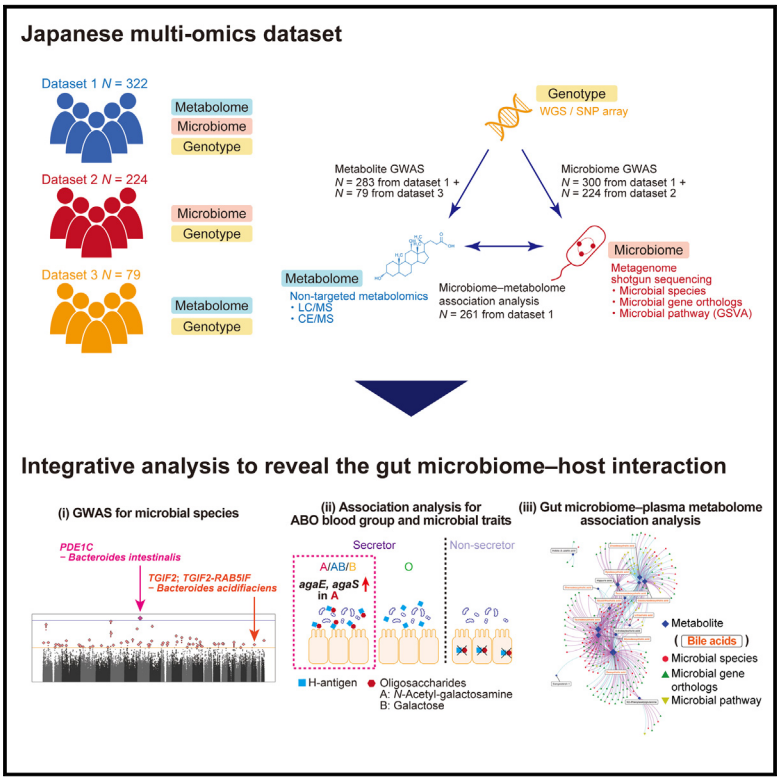
Cell Reports

Resource

Analysis of gut microbiome, host genetics, and plasma metabolites reveals gut microbiome-host interactions in the Japanese population

Graphical abstract



Highlights

- Construction of Japanese multi-omics datasets (gut microbiome, plasma metabolome, host genome)
- Genome-wide association analysis of the gut microbiome with 524 Japanese individuals
- ABO blood type was associated with the microbial genes related to N-galactosamine metabolism
- Microbiome-metabolome association analysis highlights the association involving bile acids

Authors

Yoshihiko Tomofuji, Toshihiro Kishikawa, Kyuto Sonehara, ..., Kiyoshi Takeda, Atsushi Kumanogoh, Yukinori Okada

Correspondence

ytomofuji@sg.med.osaka-u.ac.jp (Y.T.), yokada@sg.med.osaka-u.ac.jp (Y.O.)

In brief

Tomofuji et al. constructed Japanese multi-omics datasets (gut microbiome, plasma metabolome, and host genome). Their shotgun-sequencing-based gut microbiome analysis identified associations with the host genome and plasma metabolome including those involving microbial genes. Multi-omics analysis in the underrepresented population would contribute to expanding the diversity of the studied populations.

Tomofuji et al., 2023, Cell Reports *42*, 113324
November 28, 2023
https://doi.org/10.1016/j.celrep.2023.113324

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

Analysis of gut microbiome, host genetics, and plasma metabolites reveals gut microbiome-host interactions in the Japanese population

Yoshihiko Tomofuji,[1,2,3,4,21,*] Toshihiro Kishikawa,[1,5,6,21] Kyuto Sonehara,[1,2,3,4] Yuichi Maeda,[3,7,8] Kotaro Ogawa,[9] Shuhei Kawabata,[10] Eri Oguro-Igashira,[7,8] Tatsusada Okuno,[9] Takuro Nii,[7,8] Makoto Kinoshita,[9] Masatoshi Takagaki,[10] Kenichi Yamamoto,[1,11,12] Noriko Arase,[13] Mayu Yagita-Sakamaki,[7,8] Akiko Hosokawa,[9,14] Daisuke Motooka,[3,15] Yuki Matsumoto,[15] Hidetoshi Matsuoka,[16] Maiko Yoshimura,[16] Shiro Ohshima,[16] Shota Nakamura,[3,15,17] Manabu Fujimoto,[13] Hidenori Inohara,[5] Haruhiko Kishima,[10] Hideki Mochizuki,[9] Kiyoshi Takeda,[8,17,18] Atsushi Kumanogoh,[3,7,19] and Yukinori Okada[1,2,3,4,12,17,20,22,*]

[1]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[2]Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Tsurumi 230-0045, Japan
[3]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan
[4]Department of Genome Informatics, Graduate School of Medicine, the University of Tokyo, Tokyo 113-8654, Japan
[5]Department of Otorhinolaryngology-Head and Neck Surgery, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[6]Department of Head and Neck Surgery, Aichi Cancer Center Hospital, Nagoya 464-8681, Japan
[7]Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[8]Laboratory of Immune Regulation, Department of Microbiology and Immunology, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[9]Department of Neurology, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[10]Department of Neurosurgery, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[11]Department of Pediatrics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[12]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan
[13]Department of Dermatology, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[14]Department of Neurology, Suita Municipal Hospital, Suita 564-8567, Japan
[15]Department of Infection Metagenomics, Research Institute for Microbial Diseases, Osaka University, Suita 565-0871, Japan
[16]Department of Rheumatology and Allergology, NHO Osaka Minami Medical Center, Kawachinagano 586-8521, Japan
[17]Center for Infectious Disease Education and Research, Osaka University, Suita 565-0871, Japan
[18]WPI Immunology Frontier Research Center, Osaka University, Suita 565-0871, Japan
[19]Department of Immunopathology, Immunology Frontier Research Center, Osaka University, Suita 565-0871, Japan
[20]Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), Osaka University, Suita 565-0871, Japan
[21]These authors contributed equally
[22]Lead contact
*Correspondence: ytomofuji@sg.med.osaka-u.ac.jp (Y.T.), yokada@sg.med.osaka-u.ac.jp (Y.O.)
https://doi.org/10.1016/j.celrep.2023.113324

SUMMARY

Interaction between the gut microbiome and host plays a key role in human health. Here, we perform a metagenome shotgun-sequencing-based analysis of Japanese participants to reveal associations between the gut microbiome, host genetics, and plasma metabolome. A genome-wide association study (GWAS) for microbial species (n = 524) identifies associations between the *PDE1C* gene locus and *Bacteroides intestinalis* and between *TGIF2* and *TGIF2-RAB5IF* gene loci and *Bacteroides acidifiaciens*. In a microbial gene ortholog GWAS, *agaE* and *agaS*, which are related to the metabolism of carbohydrates forming the blood group A antigen, are associated with blood group A in a manner depending on the secretor status determined by the East Asian-specific *FUT2* variant. A microbiome-metabolome association analysis (n = 261) identifies associations between bile acids and microbial features such as bile acid metabolism gene orthologs including *bai* and 7β-hydroxysteroid dehydrogenase. Our publicly available data will be a useful resource for understanding gut microbiome-host interactions in an underrepresented population.

INTRODUCTION

The human gut microbiome is the collection of the microbes that reside within our gut. The gut microbiome is closely related to human health, and associations with various diseases such as inflammatory bowel diseases, autoimmune diseases, and metabolic diseases have been reported.[1–5] Although the gut microbiome interacts with the host via various mechanisms such as

the modulation of the host's immune system and production of the metabolites,[6] detailed insights into the gut microbiome-host interaction remain to be revealed.

Host genetics play a role in the interaction between the gut microbiome and the host. A twin study revealed that the abundances of the gut bacteria were more similar in monozygotic twins than in dizygotic twins, suggesting that host genetics affected the gut bacterial abundances.[7] To identify the individual genetic variants that affect gut microbial abundances, several genome-wide association studies (GWASs) have been performed.[7–17] In these studies, the *ABO* and *LCT* gene loci have been repeatedly reported.

Currently, most of the GWASs for the gut microbiome-associated traits have been conducted in European (EUR) populations and have rarely been performed in East Asian (EAS) populations, especially other than the Chinese population.[11] Given that the gut microbiome and human genetic variants are different between populations, elucidating the association between the microbial traits and host genetics in EAS populations is important to increase the diversity of the study populations and deepen insights into the gut microbiome-host interaction.

It is also important to analyze the gut microbiome-host genetics association with metagenome shotgun sequencing, which has several benefits over 16S ribosomal RNA (rRNA) sequencing. Metagenome shotgun sequencing enables us to evaluate the gut microbiome with species-level resolution, which can be hardly achieved by 16S rRNA sequencing. Given that previous metagenome-wide association studies for diseases have identified disease-gut microbiome associations that would be undetectable without species-level resolution,[2,4,18] metagenome shotgun sequencing is necessary for the comprehensive understanding of the gut microbiome-host interaction. Furthermore, metagenome shotgun sequencing enables us to obtain functional information such as microbial gene orthologs and pathways, which could not be obtained by 16S rRNA sequencing. Since the microbial gene orthologs and pathways are often shared across different microbial taxa, gene-ortholog- and pathway-level analysis may bring us functionally interpretable associations that could not have been identified by analysis solely based on microbial taxa. Despite these benefits of metagenome shotgun sequencing, most of the previous studies focusing on the gut microbiome-host genetics association utilized 16S rRNA sequencing,[8–10,16,19] and the transition from 16S rRNA sequencing to metagenome shotgun sequencing is still in progress.

Metabolites also take an important role in the gut microbiome-host interaction. The metabolic activities of the gut microbiome can contribute to human complex traits by affecting metabolites such as nutrients and bile acids.[6] Metabolites also affect the gut microbiome because they can be nutrients or cytotoxic detergents for the gut microbiome.[20] Although previous studies have investigated the gut microbiome-blood metabolite association, mainly in EUR populations,[21–24] there are still few studies focusing on the gut microbiome-metabolome association in EAS populations.[11] In addition, it was often difficult to functionally interpret the gut microbiome-metabolome association based on analysis solely focused on microbial taxa, suggesting the necessity for functional information obtained from metagenome shotgun sequencing.

Here, we investigated the gut microbiome, plasma metabolites, and host genetics of Japanese participants by gut metagenome shotgun sequencing, non-targeted metabolomic profiling, and genotyping with single-nucleotide polymorphism (SNP) array and whole-genome sequencing (WGS; Figure S1). Utilizing these large-scale and comprehensive data, we evaluated the gut microbiome-host genetics (n = 524) and -plasma metabolites (n = 261) associations.

