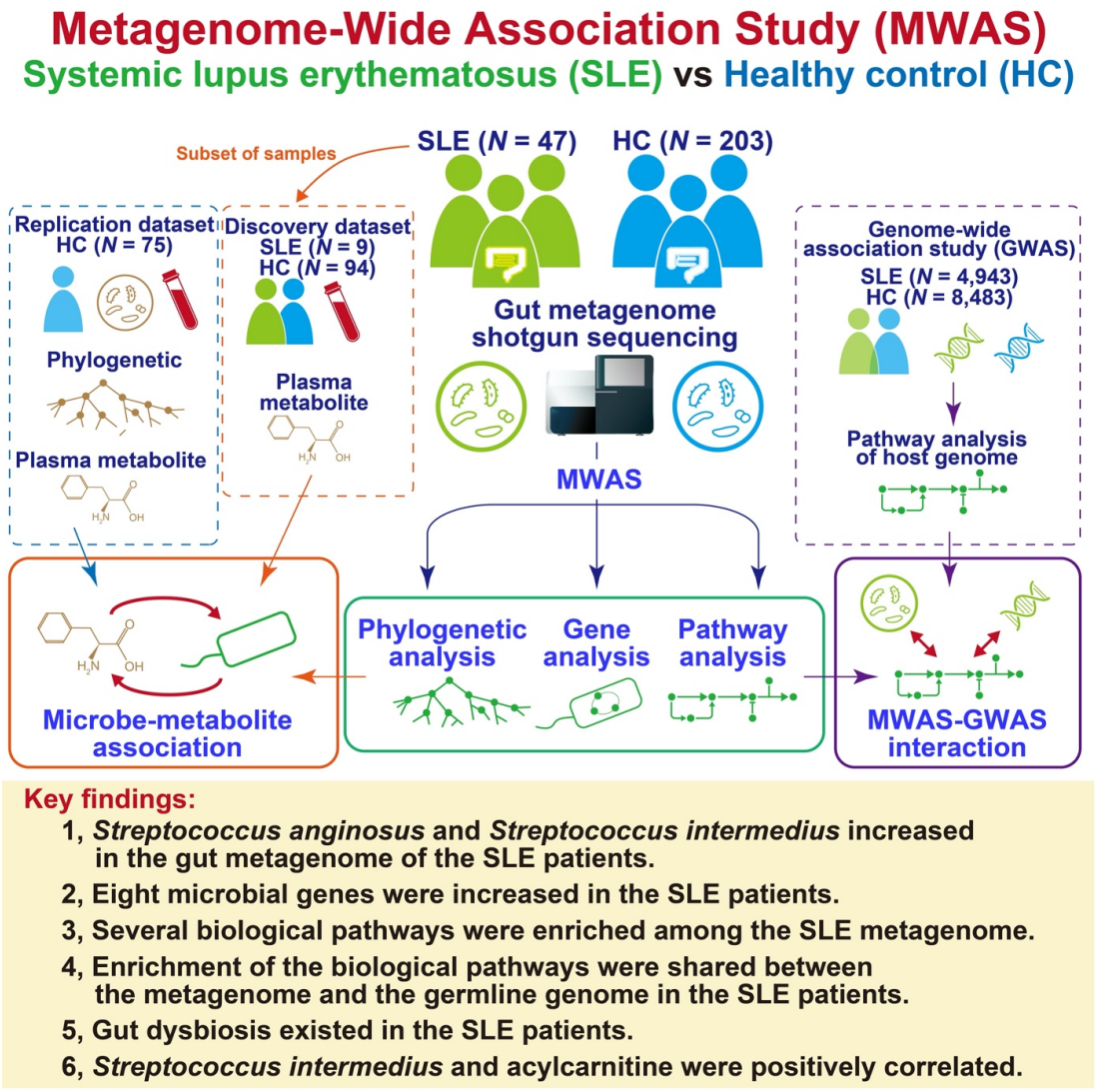
Tomofuji Y, *et al. Ann Rheum Dis* 2021;0:1–9. doi: 10.1136/annrheumdis-2021-220687