RESULTS

Genome-wide association analysis of the gut microbial species, gene orthologs, and pathways

We performed genome-wide association analysis for microbial traits (species, gene orthologs, and pathways) with two datasets generated in different periods (dataset 1 [gut microbiome, plasma metabolome, and genotype]: n = 300, dataset 2 [gut microbiome and genotype]: n = 224; Figure S1; Table S1, STAR Methods), followed by a fixed-effect meta-analysis. We analyzed 7,213,469 SNP-array-based variants that fulfilled stringent post-imputation quality control criteria (minor allele frequency [MAF] > 1% and Rsq by Minimac4 > 0.7; STAR Methods).[25] In the analysis for the 423 microbial species, an association between chr7:32016991:C>G in the *PDE1C* gene locus and *Bacteroides intestinalis* satisfied the study-wide significance threshold (Figures 1A and 1B; Tables 1 and S2; effect size = 0.987, SE = 0.152, $p = 7.2 \times 10^{-11} < 5.0 \times 10^{-8}/423$ species $= 1.18 \times 10^{-10}$). The effects of chr7:32016991:C>G on *Bacteroides intestinalis* abundance were consistent in both datasets 1 and 2 (effect size = 0.857, SE = 0.205, $p = 3.8 \times 10^{-5}$ for dataset 1 and effect size = 1.15, SE = 0.225, $p = 7.9 \times 10^{-7}$ for dataset 2; Figure S2A; Table S2) and did not depend on the data transformation methods (Figure S2B). This variant had not been reported in the previous GWAS for the gut microbial traits, possibly due to the differences in the allele frequency between populations (8.4% in EAS and <1% in EUR, gnomAD v.2.1.1), while other factors such as environmental factors and methodological differences could also contribute. Three variants located in the *PDE1C* gene locus were reported as gut microbiome-associated variants in the GWAS catalog (chr7:32214192:C>T for genus *Dialister*, $p = 2 \times 10^{-7}$; chr7:31814020:A>G for genus *Aestuariispira*, $p = 9 \times 10^{-7}$; chr7:32483218:G>A for genus *Collinsella,* $p = 5 \times 10^{-6}$),[16] supporting the association between the *PDE1C* gene locus and the gut microbiome. *PDE1C* is a gene encoding a cyclic nucleotide phosphodiesterase related to the function of various cells such as olfactory sensory neurons.[26] In the GWAS catalog, the *PDE1C* gene locus was reported to be associated with several central nervous system (CNS)-related traits such as smoking,[27] body mass index,[28] and educational attainment,[29] suggesting that association between the *PDE1C* gene locus and *Bacteroides intestinalis* might be also mediated by the CNS.

Among the associations that satisfied genome-wide significance but did not satisfy study-wide significance ($1.18 \times 10^{-10} < p < 5 \times 10^{-8}$), chr20:35208051:T>C in the *TGIF2* and *TGIF2-RAB5IF* gene loci associated with *Bacteroides acidifiaciens* (effect size = −0.434, SE = 0.078, $p = 2.6 \times 10^{-8}$; Figures 1A and

CellPress
OPEN ACCESS

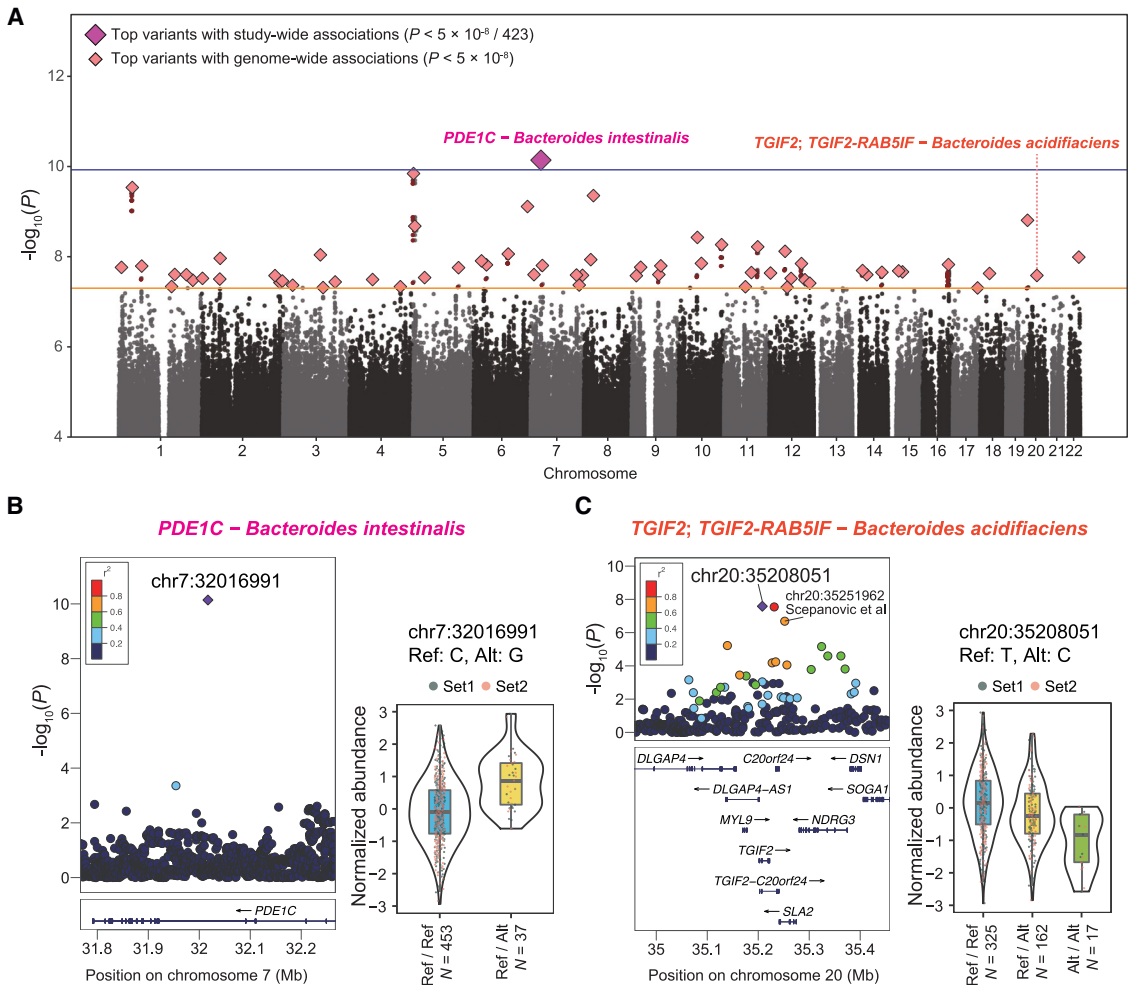


Figure 1. Genome-wide association analysis of the microbial species

(A) A Manhattan plot representing the result of the microbial species GWAS. The y axis indicates the −log10-transformed p values. The x axis indicates the genomic position of the variants. Only the trait-variant pairs satisfying $p < 1 \times 10^{-4}$ are plotted. The study-wide ($p < 5 \times 10^{-8}/423 = 1.18 \times 10^{-10}$) and genome-wide ($p < 5 \times 10^{-8}$) significances are indicated as horizontal dashed lines colored purple and orange, respectively. Lead variants satisfying study-wide or genome-wide significance are indicated as plink rhombuses. p values were calculated with the fixed-effect meta-analysis.

(B and C) Regional associations of the genetic variants at the *PDE1C* (B) and *TGIF2*/*TGIF2-RAB5IF* (C) gene loci are indicated (left). The purple diamonds indicate the lead variants. Other circles are colored by LD (r^2), with the lead variant based on the Japanese participants in the reference panel used for genotype imputation. p values were calculated with the fixed-effect meta-analysis. Violin and boxplots represent the normalized abundance of the microbial species per genotype (right). Boxplots indicate the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). The overlayed dots represent individual observations used for the association analysis. The color of the dots represents the dataset. Alt, alternative allele; GWAS, genome-wide association study; IQR, interquartile ranges; LD, linkage disequilibrium; Ref, reference allele.

See also Figures S2 and S3 and Tables S2, S3, S4, S5, and S6.

1C; Table 1) was tagged to chr20:35251962:A>C. The effects of chr20:35208051:T>C on *Bacteroides acidifiaciens* abundance were consistent in both datasets and did not depend on the data transformation methods (Figure S2; Table S2). The genetic variant chr20:35251962:A>C was previously reported to be associated with *Alistipes shahii* ($p = 3.2 \times 10^{-8}$).[8]

We also performed genome-wide association analyses for the 4,644 gut microbial gene orthologs and 146 gut microbial pathways, although no study-wide significant association was found (Figure S3A; Tables S3 and S4; significance thresholds were $p < 5.0 \times 10^{-8}/4{,}644$ gene orthologs $= 1.08 \times 10^{-11}$ and $p < 5.0 \times 10^{-8}/146$ pathway $= 3.42 \times 10^{-10}$, respectively). Inflation of the p values was not observed in the GWAS for the microbial traits (Figure S3B).

We evaluated whether the results of the previous studies could be replicated by our study. We evaluated the two gut metagenome shotgun-sequencing-based studies,[11,13] which classified the bacteria based on the NCBI taxonomy, and we could not replicate the reported genome-wide associations ($p > 0.05$; Table S5). In addition, we evaluated the association between

Table 1. Results summary of the microbiome GWAS

Bacterial species	Variant ID	rsID	Ref	Alt	Alt freq.	Effect size	SE	p	Q	p for Q	Gene
Bacteroides intestinalis	chr7:32016991	rs74338454	C	G	0.049	0.987	0.152	7.2×10^{-11}	0.17	0.68	*PDE1C*
Bacteroides acidifaciens	chr20:35208051	rs73620203	T	C	0.20	−0.434	0.078	2.6×10^{-8}	0.90	0.34	*TGIF2;TGIF2-RAB5IF*

Alt, alternative allele; freq., frequency; GWAS, genome-wide association study; Ref, reference allele; SE, standard error.

the previously reported loci ($p < 5 \times 10^{-8}$)[8–13,16] and all the microbial traits (Table S6). We found that only the *TGIF2* and *TGIF2-RAB5IF* gene loci-*Bacteroides acidifaciens* association discussed above passed the significance threshold after multiple-test correction ($p < 2.38 \times 10^{-7}$; Figure S3C). In addition, the *ABO* gene locus-*yydK* microbial gene ortholog association passed the significance threshold after per-study multiple-test correction (Figure S3D).

Association between the ABO blood group and microbial traits

The *ABO* and *LCT* gene loci were previously reported as gut microbial trait-associated loci in multiple studies.[9–13,16,17] Since the gut microbiome-associated *LCT* variant (chr2:136608646:G>A), which causes lactose intolerance, was very rare in EAS populations (0.064%, gnomAD v.2.1.1), it was difficult to evaluate the association between the *LCT* gene locus and the gut microbiome with our dataset. Therefore, we evaluated the association between the *ABO* gene locus (chr9:136146597:C>T linked to blood group A, chr9:136131322:G>T linked to blood group B, and chr9:136132908:T>TC linked to blood group O[30,31]) and the gut microbial traits. Three microbial gene orthologs and one microbial pathway were associated with the *ABO* gene locus after the correction with the number of the tested traits (Figures 2A, S4A, and S4B; $p < 1.18 \times 10^{-4}$ for the microbial species, $p < 1.08 \times 10^{-5}$ for the microbial gene orthologs, and $p < 3.42 \times 10^{-4}$ for the microbial pathways). Among these traits, *agaE*, *agaS*, and "metabolism of other amino acid" pathway were higher in blood group A than in blood groups B and O, while *yydK* was the opposite (Figures 2B, 2C, S4C, and S4D).

Among the bacterial gene orthologs involved in N-acetylgalactosamine metabolization, *agaE* and *agaS* associated with chr9:136146597:C>T were gene orthologs coding an N-acetylgalactosamine PTS system EIID component and D-galactosamine 6-phosphate deaminase, respectively.[32] agaE is a component of the N-acetylgalactosamine transport system that is necessary for bacteria to import the N-acetylgalactosamine. agaS is an enzyme necessary for metabolizing the galactosamine 6-phosphate, an N-acetylgalactosamine-derived product. Given that N-acetylgalactosamine, a terminal carbohydrate forming the antigen of blood group A, is also synthesized on the mucosal surfaces of the gut and secreted (Figure 2D), bacteria with *agaE* and *agaS* can utilize N-acetylgalactosamine as a nutrient, and their fitness can be affected by the ABO blood group of their hosts.