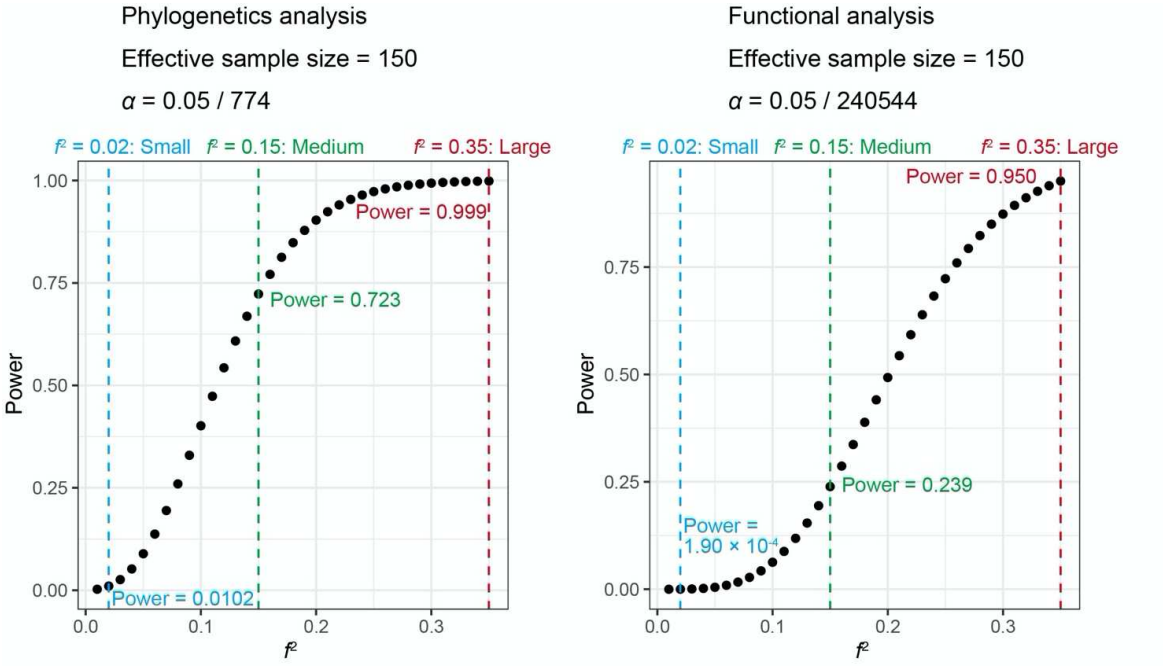
Supplementary Figure 6. β-diversities of the phylogenetic relative abundance data at L2-L6 levels.

The result of NMDS based on Bray-Curtis distance is represented. Blue and green dots represent HC and SLE subjects. HC, healthy control; NMDS, non-metric multidimensional scaling; SLE, systemic lupus erythematosus.

Tomofuji Y, *et al. Ann Rheum Dis* 2021;0:1–9. doi: 10.1136/annrheumdis-2021-220687



Supplementary Figure 7. Graphical summary of this study.



Supplementary Figure 8. Power calculation of the phylogenetic and functional analysis. Power calculation were performed for phylogenetic and functional analysis. $f^2 \geq 0.02$, $f^2 \geq 0.15$, $f^2 \geq 0.35$ represent small, medium, and large effect sizes, respectively. f^2, Cohen's f^2.

Supplementary Table 1. Characteristics of the study population.

	SLE (*N* = 47)	Control (*N* = 203)
Mean age (sd)	42.9 (15.9)	35.4 (12.0)
Sequencing group1	5 (10.6%)	27 (13.3%)
Sequencing group2	10 (21.3%)	71 (35.0%)
Female	43 (91.5%)	104 (51.2%)
Newly diagnosed	18 (38.3%)	-
Not newly diagnosed	29 (61.7%)	-
Treatment		-
Non-treatment	14 (29.8%)	-
Steroid	32 (68.1%)	-
Hydroxychloroquine	7 (14.9%)	-
FK506	5 (10.6%)	-
Ciclosporin	1 (2.1%)	-
Mycophenolate mofetil	8 (17%)	-
Cyclophosphamide	1 (2.1%)	-
Antibiotics use [*1]	11 (23.4%)	-
Proton pump inhibitor use	27 (57.4%)	-
Lupus nephritis	20 (42.6 %)	-
Mean SLE-DAI (sd)	11.1 (8.7)	-

sd, standard deviation; SLE, systemic lupus erythematosus; SLE-DAI, SLE Disease Activity Index.

[*1] Information about antibiotics usage could not be obtained for an SLE patient.

Supplementary Table 2. Result of the sub-analysis for the clades with significant SLE case-control discrepancy.

Streptococcus anginosus

Case-control comparison			
Sample set	Effect size	SE	$P_{microbe}$
All	0.617	0.146	3.7×10^{-5}
Remove male	0.647	0.194	0.0011
Remove patients with antibiotics	0.528	0.150	5.4×10^{-4}
Remove patients with proton pump inhibitor	0.487	0.174	0.0056
Remove patients with treatment for SLE	0.525	0.186	0.0053
comparison within the case			
Variable	Effect size	SE	$P_{microbe}$
Not newly diagnosed	0.373	0.705	0.60
Lupus nephritis	0.046	0.421	0.92
SLE-DAI ≥ 8	0.029	0.037	0.45

Streptococcus intermedius

Case-control comparison			
Sample set	Effect size	SE	$P_{microbe}$
All	0.579	0.143	7.5×10^{-5}
Remove male	0.463	0.161	0.0048
Remove patients with antibiotics	0.494	0.148	0.0010
Remove patients with proton pump inhibitor	0.654	0.190	7.0×10^{-4}
Remove patients with treatment for SLE	0.537	0.203	0.0088
comparison within the case			
Variable	Effect size	SE	$P_{microbe}$
Not newly diagnosed	-0.522	0.578	0.38
Lupus nephritis	0.415	0.335	0.23
SLE-DAI ≥ 8	0.0475	0.029	0.12

SE, standard error; SLE, systemic lupus erythematosus.

Supplementary Table 3. Result of the sub-analysis for the genes with significant SLE case-control discrepancy.