In previous studies for EUR populations, a loss-of-function variant of the *FUT2* gene, chr19:49206674:G>A, had significant effects on the association between the ABO blood group and gut microbial taxa because FUT2 is necessary to synthesize H-antigen on mucosal surfaces of the gut[10,12,13] (Figure 2D). Although chr19:49206674:G>A is very rare in EAS populations, chr19:49206631:A>T is an EAS-specific variant linked to the non-functional form of FUT2.[33,34] Therefore, we evaluated the effect of the secretor status determined by FUT2 (secretor, A/A or A/T for chr19:49206631:A>T; non-secretor, T/T for chr19:49206631:A>T) on the association between the ABO blood group and *agaE* and *agaS* abundances. We found that the abundances of *agaE* and *agaS* were significantly higher in the secretor than in the non-secretor in blood group A ($p = 1.6 \times 10^{-3}$ and 6.0×10^{-3}, respectively, for *agaE* and *agaS*; Table S7), and the relatively high abundance of these microbial traits in blood group A was not observed for the non-secretors (Figures 2E and 2F). Therefore, it was suggested that the association between blood group A and *agaE* and *agaS* was dependent on N-acetylgalactosamine in the gut. As for the association of the ABO blood group and *yydK* and the "metabolism of other amino acid" pathway, a significant contribution of the secretor status was not detected (Figures S4E and S4F; Table S7).

To evaluate which microbial taxa had *agaE* and *agaS* in the Japanese gut, we checked the Japanese Metagenome Assembled Genomes (JMAG), a database of the prokaryotic metagenome-assembled genomes (MAGs) recovered from 787 Japanese gut metagenome shotgun sequencing data,[35] including those used in this study. The most major origin of *agaE* and *agaS* was *Collinsella* at the genus level (23.5%), suggesting that the relatively high abundance of *Collinsella* in blood group A reported in previous studies[12,13] was possibly driven by *agaE* and *agaS* (Figures S5A–S5C). We also evaluated which microbial taxa had *yydK*, a GntR family transcriptional regulator gene ortholog. We found that the most major origin of *yydK* was *Bifidobacterium bifidum* at the species level (24.7%) which was reported to be decreased in blood group A,[13] indicating consistency with the previous study (Figures S5A–S5C).

In summary, we identified associations between microbial gene orthologs and pathways and the ABO blood group, including those that were functionally interpretable and could contribute to the mechanistic insight into the previously reported ABO blood group-microbial taxa associations.

Association between the gut microbial traits and plasma metabolites

Next, we focused on plasma metabolites as another interface of the gut microbiome-host interaction. We utilized a plasma metabolite dataset based on a comprehensive non-targeted metabolomics approach combining capillary electrophoresis time-of-flight mass spectrometry (CE-TOFMS) and liquid chromatography TOFMS (LC-TOFMS).[36] Then, we evaluated the

CellPress
OPEN ACCESS

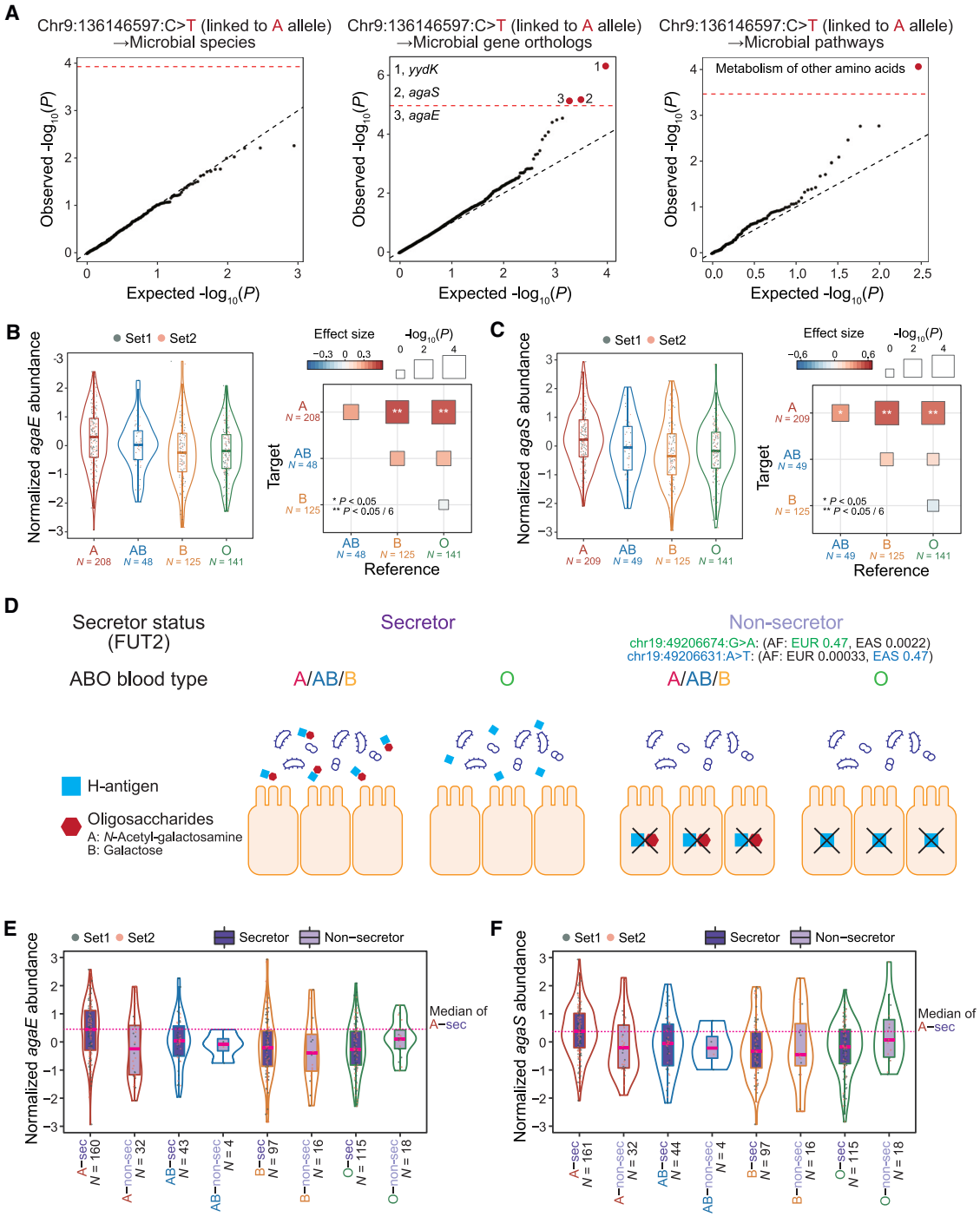


Figure 2. Association between the ABO blood group and microbial gene orthologs related to N-acetylgalactosamine metabolism

(A) Q-Q plots represent the associations between chr9:136146597:C>T (linked to blood group A allele) and microbial species (left), microbial gene orthologs (middle), and microbial pathways (right). The x axis indicates expected log-transformed p values, and the y axis indicates log-transformed observed p values. The diagonal dashed line represents y = x, which corresponds to the null hypothesis. The significance thresholds ($p < 0.05$/number of the tested traits) are indicated as horizontal dashed lines colored red. p values were calculated with the fixed-effect meta-analysis.

(B and C) Violin and boxplots represent the normalized abundance of *agaE* (B) and *agaS* (C) per blood group (left). Boxplots indicate the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). The overlayed dots represent individual observations used for the association analysis. The color of the dots represents the dataset. The associations between the microbial gene orthologs and the ABO blood group are also shown as boxes (right). The size and color of the boxes indicate the p values

(legend continued on next page)

association between 363 plasma metabolites and gut microbial traits (450 microbial species, 4760 microbial gene orthologs, and 148 microbial pathways; dataset 1: n = 261; Table S1). We found 246, 224, and 133 significant associations, respectively, for each class of microbial traits (Bonferroni-corrected $p < 0.05$; Figure 3A; Tables S8, S9, and S10). To get the whole picture of the gut microbiome-plasma metabolome association, we constructed a network plot from the significant associations. We found that a limited set of metabolites such as bile acids, indole-3-acetic acid, hippuric acid, 3-indoxylsulfurix acid, and N2-phenylacetylglutamine were associated with a large number of microbial traits (Figure 3B). To evaluate whether the genetic variants could contribute to the microbiome-metabolome association, we performed a plasma metabolite GWAS. Although we found EAS-specific metabolite-related variants that had pleiotropic associations with multiple complex traits (Figure S6; Tables S1 and S11; Data S1), no genetic variants had pleiotropic associations with both the microbial features and plasma metabolites pairs connected in the network ($p > 1 \times 10^{-5}$ for either of the microbial traits and plasma metabolites). The genetic variants that determine ABO blood groups did not have significant associations with the plasma metabolites (Figure S7).

bai gene orthologs take part in the conversion of primary bile acids (e.g., cholic acid and chenodeoxycholic acid) to secondary bile acids (e.g., deoxycholic acid and lithocholic acid).[20] In our analysis, five *bai* gene orthologs that were included in the network had significant positive associations with the deoxycholic acid and nominal negative associations with the chenodeoxycholic acid (Figure 3C). In addition, 7β-hydroxysteroid dehydrogenase, a hydroxysteroid dehydrogenase (HSDH) taking part in the conversion of chenodeoxycholic acid to ursodeoxycholic acid,[20] had a nominal positive association with ursodeoxycholic acid. It was reported that bile acid could affect the spore germination of bacteria (e.g., deoxycholic acid and chenodeoxycholic acid could promote and inhibit, respectively, the spore germination of *Clostridium difficile*).[37,38] We found that the abundances of the spore germination-related gene orthologs had a significant positive association with the deoxycholic acids and a nominal negative association with the chenodeoxycholic acid. Therefore, the function of the individual microbial gene orthologs was reflected in the gut microbiome-plasma metabolome interaction in humans.

To evaluate the association between bile acids and the overall gut microbial community, we performed a linear regression analysis between the α-diversity of the gut bacteria and plasma bile acids. We found multiple significant associations such as a positive association with deoxycholic acid and a negative association with chenodeoxycholic acid (Figure 3D).

DISCUSSION

In this study, we evaluated the association between the gut microbiome and host factors, namely plasma metabolome and host genetics, with multi-omics data for the Japanese population. We identified gut microbiome-associated variants that had study-wide significance or were replicated by previous study. We also identified the association between the ABO blood group and gut microbial traits such as *agaE* and *agaS*. In the microbiome-metabolome association analysis, we revealed that a specific set of metabolites, such as bile acid, had a significant association with a large number of gut microbial traits, including functionally interpretable gene orthologs.

A genetic variant, chr7:32016991:C>G, in the *PDE1C* gene locus associated with *Bacteroides intestinalis* was an EAS-specific variant, emphasizing the importance of performing gut microbiome GWASs in underrepresented populations, including EAS populations. Although we also identified genome-wide associations between the microbial traits and genetic variants, we should carefully interpret such associations because the reproducibility of GWAS hits that did not satisfy the study-wide significance have often been reported to be low.[39]

The association between the *ABO* gene locus and *agaE* and *agaS* again emphasizes the strength of metagenome shotgun sequencing, given that gene-ortholog- and pathway-level information could not be obtained from 16S rRNA sequencing. The abundances of *agaE* and *agaS* tended to be high in secretors with blood group A, suggesting that the abundance of N-acetylgalactosamine in the gut could positively affect the abundance of bacteria with the capacity to utilize N-acetylgalactosamine. This candidate molecular mechanism was also suggested in a multi-omics study in pigs.[32] The ABO blood group has associations with various diseases such as infectious diseases, cancers, and metabolic diseases,[40] while the biological mechanisms of such associations are still not understood. Given that the gut microbiota can interact with the host through mechanisms such as metabolite production and stimulation of the immune system, ABO-blood-type-associated differences of specific gut microbes may influence disease risks. For example, a relatively high prevalence of myocardial infarction in blood-type-A individuals[41] and an increased abundance of the bacteria with *ageE* and *agaS*, such as *Collinsella* and *Escherichia*, in the gut of individuals with atherosclerotic cardiovascular diseases[42,43] may suggest a potential link between blood type, gut microbiome, and diseases. Given that there is a global heterogeneity of the gut microbiome, the frequency of ABO blood groups, and ABO blood group-disease associations[41] among

and effect sizes in the linear regression, respectively. In the linear regression analysis, target and reference blood groups are treated as 1 and 0, respectively. p values were calculated with the fixed-effect meta-analysis. *$p < 0.05$; **$p < 0.05/6$.

(D) A schematic illustration of the relationship between ABO blood group, secretor status, and oligosaccharides in the gut.

(E and F) Violin and boxplots represent the normalized abundance of *agaE* (E) and *agaS* (F) stratified by ABO blood groups and secretor status. Boxplots indicate the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). The overlayed dots represent individual observations used for the association analysis. The color of the dots represents the dataset. The median values in those who are blood group A and secretor are indicated as horizontal dashed lines colored red. AF, allele frequency; EAS, East Asian; EUR, European; IQR, interquartile ranges; sec, secretor.

See also Figures S4 and S5 and Table S7.

CellPress
OPEN ACCESS

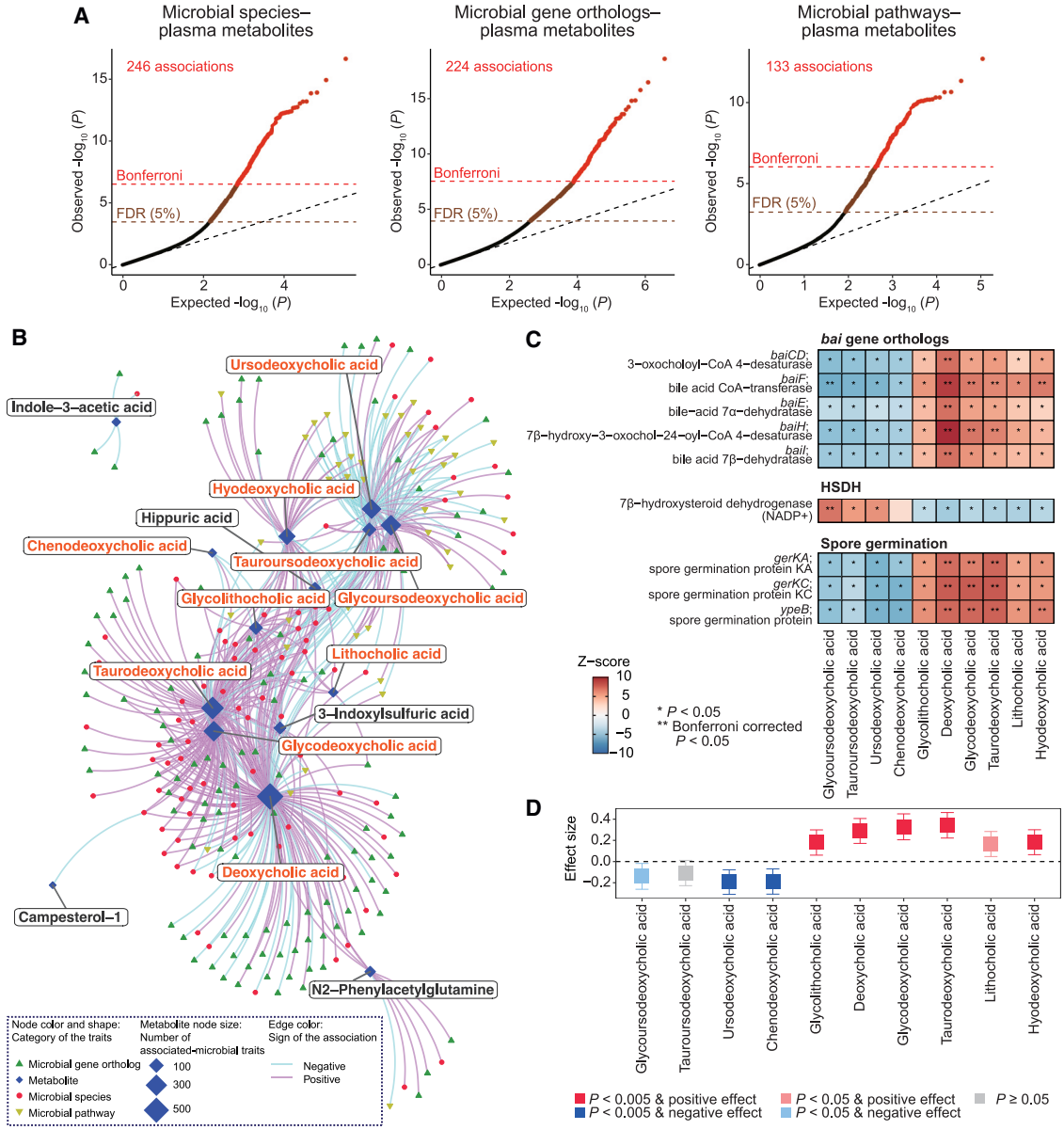


Figure 3. Gut microbiome-plasma metabolome association analysis

(A) Q-Q plots of the p values from the microbial traits-metabolites association analysis (left, bacterial species; middle, microbial gene orthologs; right, microbial pathway). The x axis indicates expected log-transformed p values, and the y axis indicates log-transformed observed p values. The diagonal dashed line represents y = x, which corresponds to the null hypothesis. The horizontal red dashed line indicates the Bonferroni-corrected threshold ($\alpha = 0.05$), and the brown dashed line indicates the FDR threshold (FDR = 0.05) calculated with the Benjamini-Hochberg method. The microbial traits-metabolites pairs with FDR <0.05 are plotted as brown dots, and the other microbial traits-metabolites pairs are plotted as black dots. p values were calculated with the linear regression analysis (Wald's test).

(B) A network plot for the significant microbial traits-metabolites associations under the Bonferroni-corrected threshold ($\alpha = 0.05$). The shape and color of the nodes represent the category of the traits. As for the metabolites, the sizes of the nodes represent the number of significantly associated microbial traits. The color of the edges represents the signs of the associations. The names of the bile acids are indicated in orange, while other metabolite names are indicated in black.

(C) Heatmaps represent the association between microbial gene orthologs and plasma bile acids. The colors of the tiles indicate the *Z* scores in the association tests. *p < 0.05; **p satisfied the Bonferroni-corrected threshold in (A). p values were calculated with the linear regression analysis (Wald's test).

(D) A forest plot indicates the association between α-diversity of the gut microbiome and plasma bile acids. The boxes and error bars indicate the effect sizes and 95% confidence interval in the linear regression analysis, respectively. The colors of the boxes indicate the sign of the effects and whether the test satisfied the nominal or multiple-test-corrected significance thresholds. p values were calculated with the linear regression analysis (Wald's test). FDR, false discovery ratio; HSDH, hydroxysteroid dehydrogenase; KO, KEGG Orthology.

See also Tables S8, S9, and S10.

the populations, analysis with underrepresented populations should be continued.

In the microbiome-metabolome associations, we identified associations between the microbial gene orthologs related to bile acid metabolism and plasma bile acid abundances possibly because microbial gene orthologs related to bile acid could affect both bile acid abundances and the fitness of the bacteria. We also identified the association of α-diversity with bile acids, such as a positive association with deoxycholic acid. Deoxycholic acid is a major secondary bile acid in the human gut and a cytotoxic detergent for some kinds of bacteria.[20] A possible explanation for this association was that moderate selection pressure by the cytotoxic activity of the deoxycholic acid was necessary to keep the diversity of the gut microbiome or that the diversity of the gut microbiome was linked to the production of the secondary bile acid by the bacteria. Decreases in α-diversity have been reported for various disease conditions like inflammatory bowel diseases (IBDs).[1,44] Given that coupled changes of the α-diversity and deoxycholic acid were reported in IBD,[44,45] the link between α-diversity and deoxycholic acid might contribute to the etiology of IBDs.

In summary, our metagenome shotgun-sequencing-based multi-omics study with the Japanese dataset identified EAS-specific or functionally interpretable gut microbiome-host factor associations that underscored the importance of analyzing currently underrepresented populations with a metagenome shotgun-sequencing-based approach.

Limitations of the study

Our sample size was not that large compared with previous studies for well-studied populations such as EUR and Chinese, which could be a limitation of this study. Although we confirmed the consistency of the association between genetic variants and gut microbiome in two datasets, further validation with another cohort would be warranted. Given that there is substantial heterogeneity of the gut microbiome and host factors, even within EAS populations,[35,46] it will be necessary to continue to build large datasets for underrepresented populations and perform comparative and meta-analyses to evaluate the reproducibility and heterogeneity of gut microbiome-host factor associations across diverse populations. We consider that our results, which are publicly available, would contribute to expanding the diversity of current study populations and be a useful resource for future studies. Although not investigated in this study, considering that bacteriophages are also involved in microbiome-host interactions,[47] it would be necessary to include them in analyses in the future, despite our research being primarily focused on bacteria.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Subject participation
- METHOD DETAILS
  - Metagenome shotgun sequencing
  - Quality control of metagenome shotgun sequencing reads
  - Taxonomic annotation of metagenome and abundance quantification
  - Functional annotation and abundance calculation
  - QC and normalization of the bacterial and gene ortholog abundance data
  - Quantification of microbial pathways based on gene set variance analysis
  - Calculation of α-diversity of the metagenome
  - Plasma metabolome profiling based on the CE-TOFMS and LC-TOFMS
  - Genotyping of the samples based on the SNP array
  - Genotyping of the samples based on the whole genome sequencing
  - GWAS for the microbial traits
  - Association between the ABO blood group and microbial traits
  - GWAS for the plasma metabolites, PheWAS, and co-localization analysis
  - Microbiome–metabolome association analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113324.

ACKNOWLEDGMENTS

We would like to thank all the participants involved in this study. This research was supported by the Japan Society for the Promotion of Science KAKENHI (22H00476); the Japan Agency for Medical Research and Development (AMED; JP21gm4010006, JP22km0405211, JP22ek0410075, JP22km0405217, JP22ek0109594, JP223fa627002, JP223fa627010, and JP233fa627011); JST grant number JPMJPF2101; JST Moonshot R&D (JPMJMS2021 and JPMJMS2024); Takeda Science Foundation; the Bioinformatics Initiative of Osaka University Graduate School of Medicine; and the Institute for Open and Transdisciplinary Research Initiatives and Center for Infectious Disease Education and Research (CiDER), Osaka University.