Spaf_0732			
Case-control comparison			
Sample set	Effect size	SE	P_{KEGG}
All	0.760	0.149	7.1×10^{-7}
Remove male	0.774	0.184	5.2×10^{-5}
Remove patients with antibiotics	0.612	0.158	1.4×10^{-4}
Remove patients with proton pump inhibitor	0.664	0.189	5.4×10^{-4}
Remove patients with treatment for SLE	0.583	0.208	0.0056
comparison within the case			
Variable	Effect size	SE	P_{KEGG}
Not newly diagnosed	0.499	0.449	0.28
Lupus nephritis	0.092	0.265	0.73
SLE-DAI ≥ 8	-0.006	0.021	0.77

Spaf_0813			
Case-control comparison			
Sample set	Effect size	SE	P_{KEGG}
All	0.781	0.149	3.6×10^{-7}
Remove male	0.679	0.191	5.5×10^{-4}
Remove patients with antibiotics	0.524	0.155	8.6×10^{-4}
Remove patients with proton pump inhibitor	0.464	0.185	0.013
Remove patients with treatment for SLE	0.524	0.206	0.012
comparison within the case			
Variable	Effect size	SE	P_{KEGG}
Not newly diagnosed	-0.121	0.552	0.83
Lupus nephritis	-0.621	0.275	0.038
SLE-DAI ≥ 8	0.004	0.025	0.87

HMPREF0833_10122

Case-control comparison			
Sample set	Effect size	SE	P_{KEGG}
All	0.701	0.135	5.0×10^{-7}
Remove male	0.616	0.167	3.5×10^{-4}
Remove patients with antibiotics	0.544	0.144	2.0×10^{-4}
Remove patients with proton pump inhibitor	0.628	0.173	3.7×10^{-4}
Remove patients with treatment for SLE	0.664	0.192	6.7×10^{-4}
comparison within the case			
Variable	Effect size	SE	P_{KEGG}
Not newly diagnosed	-0.027	0.429	0.95
Lupus nephritis	-0.074	0.244	0.76
SLE-DAI ≥ 8	-0.004	0.020	0.84

HMPREF0833_10143

Case-control comparison			
Sample set	Effect size	SE	P_{KEGG}
All	0.762	0.149	7.0×10^{-7}
Remove male	0.641	0.179	5.1×10^{-4}
Remove patients with antibiotics	0.616	0.160	1.6×10^{-4}
Remove patients with proton pump inhibitor	0.603	0.192	0.0020
Remove patients with treatment for SLE	0.600	0.216	0.0060
comparison within the case			
Variable	Effect size	SE	P_{KEGG}
Not newly diagnosed	-0.147	0.417	0.73
Lupus nephritis	-0.169	0.235	0.48
SLE-DAI ≥ 8	0.005	0.019	0.78

HMPREF0833_10371

Case-control comparison			
Sample set	Effect size	SE	P_{KEGG}
All	0.821	0.148	8.1×10^{-8}
Remove male	0.923	0.182	1.6×10^{-6}
Remove patients with antibiotics	0.563	0.156	3.9×10^{-4}
Remove patients with proton pump inhibitor	0.608	0.176	6.7×10^{-4}
Remove patients with treatment for SLE	0.659	0.192	7.1×10^{-4}
comparison within the case			
Variable	Effect size	SE	P_{KEGG}
Not newly diagnosed	0.113	0.556	0.84
Lupus nephritis	0.135	0.316	0.67
SLE-DAI ≥ 8	0.004	0.026	0.89

HMPREF0833_10389

Case-control comparison			
Sample set	Effect size	SE	P_{KEGG}
All	0.770	0.152	9.1×10^{-7}
Remove male	0.782	0.193	9.1×10^{-5}
Remove patients with antibiotics	0.539	0.163	0.0011
Remove patients with proton pump inhibitor	0.682	0.194	5.3×10^{-4}
Remove patients with treatment for SLE	0.564	0.214	0.0091
comparison within the case			
Variable	Effect size	SE	P_{KEGG}
Not newly diagnosed	-0.091	0.573	0.88
Lupus nephritis	-0.297	0.319	0.37
SLE-DAI ≥ 8	0.000	0.026	1.00

HMPREF0833_10659

Case-control comparison			
Sample set	Effect size	SE	P_{KEGG}
All	0.716	0.138	4.7×10^{-7}
Remove male	0.662	0.168	1.4×10^{-4}
Remove patients with antibiotics	0.560	0.148	2.1×10^{-4}
Remove patients with proton pump inhibitor	0.559	0.178	0.0020
Remove patients with treatment for SLE	0.491	0.199	0.014
comparison within the case			
Variable	Effect size	SE	P_{KEGG}
Not newly diagnosed	0.568	0.409	0.18
Lupus nephritis	-0.223	0.241	0.37
SLE-DAI ≥ 8	-0.019	0.019	0.34

HMPREF0833_10768

Case-control comparison			
Sample set	Effect size	SE	P_{KEGG}
All	0.850	0.141	6.7×10^{-9}
Remove male	0.873	0.181	4.5×10^{-6}
Remove patients with antibiotics	0.681	0.151	1.1×10^{-5}
Remove patients with proton pump inhibitor	0.700	0.180	1.3×10^{-4}
Remove patients with treatment for SLE	0.738	0.198	2.6×10^{-4}
comparison within the case			
Variable	Effect size	SE	P_{KEGG}
Not newly diagnosed	0.266	0.433	0.55
Lupus nephritis	-0.151	0.247	0.55
SLE-DAI ≥ 8	-0.010	0.020	0.61

SE, standard error; SLE, systemic lupus erythematosus; SLE-DAI, SLE Disease Activity Index.