AUTHOR CONTRIBUTIONS

Y.T., T.K., and Y.O. designed the study. Y.T., T.K., K.S., and Y.O. conducted the data analysis. Y.T. and Y.O. wrote the manuscript. Y.T., T.K., Y. Maeda, T.N., E.O.-I., D.M., Y. Matsumoto, and S.N. conducted the experiments. Y.T., T.K., K.S., Y. Maeda, K.O., S.K., E.O.-I., T.O., T.N., M.K., M.T., K.Y., N.A., M.Y.-S., A.H., H. Matsuoka, M.Y., and S.O. collected and managed the samples. M.F., H.I., H.K., H. Mochizuki, K.T., A.K., and Y.O. supervised the study. All authors contributed to the article and approved the submitted version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 5, 2023
Revised: September 11, 2023
Accepted: October 6, 2023
Published: November 6, 2023

REFERENCES

1. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature *569*, 655–662. https://doi.org/10.1038/s41586-019-1237-9.
2. Kishikawa, T., Maeda, Y., Nii, T., Motooka, D., Matsumoto, Y., Matsushita, M., Matsuoka, H., Yoshimura, M., Kawada, S., Teshigawara, S., et al. (2020). Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population. Ann. Rheum. Dis. *79*, 103–111. https://doi.org/10.1136/annrheumdis-2019-215743.
3. Kishikawa, T., Ogawa, K., Motooka, D., Hosokawa, A., Kinoshita, M., Suzuki, K., Yamamoto, K., Masuda, T., Matsumoto, Y., Nii, T., et al. (2020). A Metagenome-Wide Association Study of Gut Microbiome in Patients With Multiple Sclerosis Revealed Novel Disease Pathology. Front. Cell. Infect. Microbiol. *10*, 585973. https://doi.org/10.3389/fcimb.2020.585973.
4. Tomofuji, Y., Maeda, Y., Oguro-Igashira, E., Kishikawa, T., Yamamoto, K., Sonehara, K., Motooka, D., Matsumoto, Y., Matsuoka, H., Yoshimura, M., et al. (2021). Metagenome-wide association study revealed disease-specific landscape of the gut microbiome of systemic lupus erythematosus in Japanese. Ann. Rheum. Dis. *80*, 1575–1583. https://doi.org/10.1136/annrheumdis-2021-220687.
5. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60. https://doi.org/10.1038/nature11450.
6. Holmes, E., Li, J.V., Marchesi, J.R., and Nicholson, J.K. (2012). Gut Microbiota Composition and Activity in Relation to Host Metabolic Phenotype and Disease Risk. Cell Metabol. *16*, 559–564. https://doi.org/10.1016/j.cmet.2012.10.007.
7. Goodrich, J.K., Davenport, E.R., Beaumont, M., Jackson, M.A., Knight, R., Ober, C., Spector, T.D., Bell, J.T., Clark, A.G., and Ley, R.E. (2016). Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe *19*, 731–743. https://doi.org/10.1016/j.chom.2016.04.017.
8. Scepanovic, P., Hodel, F., Mondot, S., Partula, V., Byrd, A., Hammer, C., Alanio, C., Bergstedt, J., Patin, E., Touvier, M., et al. (2019). A comprehensive assessment of demographic, environmental, and host genetic associations with gut microbiome diversity in healthy individuals. Microbiome *7*, 130. https://doi.org/10.1186/s40168-019-0747-x.
9. Kurilshikov, A., Medina-Gomez, C., Bacigalupe, R., Radjabzadeh, D., Wang, J., Demirkan, A., Le Roy, C.I., Raygoza Garay, J.A., Finnicum, C.T., Liu, X., et al. (2021). Large-scale association analyses identify host factors influencing human gut microbiome composition. Nat. Genet. *53*, 156–165. https://doi.org/10.1038/s41588-020-00763-1.
10. Rühlemann, M.C., Hermes, B.M., Bang, C., Doms, S., Moitinho-Silva, L., Thingholm, L.B., Frost, F., Degenhardt, F., Wittig, M., Kässens, J., et al. (2021). Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. Nat. Genet. *53*, 147–155. https://doi.org/10.1038/s41588-020-00747-1.
11. Liu, X., Tong, X., Zou, Y., Lin, X., Zhao, H., Tian, L., Jie, Z., Wang, Q., Zhang, Z., Lu, H., et al. (2022). Mendelian randomization analyses support causal relationships between blood metabolites and the gut microbiome. Nat. Genet. *54*, 52–61. https://doi.org/10.1038/s41588-021-00968-y.
12. Qin, Y., Havulinna, A.S., Liu, Y., Jousilahti, P., Ritchie, S.C., Tokolyi, A., Sanders, J.G., Valsta, L., Brożyńska, M., Zhu, Q., et al. (2022). Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. Nat. Genet. *54*, 134–142. https://doi.org/10.1038/s41588-021-00991-z.
13. Lopera-Maya, E.A., Kurilshikov, A., van der Graaf, A., Hu, S., Andreu-Sánchez, S., Chen, L., Vila, A.V., Gacesa, R., Sinha, T., Collij, V., et al. (2022). Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. Nat. Genet. *54*, 143–151. https://doi.org/10.1038/s41588-021-00992-y.
14. Blekhman, R., Goodrich, J.K., Huang, K., Sun, Q., Bukowski, R., Bell, J.T., Spector, T.D., Keinan, A., Ley, R.E., Gevers, D., and Clark, A.G. (2015). Host genetic variation impacts microbiome composition across human body sites. Genome Biol. *16*, 191. https://doi.org/10.1186/s13059-015-0759-1.
15. Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. Nature *555*, 210–215. https://doi.org/10.1038/nature25973.
16. Hughes, D.A., Bacigalupe, R., Wang, J., Rühlemann, M.C., Tito, R.Y., Falony, G., Joossens, M., Vieira-Silva, S., Henckaerts, L., Rymenans, L., et al. (2020). Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. Nat. Microbiol. *5*, 1079–1087. https://doi.org/10.1038/s41564-020-0743-8.
17. Bonder, M.J., Kurilshikov, A., Tigchelaar, E.F., Mujagic, Z., Imhann, F., Vila, A.V., Deelen, P., Vatanen, T., Schirmer, M., Smeekens, S.P., et al. (2016). The effect of host genetics on the gut microbiome. Nat. Genet. *48*, 1407–1412. https://doi.org/10.1038/ng.3663.
18. Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., Wu, X., Li, J., Tang, L., Li, Y., et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat. Med. *21*, 895–905. https://doi.org/10.1038/nm.3914.
19. Ishida, S., Kato, K., Tanaka, M., Odamaki, T., Kubo, R., Mitsuyama, E., Xiao, J.Z., Yamaguchi, R., Uematsu, S., Imoto, S., and Miyano, S. (2020). Genome-wide association studies and heritability analysis reveal the involvement of host genetics in the Japanese gut microbiota. Commun. Biol. *3*, 686. https://doi.org/10.1038/s42003-020-01416-z.
20. Cai, J., Sun, L., and Gonzalez, F.J. (2022). Gut microbiota-derived bile acids in intestinal immunity, inflammation, and tumorigenesis. Cell Host Microbe *30*, 289–300. https://doi.org/10.1016/j.chom.2022.02.004.
21. Wilmanski, T., Rappaport, N., Earls, J.C., Magis, A.T., Manor, O., Lovejoy, J., Omenn, G.S., Hood, L., Gibbons, S.M., and Price, N.D. (2019). Blood metabolome predicts gut microbiome α-diversity in humans. Nat. Biotechnol. *37*, 1217–1228. https://doi.org/10.1038/s41587-019-0233-9.
22. Visconti, A., Le Roy, C.I., Rosa, F., Rossi, N., Martin, T.C., Mohney, R.P., Li, W., de Rinaldis, E., Bell, J.T., Venter, J.C., et al. (2019). Interplay between the human gut microbiome and host metabolism. Nat. Commun. *10*, 4505. https://doi.org/10.1038/s41467-019-12476-z.
23. Vojinovic, D., Radjabzadeh, D., Kurilshikov, A., Amin, N., Wijmenga, C., Franke, L., Ikram, M.A., Uitterlinden, A.G., Zhernakova, A., Fu, J., et al. (2019). Relationship between gut microbiota and circulating metabolites in population-based cohorts. Nat. Commun. *10*, 5813. https://doi.org/10.1038/s41467-019-13721-1.
24. Dekkers, K.F., Sayols-Baixeras, S., Baldanzi, G., Nowak, C., Hammar, U., Nguyen, D., Varotsis, G., Brunkwall, L., Nielsen, N., Eklund, A.C., et al. (2022). An online atlas of human plasma metabolite signatures of gut microbiome composition. Nat. Commun. *13*, 5370. https://doi.org/10.1038/s41467-022-33050-0.
25. Tomofuji, Y., Sonehara, K., Kishikawa, T., Maeda, Y., Ogawa, K., Kawabata, S., Nii, T., Okuno, T., Oguro-Igashira, E., Kinoshita, M., et al. (2023). Reconstruction of the personal information from human genome reads in gut metagenome sequencing data. Nat. Microbiol. *8*, 1079–1094. https://doi.org/10.1038/s41564-023-01381-3.
26. Cygnar, K.D., and Zhao, H. (2009). Phosphodiesterase 1C is dispensable for rapid response termination of olfactory sensory neurons. Nat. Neurosci. *12*, 454–462. https://doi.org/10.1038/nn.2289.
27. Saunders, G.R.B., Wang, X., Chen, F., Jang, S.-K., Liu, M., Wang, C., Gao, S., Jiang, Y., Khunsriraksakul, C., Otto, J.M., et al. (2022). Genetic diversity fuels gene discovery for tobacco and alcohol use. Nature *612*, 720–724. https://doi.org/10.1038/s41586-022-05477-4.

28. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am. J. Hum. Genet. *104*, 65–75. https://doi.org/10.1016/j.ajhg.2018.11.008.

29. Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S.M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T., et al. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. Nat. Genet. *54*, 437–449. https://doi.org/10.1038/s41588-022-01016-z.

30. Nakao, M., Matsuo, K., Hosono, S., Ogata, S., Ito, H., Watanabe, M., Mizuno, N., Iida, S., Sato, S., Yatabe, Y., et al. (2011). ABO blood group alleles and the risk of pancreatic cancer in a Japanese population. Cancer Sci. *102*, 1076–1080. https://doi.org/10.1111/j.1349-7006.2011.01907.x.

31. Masuda, M., Okuda, K., Ikeda, D.D., Hishigaki, H., and Fujiwara, T. (2015). Interaction of genetic markers associated with serum alkaline phosphatase levels in the Japanese population. Hum. Genome Var. *2*, 15019. https://doi.org/10.1038/hgv.2015.19.

32. Yang, H., Wu, J., Huang, X., Zhou, Y., Zhang, Y., Liu, M., Liu, Q., Ke, S., He, M., Fu, H., et al. (2022). ABO genotype alters the gut microbiota by regulating GalNAc levels in pigs. Nature *606*, 358–367. https://doi.org/10.1038/s41586-022-04769-z.

33. Kudo, T., Iwasaki, H., Nishihara, S., Shinya, N., Ando, T., Narimatsu, I., and Narimatsu, H. (1996). Molecular Genetic Analysis of the Human Lewis Histo-blood Group System: II. SECRETOR GENE INACTIVATION BY A NOVEL SINGLE MISSENSE MUTATION A385T IN JAPANESE NONSECRETOR INDIVIDUALS (*). J. Biol. Chem. *271*, 9830–9837. https://doi.org/10.1074/jbc.271.16.9830.

34. Koda, Y., Soejima, M., Liu, Y., and Kimura, H. (1996). Molecular basis for secretor type alpha(1,2)-fucosyltransferase gene deficiency in a Japanese population: a fusion gene generated by unequal crossover responsible for the enzyme deficiency. Am. J. Hum. Genet. *59*, 343–350.

35. Tomofuji, Y., Kishikawa, T., Maeda, Y., Ogawa, K., Otake-Kasamoto, Y., Kawabata, S., Nii, T., Okuno, T., Oguro-Igashira, E., Kinoshita, M., et al. (2022). Prokaryotic and viral genomes recovered from 787 Japanese gut metagenomes revealed microbial features linked to diets, populations, and diseases. Cell Genom. *2*, 100219. https://doi.org/10.1016/j.xgen.2022.100219.

36. Kishikawa, T., Maeda, Y., Nii, T., Arase, N., Hirata, J., Suzuki, K., Yamamoto, K., Masuda, T., Ogawa, K., Tsuji, S., et al. (2021). Increased levels of plasma nucleotides in patients with rheumatoid arthritis. Int. Immunol. *33*, 119–124. https://doi.org/10.1093/intimm/dxaa059.

37. Sorg, J.A., and Sonenshein, A.L. (2010). Inhibiting the Initiation of Clostridium difficile Spore Germination using Analogs of Chenodeoxycholic Acid, a Bile Acid. J. Bacteriol. *192*, 4983–4990. https://doi.org/10.1128/JB.00610-10.

38. Sorg, J.A., and Sonenshein, A.L. (2008). Bile Salts and Glycine as Cogerminants for Clostridium difficile Spores. J. Bacteriol. *190*, 2505–2512. https://doi.org/10.1128/JB.01765-07.

39. Sanna, S., Kurilshikov, A., van der Graaf, A., Fu, J., and Zhernakova, A. (2022). Challenges and future directions for studying effects of host genetics on the gut microbiome. Nat. Genet. *54*, 100–106. https://doi.org/10.1038/s41588-021-00983-z.

40. Abegaz, S.B. (2021). Human ABO Blood Groups and Their Associations with Different Diseases. BioMed Res. Int. *2021*, 6629060. https://doi.org/10.1155/2021/6629060.

41. Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. Nat. Genet. *53*, 1415–1424. https://doi.org/10.1038/s41588-021-00931-x.

42. Karlsson, F.H., Fåk, F., Nookaew, I., Tremaroli, V., Fagerberg, B., Petranovic, D., Bäckhed, F., and Nielsen, J. (2012). Symptomatic atherosclerosis is associated with an altered gut metagenome. Nat. Commun. *3*, 1245. https://doi.org/10.1038/ncomms2266.

43. Jie, Z., Xia, H., Zhong, S.-L., Feng, Q., Li, S., Liang, S., Zhong, H., Liu, Z., Gao, Y., Zhao, H., et al. (2017). The gut microbiome in atherosclerotic cardiovascular disease. Nat. Commun. *8*, 845. https://doi.org/10.1038/s41467-017-00900-1.

44. Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., Vatanen, T., Hall, A.B., Mallick, H., McIver, L.J., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat. Microbiol. *4*, 293–305. https://doi.org/10.1038/s41564-018-0306-4.

45. Sinha, S.R., Haileselassie, Y., Nguyen, L.P., Tropini, C., Wang, M., Becker, L.S., Sim, D., Jarr, K., Spear, E.T., Singh, G., et al. (2020). Dysbiosis-Induced Secondary Bile Acid Deficiency Promotes Intestinal Inflammation. Cell Host Microbe *27*, 659–670.e5. https://doi.org/10.1016/j.chom.2020.01.021.

46. Nishijima, S., Suda, W., Oshima, K., Kim, S.-W., Hirose, Y., Morita, H., and Hattori, M. (2016). The gut microbiome of healthy Japanese and its microbial and functional uniqueness. DNA Res. *23*, 125–133. https://doi.org/10.1093/dnares/dsw002.

47. Shkoporov, A.N., and Hill, C. (2019). Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. Cell Host Microbe *25*, 195–209. https://doi.org/10.1016/j.chom.2019.01.017.

48. Tomofuji, Y., Kishikawa, T., Maeda, Y., Ogawa, K., Nii, T., Okuno, T., Oguro-Igashira, E., Kinoshita, M., Yamamoto, K., Sonehara, K., et al. (2022). Whole gut virome analysis of 476 Japanese revealed a link between phage and autoimmune disease. Ann. Rheum. Dis. *81*, 278–288. https://doi.org/10.1136/annrheumdis-2021-221267.

49. Kishikawa, T., Arase, N., Tsuji, S., Maeda, Y., Nii, T., Hirata, J., Suzuki, K., Yamamoto, K., Masuda, T., Ogawa, K., et al. (2021). Large-scale plasma-metabolome analysis identifies potential biomarkers of psoriasis and its clinical subtypes. J. Dermatol. Sci. *102*, 78–84. https://doi.org/10.1016/j.jdermsci.2021.03.006.

50. Sonehara, K., Sakaue, S., Maeda, Y., Hirata, J., Kishikawa, T., Yamamoto, K., Matsuoka, H., Yoshimura, M., Nii, T., Ohshima, S., et al. (2022). Genetic architecture of microRNA expression and its link to complex diseases in the Japanese population. Hum. Mol. Genet. *31*, 1806–1820, ddab361.. https://doi.org/10.1093/hmg/ddab361

51. Maeda, Y., Kurakawa, T., Umemoto, E., Motooka, D., Ito, Y., Gotoh, K., Hirota, K., Matsushita, M., Furuta, Y., Narazaki, M., et al. (2016). Dysbiosis Contributes to Arthritis Development via Activation of Autoreactive T Cells in the Intestine: DYSBIOSIS CONTRIBUTES TO ARTHRITIS DEVELOPMENT. Arthritis Rheumatol. *68*, 2646–2661. https://doi.org/10.1002/art.39783.

52. Kishikawa, T., Tomofuji, Y., Inohara, H., and Okada, Y. (2022). OMARU: a robust and multifaceted pipeline for metagenome-wide association study. NAR Genom. Bioinform. *4*, lqac019. https://doi.org/10.1093/nargab/lqac019.

53. Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864. https://doi.org/10.1093/bioinformatics/btr026.

54. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

55. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

56. Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. *37*, 179–185. https://doi.org/10.1038/s41587-018-0008-8.

57. Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D., Dunn, M., Mkandawire, T.T., Zhu, A., Shao, Y., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat. Biotechnol. *37*, 186–192. https://doi.org/10.1038/s41587-018-0009-7.

58. Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504. https://doi.org/10.1038/s41586-019-0965-1.

59. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

60. Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. *38*, e132. https://doi.org/10.1093/nar/gkq275.

61. Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

62. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

63. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507. https://doi.org/10.1038/nprot.2011.457.

64. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinf. *14*, 7. https://doi.org/10.1186/1471-2105-14-7.

65. Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J.Z., Abe, F., and Osawa, R. (2016). Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. BMC Microbiol. *16*, 90. https://doi.org/10.1186/s12866-016-0708-5.

66. Sugimoto, M., Wong, D.T., Hirayama, A., Soga, T., and Tomita, M. (2010). Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. Metabolomics *6*, 78–95. https://doi.org/10.1007/s11306-009-0178-y.

67. Baran, R., Kochi, H., Saito, N., Suematsu, M., Soga, T., Nishioka, T., Robert, M., and Tomita, M. (2006). MathDAMP: a package for differential analysis of metabolite profiles. BMC Bioinf. *7*, 530. https://doi.org/10.1186/1471-2105-7-530.

68. Wallace, W.E., Kearsley, A.J., and Guttman, C.M. (2004). An Operator-Independent Approach to Mass Spectral Peak Identification and Integration. Anal. Chem. *76*, 2446–2452. https://doi.org/10.1021/ac0354701.

69. Reijeng, J.C., Martens, J.H.P.A., Giuliani, A., and Chiari, M. (2002). Pherogram normalization in capillary electrophoresis and micellar electrokinetic chromatography analyses in cases of sample matrix-induced migration time shifts. J. Chromatogr. B *770*, 45–51. https://doi.org/10.1016/S0378-4347(01)00527-8.

70. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

71. Sakaue, S., Yamaguchi, E., Inoue, Y., Takahashi, M., Hirata, J., Suzuki, K., Ito, S., Arai, T., Hirose, M., Tanino, Y., et al. (2021). Genetic determinants of risk in autoimmune pulmonary alveolar proteinosis. Nat. Commun. *12*, 1032. https://doi.org/10.1038/s41467-021-21011-y.

72. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909. https://doi.org/10.1038/ng1847.

73. Okada, Y., Momozawa, Y., Sakaue, S., Kanai, M., Ishigaki, K., Akiyama, M., Kishikawa, T., Arai, Y., Sasaki, T., Kosaki, K., et al. (2018). Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. Nat. Commun. *9*, 1631. https://doi.org/10.1038/s41467-018-03274-0.

74. Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat. Commun. *10*, 4393. https://doi.org/10.1038/s41467-019-12276-5.

75. Tadaka, S., Katsuoka, F., Ueki, M., Kojima, K., Makino, S., Saito, S., Otsuki, A., Gocho, C., Sakurai-Yageta, M., Danjoh, I., et al. (2019). 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. Hum. Genome Var. *6*, 28–29. https://doi.org/10.1038/s41439-019-0059-5.

76. Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat. Commun. *10*, 5436. https://doi.org/10.1038/s41467-019-13225-y.

77. Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2015). minimac2: faster genotype imputation. Bioinformatics *31*, 782–784. https://doi.org/10.1093/bioinformatics/btu704.

78. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. *103*, 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.

79. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

80. Han, B., and Eskin, E. (2011). Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. Am. J. Hum. Genet. *88*, 586–598. https://doi.org/10.1016/j.ajhg.2011.04.014.

81. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336–2337. https://doi.org/10.1093/bioinformatics/btq419.

82. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164. https://doi.org/10.1093/nar/gkq603.

83. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

84. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

85. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

86. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

87. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. *10*, e1004383. https://doi.org/10.1371/journal.pgen.1004383.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human DNA extracted from blood	This study	N/A
Critical commercial assays		
KAPA Hyper Prep Kit	illumina	Cat#KK8504
Glass beads (diameter 0.1 mm)	biospec	Cat#11079101
HMT's metabolome analysis services (Dual Scan)	Human Metabolome Technologies Inc.	https://humanmetabolome.com/ap/service/dualscan/
Deposited data		
Metagenome shotgun sequencing data	Tomofuji et al.[35]	National Bioscience Database Center (NBDC) Human Database #hum0197
Metagenome shotgun sequencing data	Kishikawa et al.[2]	National Bioscience Database Center (NBDC) Human Database #hum0197
Metagenome shotgun sequencing data	Kishikawa et al.[3]	National Bioscience Database Center (NBDC) Human Database #hum0197
Metagenome shotgun sequencing data	Tomofuji et al.[4]	National Bioscience Database Center (NBDC) Human Database #hum0197
Metagenome shotgun sequencing data	Tomofuji et al.[48]	National Bioscience Database Center (NBDC) Human Database #hum0197
JMAG	Tomofuji et al.[35]	National Bioscience Database Center (NBDC) Human Database #hum0197
Software and algorithms		
ANNOVAR	Wang et al.[82]	https://annovar.openbioinformatics.org/en/latest/
bcl2fastq	illumina	https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software/downloads.html
BMTagger	ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/	ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/
bowtie2	Langmead and Salzberg[55]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BWA-MEM	https://github.com/lh3/bwa	https://github.com/lh3/bwa
CheckM	Parks et al.[84]	https://github.com/Ecogenomics/CheckM
coloc	Giambartolomei et al.[84]	https://chr1swallace.github.io/coloc/articles/a01_intro.html
DIAMOND	Buchfink et al.[62]	https://github.com/bbuchfink/diamond
EIGENSTRAT	Price et al.[72]	https://www.hsph.harvard.edu/alkes-price/software/
GATK	Broad institute	https://gatk.broadinstitute.org/hc/en-us
ggraph	https://github.com/thomasp85/ggraph	https://github.com/thomasp85/ggraph
GSVA	Hänzelmann et al.[64]	https://github.com/rcastelo/GSVA
MEGAHIT	Li et al.[59]	https://github.com/voutcn/megahit
metafor	https://wviechtb.github.io/metafor/	https://wviechtb.github.io/metafor/
METASOFT	Han et al.[80]	http://genetics.cs.ucla.edu/meta
Minimac4	Fuchsberger et al.[77]	https://github.com/statgen/Minimac4
MMseqs2	Steinegger & Söding[85]	https://github.com/soedinglab/MMseqs2
ncbi-blast-plus	Camacho et al.[86]	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE = BlastDocs&DOC_TYPE = Download

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PEER	Stegle et al.[63]	https://github.com/PMBio/peer
PLINK	Purcell et al.[70]	https://www.cog-genomics.org/plink/
PLINK2	Chang et al.[79]	https://www.cog-genomics.org/plink/
PRINSEQ	Schmieder and Edwards[53]	http://prinseq.sourceforge.net/
Python	Python Software Foundation	https://www.python.org/downloads/release/python-376/
R	The R Foundation for Statistical Computing	https://www.r-project.org
SHAPEIT4	Delaneau et al.[76]	https://github.com/odelaneau/shapeit4
vegan	http://CRAN.R-project.org/package = vegan	http://CRAN.R-project.org/package = vegan

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Yukinori Okada (yokada@sg.med.osaka-u.ac.jp).