Supplementary Table 4. Pathways with significant SLE case-control discrepancy.

KEGG pathway	Set size	$P_{pathway}$	q	Definition
ko00500	2,611	1.5×10^{-5}	0.0015	Starch and sucrose metabolism
B3	4,913	2.3×10^{-5}	0.0015	Energy metabolism
ko00920	805	1.4×10^{-4}	0.0048	Sulfur metabolism
ko02040	320	1.5×10^{-4}	0.0048	Flagellar assembly
ko02010	6,502	0.0011	0.029	ABC transporter
B25	659	0.0017	0.035	Cell motility
ko00460	626	0.0020	0.036	Cyanoamino acid metabolism

KEGG, Kyoto Encyclopedia of Genes and Genomes.

Supplementary Table 5. SLE case-control and within-case comparisons for phylogenetic α-diversity.

Case-control comparison

	Phylum (L2)			Class (L3)		
Sample set	Effect size	SE	$P_{\alpha\text{-diversity}}$	Effect size	SE	$P_{\alpha\text{-diversity}}$
All	-0.032	0.022	0.14	-0.075	0.032	0.019
Remove male	-0.040	0.025	0.11	-0.083	0.037	0.027
Remove patients with antibiotics	-0.021	0.024	0.39	-0.059	0.035	0.095
Remove patients with proton pump inhibitor	-0.043	0.030	0.15	-0.130	0.043	0.0028
Remove patients with treatment for SLE	-0.031	0.034	0.37	-0.140	0.048	0.0040
	Order (L4)			Family (L5)		
Sample set	Effect size	SE	$P_{\alpha\text{-diversity}}$	Effect size	SE	$P_{\alpha\text{-diversity}}$
All	-0.072	0.034	0.038	-0.181	0.038	3.6×10^{-6}
Remove male	-0.083	0.040	0.042	-0.191	0.042	1.4×10^{-5}
Remove patients with antibiotics	-0.056	0.039	0.15	-0.142	0.042	7.9×10^{-4}
Remove patients with proton pump inhibitor	-0.135	0.047	0.0043	-0.265	0.052	7.9×10^{-7}
Remove patients with treatment for SLE	-0.138	0.052	0.0090	-0.223	0.060	2.4×10^{-4}
	Genus (L6)			Species (L7)		
Sample set	Effect size	SE	$P_{\alpha\text{-diversity}}$	Effect size	SE	$P_{\alpha\text{-diversity}}$
All	-0.332	0.054	4.2×10^{-9}	-0.164	0.057	0.0042
Remove male	-0.337	0.060	8.7×10^{-8}	-0.157	0.063	0.014
Remove patients with antibiotics	-0.278	0.060	5.5×10^{-6}	-0.122	0.063	0.052
Remove patients with proton pump inhibitor	-0.396	0.076	5.1×10^{-7}	-0.270	0.078	6.1×10^{-4}
Remove patients with treatment for SLE	-0.335	0.088	1.8×10^{-4}	-0.206	0.090	0.022

comparison within the case

	Phylum (L2)			Class (L3)		
Variable	Effect size	SE	$P_{\alpha\text{-diversity}}$	Effect size	SE	$P_{\alpha\text{-diversity}}$
Not newly diagnosed	0.048	0.060	0.43	0.168	0.092	0.074
Lupus nephritis	-0.007	0.050	0.89	0.015	0.080	0.85
SLE-DAI ≥ 8	-0.048	0.058	0.42	-0.034	0.093	0.71
	Order (L4)			Family (L5)		
Variable	Effect size	SE	$P_{\alpha\text{-diversity}}$	Effect size	SE	$P_{\alpha\text{-diversity}}$
Not newly diagnosed	0.173	0.101	0.093	0.087	0.098	0.38
Lupus nephritis	0.028	0.087	0.75	0.022	0.082	0.79
SLE-DAI ≥ 8	-0.046	0.102	0.66	0.035	0.096	0.72
	Genus (L6)			Species (L7)		
Variable	Effect size	SE	$P_{\alpha\text{-diversity}}$	Effect size	SE	$P_{\alpha\text{-diversity}}$
Not newly diagnosed	0.039	0.138	0.78	0.012	0.146	0.93
Lupus nephritis	-0.027	0.115	0.82	0.064	0.122	0.60
SLE-DAI ≥ 8	0.093	0.134	0.49	0.205	0.140	0.15

SE, standard error; SLE, systemic lupus erythematosus.

Supplementary Table 6. SLE case-control and within-case comparisons for genetic α-diversity.