Materials availability

The materials that support the findings of this study are available from the corresponding authors upon reasonable request. Please contact the Lead Contact for additional information.

Data and code availability

- The metagenome shotgun sequencing data used in this study are under controlled access in the Japanese Genotype-Phenotype Archive (JGA) with accession numbers JGAS000205, JGAS000260, JGAS000316, JGAS000531, and JGAS000415 to protect the participants' privacy.[25] Researchers who comply with NBDC's data terms of use can apply for access to the data. The results of the association analysis are publicly available in NBDC Human Database (http://humandbs.biosciencedbc.jp/) with the accession number of hum0197.
- This paper does not report original code. We used publicly available software in this study. Please see the METHOD DETAILS section for further details.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subject participation

625 Japanese participants recruited in the previous studies were included in this study.[2–4,25,35,36,48–50] Among these participants, 524, 362, 524, and 362 participants had included in the analysis with metagenome shotgun sequencing, plasma metabolome analysis, SNP array-based genotyping, and WGS data, respectively. All participants provided written informed consent before participation. The study protocol was approved by the ethics committees of Osaka University and related medical institutions.

METHOD DETAILS

Metagenome shotgun sequencing

Phenol-chloroform DNA extraction and subsequent metagenome shotgun sequencing were performed in the previous studies.[2–4,35,48] Briefly, for the sequencing batches 2–4, 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 h after collection. After washing with 1 mL of PBS (−), 200 μL of the homogenates were used for further DNA extraction.

For the sequencing batches 1 and 5, fecal samples had been stored at −80°C within 6 h after production or immediately frozen after production in an insulated container for storage at −20°C and subsequently stored at −80°C within 24 h after production. For these samples stored without RNAlater, RNAlater (Ambion) was added to make 10-fold dilutions of homogenates before the DNA extraction. After washing with 1 mL of PBS (−), 200 μL of the homogenates were used for further DNA extraction.

DNA was extracted according to a previously described method.[51] Briefly, 300 μL of sodium dodecyl sulfate–Tris solution, 0.3 g glass beads (diameter 0.1 mm) (BioSpec), and 500 μL EDTA-Tris-saturated phenol were added to the suspension, and the mixture

was vortexed vigorously using a FastPrep-24 (MP Biomedicals) at 5.0 power level for 30 s. After centrifugation at 20,000 g for 5 min 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.

The amount of dsDNA was quantified with Qubit Fluorometer (Thermo Fisher Scientific). After the sonication with the ME220 (Covaris), a shotgun sequencing library was constructed using the KAPA Hyper Prep Kit (KAPA Biosystems) following the manufacturer's instructions. The library quality was evaluated with LabChip GX Touch (PerkinElmer). The amount of the library was quantified with the Qubit Fluorometer and KAPA Library Quantification Kits (KAPA Biosystems). 150-bp paired-end reads were generated on HiSeq 3000 or NovaSeq 6000. Samples in each dataset were barcoded, pooled, and sequenced simultaneously in a single run without library replication. All the sequencing was performed in the Department of Infection Metagenomics/Next-Generation Sequencing Core Facility, Research Institute for Microbial Diseases, Osaka University (Suita, Japan). The sequence reads were converted to the FASTQ format using bcl2fastq (version 2.19). Further run information is described in Table S12.

The five sequencing runs were grouped into two datasets based on the sequencing period, sequencer, and other modality of data (Figure S1). Dataset 1 was a gut microbiome dataset with plasma metabolome and genotype information and it was sequenced with Hiseq3000 in 2019. On the other hand, Dataset 2 was a gut microbiome dataset with only genotype information, which was sequenced with Novaseq6000 between 2020 and 2021.

Quality control of metagenome shotgun sequencing reads

We performed a series of QC steps to maximize the quality of the datasets as previously described.[2–4,52] The main steps in the QC process were: (i) trimming of low-quality bases, (ii) removal of duplicated reads, and (iii) identification and masking of human reads. We marked duplicate reads using PRINSEQ-lite[53] (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[54] (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 the same sequences. As a final QC step, we aligned the quality-filtered reads to the human reference genome (hg38) using bowtie2[55] (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

We performed taxonomic annotation of metagenome and abundance quantification as previously described. We used curated reference microbial genomes as previously described.[2] The reference microbial genomes of the Japanese population constructed by Nishijima et al.[46] were combined with the genomes identified from the cultivated or uncultivated human gut bacteria projects.[56–58] 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 the "-R 3" option. 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 species-level clades was calculated. Then, we detected and removed 10 and 7 outlier samples by principal component analysis (PCA; PC 1–6) based on the species-level abundance separately for datasets 1 and 2.

Functional annotation and abundance calculation

We performed functional annotation of metagenome and abundance quantification as previously described.[2–4,52] *De novo* assembly of the filtered paired-end reads into contigs was conducted using MEGAHIT[59] (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[60] (version 3.38; parameters: -a -k -f G). Next, we annotated the ORF catalog with the Kyoto Encyclopedia of Genes and Genomes (KEGG) protein database (https://www.kegg.jp).[61] 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 the KEGG protein database with DIAMOND[62] using BLASTP (version v0.9.32.133; parameters: -f 6 -b 15.0 -k 1 -e 1e-6 –subject-cover 50). 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.

QC and normalization of the bacterial and gene ortholog abundance data

After sample QC, we performed the QC and normalization of the bacterial abundance data. We removed clades detected (i) in less than 50% of the samples, (ii) in no sample in any of the sequencing batches, or (iii) with an average relative abundance of less than 0.001% of the total abundance. The thresholds for the detection ratio and average relative abundance were set based on the recent gut microbiome GWAS.[11,13] After selection, 450 and 453 species-level clades were retained for datasets 1 and 2, respectively. Then, bacterial abundances were log-transformed and outliers (outside the range of mean±5s.d.) were removed per clade. We added the pseudo-counts (half of the minimum non-zero value) to the zeros before the log transformation. Then, log-transformed bacterial

abundances were corrected for the covariates using PEER[63] (version 1.3) accounting for 30 unobserved confounders as well as the known confounders, such as age, age,[2] sex, sequencing batches, facility, diseases, and three genotype-based principal components. By utilizing PEER, we removed unwanted variations of the data. The residuals were standardized with the inverse normal transformation.

We also performed the normalization of the microbial gene ortholog abundance data. We summed up the gene abundance data for each KEGG Orthology (KO) and removed gene orthologs detected (i) in less than 50% of the samples or (ii) in no sample in any of the sequencing batches. After selection, 4,760 and 4,829 microbial gene orthologs were retained for datasets 1 and 2, respectively. Then, log transformation, covariate adjustment, and inverse normal transformation were performed as done for the species-level clades.

Quantification of microbial pathways based on gene set variance analysis

For obtaining the pathway scores of each sample, we performed gene set variance analysis (GSVA; version 3.8–0).[64] As an input for the GSVA, we utilized the gene abundance data before the aggregation into orthologs. We removed genes detected (i) in less than 20% of the samples or (ii) in no sample in any of the sequencing batches and performing log-transformation. We added the pseudo-counts (half of the minimum non-zero value) to the zeros before the log transformation. The KEGG gene sets were defined according to the KEGG pathway. Gene sets that contained over 30,000 genes or under 10 genes were excluded from the calculation of the pathway scores, resulting in 148 and 152 microbial pathways respectively for datasets 1 and 2. After the removal of outliers (outside the range of mean±5s.d.), pathway scores were corrected for the covariates using PEER accounting for 15 unobserved confounders as well as the known confounders, such as age, age,[2] sex, sequencing batches, facility, diseases, and three genotype-based principal components. By utilizing PEER, we removed unwanted variations of the data. We included age[2] as one of the covariates to correct for non-linear age-related changes in the gut microbiome.[65] The residuals were standardized with the inverse normal transformation.

Calculation of α-diversity of the metagenome

Quality-controlled reads were down-sampled to 9,000,000 paired-ends reads to adjust the differences in the library sizes between the samples. Then, the down-sampled reads were used for the quantification of the species-level clades as described above. The resulting abundance data, before clade-level QC, was subjected to the diversity function in the R package vegan (version 2.5_6) to calculate the Shannon index.

Plasma metabolome profiling based on the CE-TOFMS and LC-TOFMS

Plasma metabolite profiling was performed in previous studies.[36,49] Briefly, plasma samples from the participants were collected at collaborating facilities. Metabolite extraction and metabolome analysis were conducted at Human Metabolome Technologies (HMT), Japan. For CE-TOFMS analysis, 50 μL of serum was added to 450 μL of methanol containing internal standards (H3304-1002, HMT) at 0°C to inactivate enzymes. The internal standards were L-methionine sulfone and D-camphor-10-sulfonic acid for cationic mode and anionic mode, respectively. 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 resuspended in 50 μL of Milli-Q water for CE-MS analysis.

For LC-TOFMS analysis, 500 μL of serum was added to 1500 μL of 1% formic acid/acetonitrile containing internal standard solution (Solution ID: H3304-1002, HMT) at 0°C to inactivate enzymes. D-camphor-10-sulfonic acid was used for the internal standard in both the positive and negative modes. The solution was thoroughly mixed and centrifuged at 2300 × g and 4°C for 5 min. The supernatant was filtrated by using a Hybrid SPE phospholipid (55261-U, Supelco, Bellefonte, PA, USA) to remove phospholipids. The filtrate was desiccated and dissolved with 100 μL of iso-propanol/Milli-Q for LC-MS analysis.

Metabolome analysis was conducted with 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 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 an ODS column (2 × 50 mm, 2 μm). The equilibration time was 7.5 min. In this service, the sensitivity of the analysis was checked by the signals from the internal standards rather than using pooled QC samples. In addition, for the LC-TOFMS analysis, the sensitivity of the analysis was confirmed by measuring the D-camphor-10-sulfonic acid solution for every ten samples. Note that all the samples were measured in a single experiment for each batch to mitigate the measurement noises.

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) as previously described.[66] Briefly, the raw data was subjected to the noise-filtering, baseline correction, peak

detection and integration of the peak area from sliced electropherograms (the width of each electropherogram was 0.02 m/z). The accurate m/z value for each peak detected within the time domain was calculated with Gaussian curve-fitting to the mass spectrum on the m/z domain peak. The alignment of peaks in multiple measurements was done by dynamic programming (DP)-based techniques[67] with slight modifications. The method picked up a few representative peaks using the Douglas-Peucker algorithm[68] from unit m/z electropherograms, found corresponding peaks across multiple samples by DP, and optimized the numerical parameters of the normalization function for CE-migration.[69] Instead of representative peaks, the service used the detected peaks with accurate m/z values and regarded the peaks whose m/z difference was less than 20 ppm as ones that were derived from the same electropherograms. Signal peaks corresponding to isotopomers, adduct ions, and other product ions of known metabolites were excluded.

The remaining peaks were annotated according to the HMT metabolite database based on their m/z values with the MTs and RTs determined by TOFMS. The tolerance range for the peak annotation was configured at ±0.5 min for MT and ±10 ppm for m/z at CE-TOFMS, ±0.3 min for RT and ±25 ppm for m/z at LC-TOFMS, respectively. Areas of the annotated peaks were normalized based on the levels of the internal standard for each modality (CE-TOFMS, L-methionine sulfone and D-camphor-10-sulfonic acid for cationic mode and anionic mode, respectively; LC-TOFMS, D-camphor-10-sulfonic acid) and sample amounts to obtain relative levels of each metabolite. Then, we detected outlier samples by PCA (PC 1–6) and two samples in dataset 1 were removed because they were outliers in the PCA analyses.

After sample QC, we performed the normalization of the plasma metabolite abundances. We removed metabolites detected in less than 30% of the samples. After metabolite QC, 363 and 368 metabolites were retained for datasets 1 and 3, respectively. Then, metabolite abundances were log-transformed and outliers (outside the range of mean±5s.d.) were removed per clades. We added the pseudo-counts (half of the minimum non-zero value) to the zeros before the log transformation. Then, log-transformed metabolite abundances were corrected for the covariates using PEER accounting for 30 unobserved confounders as well as the known confounders, such as age, age,[2] sex, facility, diseases, and three genotype-based principal components. By utilizing PEER, we removed unwanted variations of the data. The residuals were standardized with the inverse normal transformation.

Genotyping of the samples based on the SNP array

In this study, we utilized both the previously published[25] and newly generated SNP array-based genotype data. We performed SNP array-based genotyping using Infinium Asian Screening Array (Illumina, San Diego, CA, USA). This genotyping array was built using an EAS reference panel including whole genome sequences, which enabled effective genotyping in EAS populations.

We applied stringent quality control filters to the genotyping dataset using PLINK (version 1.90b4.4)[70] as described elsewhere.[71] We excluded individuals with a genotyping call rate of <0.98. All the individuals were estimated to be of EAS ancestry, based on the PCA with the samples of the 1KG dataset using EIGENSTRAT[72] (version 6.1.4). We further excluded SNPs with (i) call rate <0.99, (ii) minor allele count <5, and (iii) P-values for Hardy-Weinberg equilibrium $<1.0 \times 10^{-5}$. For pairs of closely related individuals (PI_HAT calculated by PLINK >0.185), we removed either of the related individuals.

We performed genome-wide genotype imputation to estimate untyped variants computationally. We used the combined reference panel of 1KG Project Phase 3 version 5 genotype (n = 2,504) and Japanese WGS data (n = 1,037)[73,74] as a haplotype reference for genotype imputation. First, we excluded SNPs with >7.5% allele frequency difference with the representative reference datasets of Japanese ancestry, namely the combined reference panel aforementioned[73,74] and the allele frequency panel of Tohoku Medical Megabank Project.[75] Second, we conducted haplotype estimation to improve imputation performance using SHAPEIT (version 4.2.1)[76] with haplotype reference. After the prephasing, we used Minimac4 (version 1.0.1)[77] for genotype imputation. The variants imputed with Rsq >0.7 were used for the downstream analysis.

Genotyping of the samples based on the whole genome sequencing

In this study, we utilized both the previously published[50] and newly generated whole genome sequencing data. DNA samples extracted from whole blood were sequenced at Macrogen Japan Corporation. DNA quantity was measured by Picogreen, and degradation of DNA was assessed by gel electrophoresis. All libraries were constructed using the TruSeq DNA PCR-Free Library Preparation Kit according to the manufacturer's protocols. Libraries were sequenced on HiSeqX (Illumina, San Diego, CA, USA) with a mean coverage of 16.4×. The reads produced by HiSeqX were processed as previously described.[50] Briefly, sequenced reads were aligned against the reference human genome with the decoy sequence (GRCh37, human_g1k_v37_decoy) using BWA-MEM (version 0.7.13). Duplicated reads were removed using Picard MarkDuplicates (version 2.10.10). After Base-quality score recalibration implemented in GATK (versions 3.8–0), we generated individual variant call results using HaplotypeCaller and performed multi-sample joint-calling of the variants via GenotypeGVCFs. We set genotypes satisfying any of the following criteria as missing: (i) DP < 5, (ii) GQ < 20, or (iii) DP > 60 and GQ < 95, then removed variants with low genotyping call rates (<0.90). We performed Variant Quality Score Recalibration for SNVs and short indels according to the GATK Best Practice recommendations and adopted the variants, which passed the QC criteria. We further removed the variants (i) located in the low complexity regions, (ii) with ExcessHet >60, or (iii) with Hardy–Weinberg P-value $<1.0 \times 10^{-10}$. We kept only those presenting a non-significant difference in allele frequency ($p > 1.0 \times 10^{-10}$ provided by chi-square test) in the following representative reference datasets of Japanese ancestry: the combined reference panel of 1KG Phase 3 version 5 genotype ($N_{Japanese}$ = 104) and Japanese WGS data (N = 1037) used for the aforementioned genotype imputation,[74] and the allele frequency panel of Tohoku Medical Megabank Project[75] (ToMMo 8.3KJPN Allele Frequency

Panel, *N* = 8,380). Genotype refinement was performed using Beagle (version 5.1).[78] For pairs of closely related individuals (PI_HAT calculated by PLINK >0.185), we removed either of the related individuals.

GWAS for the microbial traits

We performed genome-wide linear regression analysis with an additive effect model using the SNP array-based imputed genotype data and normalized microbial traits using plink2 (version 2.00a3 9 Apr 2020)[79] separately for datasets 1 and 2. Then, a fixed-effect meta-analysis was performed for the variants with MAF >0.01 using the METASOFT (version 2.0.0)[80]. The genome-wide significance threshold was set at $p < 5 \times 10^{-8}$ and study-wide significances were set at $p < 5 \times 10^{-8}$/number of the tested traits. We used LocusZoom (version 1.4)[81] to make the plots for the microbiome-associated loci. Annotation of the genetic variants was performed with ANNOVAR (Mon, 8 Jun 2020).[82]

As for the pair of the microbial taxa and genetic variants with genome-wide association ($p < 5 \times 10^{-8}$) in either of the two studies[11,13] which were based on the shotgun sequencing analysis with the NCBI taxonomy, we evaluated the reproducibility in our study, namely checking the significance and consistency of the effect directions. The list of the evaluated variants is described in Table S5. We also evaluated the association with all the microbial traits for the variants that were previously nominated for the association with microbial traits. We listed up such variants from the 16S rRNA-based[8–10,16] and shotgun sequencing-based[11–13] studies, including both single cohort analyses and meta-analyses, with a significance threshold of $p < 5 \times 10^{-8}$. Note that for one study,[16] a significance threshold of $p < 2.5 \times 10^{-8}$ was adopted, therefore we used that value for the study. Since some studies reported variants in LD relationships, we evaluated the number of independent variants to set appropriate significance thresholds. The 496 tested variants could be clumped into 341 variants at $r^2 = 0.1$. The summary for the included studies and list of the evaluated variants are described in Table S6.

Association between the ABO blood group and microbial traits

First, we evaluated the results of the GWAS for the genetic variants which were linked to each ABO blood group (chr9:136146597:C>T linked to the blood group A, chr9:136131322:G>T linked to the blood group B, and chr9:136132908:T>TC linked to the blood group O[30,31]) and all the microbial traits. Then, to directly evaluate the association between the ABO blood group and microbial traits, we determined the ABO blood group of the participants based on the best guess imputed genotypes of the chr9:136131322:G>T and chr9:136132908:T>TC in the *ABO* gene. Differences in the normalized microbial traits between each pair of the blood group were evaluated by the linear regression analysis (normalized microbial traits ~ blood group) for each dataset followed by meta-analysis with metafor (version 3.0–2) package for R.

We determined the secretor status of the participants based on the genotype of chr19:49206631:A>T in the *FUT2* gene based on the previous finding.[33,34] We labeled 70 participants whose imputed dosage of the T allele was $\leq$0.15 as non-secretor and 417 participants whose imputed dosage of the T allele was $\geq$0.85 as secretors. The remaining 37 participants were removed from the analysis of the secretor status. Per blood group differences in the normalized microbial traits between secretor and non-secretor were evaluated by the linear regression analysis (normalized microbial traits ~ secretor status) for each dataset followed by meta-analysis with metafor (version 3.0–2) package for R.

We utilized JMAG,[35] a database of the prokaryotes genomes reconstructed from the Japanese metagenome data, including those used in this study, to evaluate which bacterial taxa had the *agaE*, *agaS*, and *yydK* gene orthologs. In the previous study, protein-coding genes on the 19,084 MAGs were predicted by Prokka[83] (version 1.14.6) with the specification of the kingdom annotated by CheckM (version 1.0.12).[84] Then all the predicted protein sequences were dereplicated at 100% AAI by MMseqs2[85] (version 13.45111) with the following parameters; –cov-mode 1 -c 0.8 –kmer-per-seq 80 –min-seq-id 1. Blast searches to the *agaE*, *agaS*, and *yydK* sequences in the KEGG protein databases were performed for the non-redundant predicted protein sequences. First, to reduce the computation costs, only the sequences annotated as *agaE* (K02747), *agaS* (K02082), and *yydK* (K03489) were extracted from the KEGG protein database, and a small reference database was constructed from the extracted sequences. Then, the protein sequences on the MAGs were subjected to the BLASTP search against the small reference database with DIAMOND (parameters: -f 6 -b 15.0 -k 1 -e 1e-6 –subject-cover 50) with the same option as the aforementioned functional analysis. The protein sequences with a hit were further subjected to the BLASTP[86] search against the whole KEGG protein databases as done in the aforementioned gene-level quantification with DIAMOND, and the *agaE*, *agaS*, and *yydK* gene orthologs on the MAG were identified. The taxonomic information of the MAGs with *agaE*, *agaS*, and *yydK* was then extracted and summarized.

GWAS for the plasma metabolites, PheWAS, and colocalization analysis

We performed genome-wide linear regression analysis with an additive effect model using the WGS-based genotype data and normalized plasma metabolite data using plink2 separately for datasets 1 and 3. Since WGS data were available for the participants from which plasma metabolites were profiled, we utilized WGS-based genotype data rather than SNP array-based genotype data for the plasma metabolite GWAS. Then, a fixed-effect meta-analysis was performed for the variants with MAF >0.01 using the METASOFT. The genome-wide significance threshold was set at $p < 5 \times 10^{-8}$ and study-wide significances were set at $p < 5 \times 10^{-8}$/306 (number of the metabolites commonly detected in datasets 1 and 3). Annotation of the genetic variants was performed with ANNOVAR.

For PheWAS, we looked up the association of the study-wide metabolite-associated variants ($p < 5 \times 10^{-8}$/306) and their tagged variants ($r^2 > 0.6$ in Japanese) to the 155 diseases, three anthropometric traits, and 35 biomarkers from the previously released GWAS

summary statistics.[41] Then, if the metabolite-associated variants or their tagged variants had an association with the traits with $p < 1 \times 10^{-4}$, we performed colocalization analyses with coloc (version 5.1.1)[87] using the variants located within ±250kbp of the metabolite-associated variants. We used LocusZoom to make the plots for the metabolite-associated loci.

Microbiome–metabolome association analysis

We performed linear regression analysis for dataset 1 using the lm() function implemented in R with the following formula; microbial traits ~ plasma metabolites + age + age^2 + sex + sequencing batches + facility + diseases. We included age^2 as one of the covariates to correct for non-linear age-related changes in the gut microbiome.[65] Sample QC, trait QC, log-transformation, and per-trait outlier removal were performed for the microbial traits and metabolite data as described above, while PEER was not applied. FDR was calculated by the Benjamini-Hochberg procedure.

For the construction of the microbiome–metabolome network, we extracted all the microbial traits–metabolite associations that satisfied the Bonferroni-corrected significance. Then, we constructed a network plot with the ggraph package (version 2.0.4). We specified 'kk' as a layout option to place nodes based on the spring-based algorithm by Kamada and Kawai.

Association between the α-diversity and bile acids was evaluated by the linear regression with the following formula; Shannon-index ~ bile acid + age + age^2 + sex + sequencing batches + facility + diseases.

QUANTIFICATION AND STATISTICAL ANALYSIS

For GWAS of the gut microbial traits and plasma metabolites, we used fixed-effect meta-analysis for calculating test statistics. Statistical tests for each dataset were performed with the linear regression implemented in the PLINK2 software.[79] All the ABO blood type association tests were fixed-effect meta-analyses from the results of the linear regression analysis implemented in the R. For other statistical tests, we used linear regression with the lm() function and Wald's test as implemented in the R. Please also refer to figure legends and METHOD DETAILS for details of statistical analysis. Number of the samples used in the analyses are described in Table S1.