Case-control comparison			
Sample set	Effect size	SE	$P_{\alpha\text{-diversity}}$
All	-0.228	0.057	7.9×10^{-5}
Remove male	-0.210	0.063	0.0011
Remove patients with antibiotics	-0.181	0.062	0.0042
Remove patients with proton pump inhibitor	-0.344	0.078	1.7×10^{-5}
Remove patients with treatment for SLE	-0.295	0.092	0.0015
comparison within the case			
Variable	Effect size	SE	$P_{\alpha\text{-diversity}}$
Not newly diagnosed	0.038	0.138	0.78
Lupus nephritis	0.064	0.115	0.58
SLE-DAI ≥ 8	0.153	0.133	0.26

SE, standard error; SLE, systemic lupus erythematosus.

Supplementary Table 7. SLE case-control and within-case comparisons for phylogenetic β-diversity.

Group A	Group B	*P* (PERMANOVA)					
		Phylum (L2)	Class (L3)	Order (L4)	Family (L5)	Genus (L6)	Species (L7)
HC	SLE	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
HC (Only female)	SLE (Only female)	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
HC	SLE (Remove patients with antibiotics)	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
HC	SLE (Remove patients with proton pump inhibitor)	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
HC	SLE (Remove patients with treatment for SLE)	0.0013	0.0013	8.4×10^{-4}	2.0×10^{-5}	4.0×10^{-5}	3.0×10^{-5}
SLE (Newly diagnosed)	SLE (Not newly diagnosed)	0.32	0.33	0.33	0.11	0.084	0.12
SLE (without LN)	SLE (with LN)	0.94	0.72	0.67	0.70	0.69	0.57
SLE (SLE-DAI < 8)	SLE (SLE-DAI ≥ 8)	0.12	0.10	0.10	0.027	0.032	0.11

HC, healthy control; SE, standard error; SLE, systemic lupus erythematosus; LN, lupus nephritis.

Supplementary Table 8. SLE case-control and within-case comparisons for genetic β-diversity.

Group A	Group B	*P* (PERMANOVA)
HC	SLE	1.0×10^{-5}
HC (Only female)	SLE (Only female)	1.0×10^{-5}
HC	SLE (Remove patients with antibiotics)	1.0×10^{-5}
HC	SLE (Remove patients with proton pump inhibitor)	1.0×10^{-5}
HC	SLE (Remove patients with treatment for SLE)	3.0×10^{-5}
SLE (Newly diagnosed)	SLE (Not newly diagnosed)	0.23
SLE (without LN)	SLE (with LN)	0.34
SLE (SLE-DAI < 8)	SLE (SLE-DAI ≥ 8)	0.14

HC, healthy control; SE, standard error; SLE, systemic lupus erythematosus; LN, lupus nephritis.

METHODS

Patient participation

We examined the 48 SLE patients at Osaka University Hospital and the National Hospital Organization Osaka Minami Medical Center. The SLE patients were diagnosed according to the systemic lupus international collaborating clinics classification criteria (SLICC),[1]. The 205 healthy controls were enrolled at Osaka University Graduate School of Medicine, Osaka University Hospital and National Hospital Organization Osaka Minami Medical Center. The healthy controls had no personal history of immune-related diseases. Among the healthy control samples, 27 samples which belonged to the sequencing group1 were derived from a previous study,[2].

Participants with extreme diets (e.g., strict vegetarians) and the healthy controls who were treated with antibiotics for at least a month prior to sampling were removed from analysis. Characteristics of the subjects are described in **online supplementary table 1**. All subjects provided written informed consent before participation. The study protocol was approved by the ethical committees of Osaka University and related medical institutions. In microbe-metabolome association analysis, we re-analyzed phylogenetic abundance data from the 75 healthy control subjects in a previously published study,[3].

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 -80°C within 24 hours after collection. Bacterial DNA was extracted according to a previously described method,[2]. Briefly, 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 20,000 g 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 100 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 HiSeq 3000 for the sequencing group1 (average 25.8 million paired end reads per sample) and group2 (average 24.4 million paired end reads per sample), on Novaseq 6000 for the sequencing group3 (average 29.6 million paired end reads per sample). Healthy control data from the sequencing group1 was also used in the previous study,[2]. The sequence reads were converted to FASTQ format using bcl2fastq (version 2.19).

Quality control of sequencing reads

We followed a series of the QC 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 marked duplicate reads using PRINSEQ-lite (version 0.20.4; parameters: -derep 1). We trimmed the raw reads to clip Illumina adapters and cut off low-quality bases at both ends using the Trimmomatic (version 0.39; 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 performed duplicate removal by retaining only the longest read among the duplicates with same sequences. As a final QC step, we aligned the quality-filtered reads to the human reference genome (hg38) using bowtie2 (version 2.3.5) with default parameters and BMTagger (version 3.101). We kept only reads of which both paired ends failed to align in either tool.

Taxonomic annotation of metagenome and abundance quantification

To improve both the efficiency and the accuracy of taxonomic assignment, we used curated reference metagenomes as previously described,[2]. The reference metagenomes of the Japanese population constructed by Nishijima et al.,[4] were combined with the metagenomes identified 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 dataset using bowtie2 with default parameters. 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). Then, we detected outlier samples by PCA.

Functional annotation and abundance calculation

De novo assembly of the filtered paired-end reads into contigs was conducted using MEGAHIT (version 1.2.9; parameters: --min-contig-len 135). We predicted open reading frames (ORFs) on the contigs with the *ab initio* gene finder MetageneMark (version 3.38; parameters: -a -k -f G). Next, we annotated the ORF catalog with Kyoto Encyclopedia of Genes and Genomes (KEGG) protein database (https://www.kegg.jp),[8]. 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 KEGG protein database with DIAMOND using BLASTP (version v0.9.32.133; parameters: f 6 -b 15.0-k 1 -e 1e-6 --subject-cover 50). We identified 1,814,229 KEGG genes. For quantification of the

ORF abundance, we mapped the filtered paired-end reads to the assembled contigs using bowtie2 with default parameters. 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. Then, we detected outlier samples by PCA. Two healthy control samples and an SLE samples were removed from analysis because they were outliers in PCA analysis on the phylogenetic data or the KEGG gene abundance data (**online supplementary figure 1**).

Case-control association test for phylogenetic data

We normalized the relative abundance profiles using log transformation. We removed clades detected (i) in less than 20% of the samples, (ii) in no sample in either cases or controls, (iii) in no sample in any of the three sequencing groups, or (iv) with an average relative abundance of less than 0.001% of total abundance. After selection, we assessed 774 clades (12 phyla, 25 classes, 36 orders, 72 families, 178 genera, and 451 species). Case-control association tests were performed separately for each standardized clade abundance using the lm() function in the R (version 4.0.1) and the effect size of disease state was evaluated. Abundances of the taxa were set as objective variables and distributions of the residuals from linear regression models were checked manually for the significantly changed taxa to confirm the satisfaction of the assumption of normality. We adopted sex, age, sequencing group and the top 25 principal components as covariates. To evaluate the effect of potential confounding factors, we performed sub-analysis with dataset from which (i) males were removed, (ii) those treated with antibiotics were removed, (iii) those treated with proton pump inhibitors were removed or (iv) those treated for the SLE were removed. In addition, the SLE patients were stratified according to (i) newly diagnosed or not, (ii) presence of lupus nephritis, and (iii) SLE-DAI ≥ 8 or not. Then, comparison within case was performed.

Case-control association test for gene abundance data

We converted each ORF abundance to annotated gene abundance for 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, (ii) in no sample in either cases or controls, or (iii) in no sample in any of the three sequencing groups. After gene selection, we assessed 240,544 genes annotated by KEGG gene database. Case-control association tests were performed separately for each clade using the lm() function in the R and the effect size of the disease state was evaluated. Abundances of the genes were set as objective variables and distributions of the residuals from linear regression models were checked manually for the significantly changed taxa to confirm the satisfaction of the assumption of normality. We adopted sex, age, sequencing group and the top 25 principal components as covariates. To evaluate the effect of potential confounding factors, we performed sub-analysis with the dataset from which (i) males were removed, (ii) those treated with antibiotics were removed, (iii) those treated with proton pump inhibitors were removed, or (iv) those treated for SLE were removed. In addition, the SLE patients were stratified according to (i) newly diagnosed or not, or (ii) presence of lupus nephritis, and (iii) SLE-DAI ≥ 8 or not. Then, comparison within case was performed.

Biological pathway enrichment analysis of the gut microbiome

We performed a gene set enrichment analysis using the R package fgsea (version 1.16.0). 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 KEGG database were ranked based on their z-value in case-control gene association tests.

Tomofuji Y, *et al. Ann Rheum Dis* 2021;0:1–9. doi: 10.1136/annrheumdis-2021-220687

The KEGG gene sets were defined according to the KEGG pathway. False discovery ratio (FDR) was calculated by Benjamini-Hochberg procedure.

Comparison of the pathway analysis results between the MWAS and the GWAS

We assessed whether there existed shared biological pathways between the gut metagenome and the human germline genome. We compared the pathway enrichment data of the metagenome in the SLE patients with the GWAS of SLE and RA. For GWAS, we used PASCAL,[9] with summary statistics from the SLE GWAS,[10] (4,943 cases and 8,483 controls) and the RA GWAS,[11] (14,361 cases and 43,923 controls) in the European population in order to determine KEGG pathway enrichment of the human genome. Based on the biological pathway enrichment analysis from MWAS and GWAS, we classified biological pathways into the following four groups; (1) Biological pathways enriched in MWAS and GWAS results ($P_{pathway} < 0.05$ in MWAS and $P_{pathway} < 0.05$ in GWAS) (2) Biological pathways enriched only in MWAS result ($P_{pathway} < 0.05$ in MWAS and $P_{pathway} \geq 0.05$ in GWAS). (3) Biological pathways enriched only in GWAS result ($P_{pathway} \geq 0.05$ in MWAS and $P_{pathway} < 0.05$ in GWAS). (4) Biological pathways not enriched in MWAS and GWAS results ($P_{pathway} \geq 0.05$ in MWAS and $P_{pathway} \geq 0.05$ in GWAS). Then, we evaluated the enrichment of the number of biological pathways in the group (1) compared to the group (2~4) by a one-tailed Fisher's exact test (**online supplementary figure 2**).

Empirical estimation of the metagenome-wide significance threshold

We empirically estimated the statistical significance threshold separately for the phylogenetic and the gene case-control analyses, performing a phenotype permutation procedure,[12]. We randomly simulated case-control phenotypes (× 10,000 iterations) and calculated empirical null distributions of the minimum p-values (= P_{min}) in each iteration. We defined an empirical Bonferroni significance threshold at a significance level of 0.05, as the 95th

percentile of P_{min} (= P_{sig}). We calculated the P_{sig} using the Harrell-Davis distribution-free quantile estimator and calculated a 95% confidence interval for P_{sig} by a bootstrapping method in the R package Hmisc (version 4.4_0). To estimate the null distribution of the test statistics, we applied the same process used for minimum p-values to all the p-values above each rank. We defined an empirical FDR threshold of 0.05 as the 95th percentile of all the $-\log_{10}(P)$ above each rank.

Case-control difference between α-diversity and β-diversity of the metagenome

α-diversity (within-sample diversity) was calculated as a Shannon index based on the gene abundance and the six levels of phylogenetic relative abundance (L2-L7) for each sample. Case-control comparison were performed with the lm() function in the R and the effect size of disease state was evaluated. We adopted sex, age and sequencing group as covariates. To evaluate β-diversity, multidimensional scaling (MDS) on the Bray-Curtis dissimilarity was performed. For evaluating case-control differences in the dissimilarity, we performed permutational multivariate analysis of variance,[13] (PERMANOVA) with 99,999 permutations using the adonis() function in R package vegan (version 2.5_6). Sex, binary transformed age (<50 or ≥50) and sequencing group were adjusted as covariate before the variance explained by disease state was evaluated.

Participants of metabolome analysis

We used the previously obtained plasma metabolite data for microbe-metabolite association analysis,[14]. Among the participants, 31 subjects (27 healthy controls, 4 cases) from sequencing group1 and 72 subjects (67 healthy controls, 5 cases) from the sequencing group 2 were profiled their plasma metabolite by CE-TOFMS and LC-TOFMS in the previous study,[14]. In addition, 75 healthy control subjects who belong to replication dataset were also profiled their plasma metabolite in the same study.

Sample collection and metabolome profiling

Details were described in Kishikawa et al,[15]. Briefly, plasma samples were collected in BDTM P100 Blood Collection Tubes (Franklin Lakes, NJ, USA), which contains spray-dried K2 EDTA anticoagulant, proprietary proteinase inhibitors and a mechanical separator to obtain plasma. Within 1h of collection, the samples were centrifuged for 15 min at 2500 × g and stored at -80°C until analysis. Metabolite extraction and metabolome analysis were conducted at Human Metabolome Technologies (HMT), Japan. For CE-TOFMS measurement, 50 µl of plasma was added to 450 µl of methanol containing internal standards (H3304-1002, HMT) at 0°C to inactivate enzymes. The extract solution was thoroughly mixed with 500 µl of chloroform and 200 µl of Milli-Q water and centrifuged at 2300 × g and 4°C for 5 min. The 350 µl of the upper aqueous layer was centrifugally filtered through a Millipore 5-kDa cutoff filter to remove proteins. The filtrate was centrifugally concentrated and re-suspended in 50 µl of Milli-Q water for CE-MS analysis. For LC-TOFMS measurement, 500 µl of plasma was added to 1500 µl of 1% formic acid/acetonitrile containing internal standard solution (Solution ID: H3304-1002, Human Metabolome Technologies, Inc., Tsuruoka, Japan) at 0°C to inactivate enzymes. The solution was thoroughly mixed and centrifuged at 2300 × g and 4°C for 5 min. The supernatant was filtrated by using Hybrid SPE phospholipid (55261-U, Supelco, Bellefonte, PA, USA) to remove phospholipids. The filtrate was desiccated and then dissolved with 100 µl of iso- propanol/Milli-Q for LC-MS analysis.

Metabolome analysis was conducted by the Dual Scan package of HMT using CE-TOFMS and LC-TOFMS for ionic and nonionic metabolites, respectively. CE-TOFMS analysis was carried out using an Agilent CE system equipped with an Agilent 6210 TOFMS, Agilent 1100 isocratic HPLC pump, Agilent G1603A CE-MS adapter kit and Agilent G1607A CE-ESI-MS sprayer kit (Agilent Technologies, Santa Clara, CA, USA). The systems were controlled by Agilent G2201AA ChemStation software version B.03.01 for CE (Agilent

Technologies) and connected by a fused silica capillary (50 μm i.d. × 80 cm total length) with commercial electrophoresis buffer (H3301-1001 and I3302-1023 for cation and anion analyses, respectively, HMT) as the electrolyte. The spectrometer was scanned from m/z 50 to 1000. LC-TOFMS analysis was carried out using an Agilent LC System (Agilent 1200 series RRLC system SL) equipped with an Agilent 6230 TOFMS (Agilent Technologies). The systems were controlled by Agilent G2201AA ChemStation software version B.03.01 (Agilent Technologies) equipped with ODS column (2 × 50 mm, 2 μm). Peaks were extracted using MasterHands, automatic integration software (Keio University, Tsuruoka, Yamagata, Japan) to obtain peak information including m/z, peak area and migration time for CE-TOFMS measurement (MT) or retention time for LC-TOFMS measurement (RT). Signal peaks corresponding to isotopomers, adduct ions and other product ions of known metabolites were excluded, and remaining peaks were annotated according to the HMT metabolite database based on their m/z values with the MTs and RTs determined by TOFMS. Areas of the annotated peaks were normalized based on internal standard levels and sample amounts to obtain relative levels of each metabolite.

Microbe-metabolome association analysis

541 and 512 metabolites were detected in the discovery and the replication dataset. 398 and 385 metabolites which were detected in ≥20% of the samples were retained for analysis. The abundance of metabolite was normalized using log transformation. Microbe-metabolite association tests were performed separately for each microbe-metabolite pair using the lm() function in the R and the effect size of the microbe abundance was evaluated. In the discovery dataset, we adopted sex, age, sequencing group and the top principal component of microbe and the top five principal components of metabolite as covariates. In the replication dataset, we adopted sex, age and the top principal component of microbe and the top five principal

components of metabolite as covariates. FDR was calculated by Benjamini-Hochberg procedure.

Power calculation

Power calculation was performed for phylogenetic and functional analysis using the wp.regression() function in R package WebPower (version 0.6) considering of pre-estimated effect sizes ranged from 0.01 to 0.35 (represented as Cohen's f^2). To account for the case-control balance in our dataset, we calculated effective sample size as $N_{eff} = 4 / (1 / N_{case} + 1 / N_{control}) = 152.66 \simeq 150$, according to the previous report,[16]. According to Cohen's (1988) guidelines,[17], $f^2 \geq 0.02$, $f^2 \geq 0.15$, $f^2 \geq 0.35$ represent small, medium, and large effect sizes, respectively.

Patient and public involvement

This research was done without patient and public involvement. Patients and public were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results.

URLs

bcl2fastq, https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software/downloads.html

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/Genemark/meta_gmhmmp.cgi

DIAMOND, https://github.com/bbuchfink/diamond

Samtools, http://www.htslib.org/download/

bedtools, https://github.com/arq5x/bedtools2

R, https://www.r-project.org

Python, https://www.python.org/downloads/release/python-376/

Seqkit, https://bioinf.shenwei.me/seqkit/download/

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

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

fgsea, https://github.com/ctlab/fgsea

WebPower, https://CRAN.R-project.org/package=WebPower

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

KEGG, https://www.kegg.jp

REFERENCES

1 Petri M, Orbai A-M, Alarcón GS, *et al.* Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. *Arthritis & Rheumatism* 2012;64:2677–86. doi:10.1002/art.34473

2 Kishikawa T, Maeda Y, Nii T, *et al.* Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population. *Ann Rheum Dis* 2020;79:103–11. doi:10.1136/annrheumdis-2019-215743

3 Kishikawa T, Ogawa K, Motooka D, *et al.* A Metagenome-Wide Association Study of Gut Microbiome in Patients With Multiple Sclerosis Revealed Novel Disease Pathology. *Front Cell Infect Microbiol* 2020;10:585973. doi:10.3389/fcimb.2020.585973

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. doi:10.1093/dnares/dsw002

5 Almeida A, Mitchell AL, Boland M, *et al.* A new genomic blueprint of the human gut microbiota. *Nature* 2019;568:499–504. doi:10.1038/s41586-019-0965-1

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. doi:10.1038/s41587-018-0009-7

7 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. doi:10.1038/s41587-018-0008-8

8 Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. ;:4.

9 Lamparter D, Marbach D, Rueedi R, *et al.* Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol* 2016;12:e1004714. doi:10.1371/journal.pcbi.1004714

10 Julià A, López-Longo FJ, Pérez Venegas JJ, *et al.* Genome-wide association study meta-analysis identifies five new loci for systemic lupus erythematosus. *Arthritis Res Ther* 2018;20:100. doi:10.1186/s13075-018-1604-1

11 the RACI consortium, the GARNET consortium, Okada Y, *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014;506:376–81. doi:10.1038/nature12873

12 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;61:861–6. doi:10.1038/jhg.2016.72

13 Anderson MJ. A new method for non-parametric multivariate analysis of variance: NON-PARAMETRIC MANOVA FOR ECOLOGY. *Austral Ecology* 2001;26:32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x

14 Kishikawa T, Maeda Y, Nii T, *et al.* Increased levels of plasma nucleotides in patients with rheumatoid arthritis. *International Immunology* 2021;33:119–24. doi:10.1093/intimm/dxaa059

15 Kishikawa T, Arase N, Tsuji S, *et al.* Large-scale plasma-metabolome analysis identifies potential biomarkers of psoriasis and its clinical subtypes. *Journal of Dermatological Science*

16 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;26:2190–1. doi:10.1093/bioinformatics/btq340

17 Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, N.J: : L. Erlbaum Associates 1988.