Published online 7 March 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/nargab/lqac019

OMARU: a robust and multifaceted pipeline for metagenome-wide association study

Toshihiro Kishikawa[1,2,3,*], Yoshihiko Tomofuji[1,4], Hidenori Inohara[2] and Yukinori Okada[1,4,5,6,7,*]

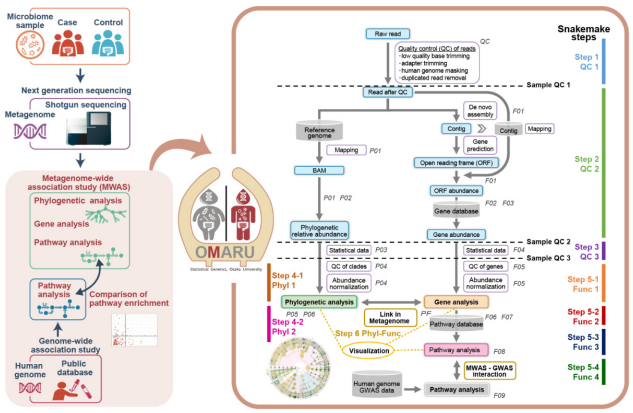
[1]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan, [2]Department of Otorhinolaryngology-Head and Neck Surgery, Osaka University Graduate School of Medicine, Suita 565-0871, Japan, [3]Department of Head and Neck Surgery, Aichi Cancer Center Hospital, Nagoya 464-8681, Japan, [4]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan, [5]Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Kanagawa 230-0045, Japan, [6]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan and [7]Center for Infectious Disease Education and Research (CiDER), Osaka University, Suita, Japan

Received December 28, 2021; Revised February 04, 2022; Editorial Decision February 14, 2022; Accepted February 18, 2022

ABSTRACT

Microbiome is an essential omics layer to elucidate disease pathophysiology. However, we face a challenge of low reproducibility in microbiome studies, partly due to a lack of standard analytical pipelines. Here, we developed OMARU (Omnibus metagenome-wide association study with robustness), a new end-to-end analysis workflow that covers a wide range of microbiome analysis from phylogenetic and functional profiling to case–control metagenome-wide association studies (MWAS). OMARU rigorously controls the statistical significance of the analysis results, including correction of hidden confounding factors and application of multiple testing comparisons. Furthermore, OMARU can evaluate pathway-level links between the metagenome and the germline genome-wide association study (i.e. MWAS-GWAS pathway interaction), as well as links between taxa and genes in the metagenome. OMARU is publicly available (https://github.com/toshi-kishikawa/OMARU), with a flexible workflow that can be customized by users. We applied OMARU to publicly available type 2 diabetes (T2D) and schizophrenia (SCZ) metagenomic data (n = 171 and 344, respectively), identifying disease biomarkers through comprehensive, multilateral, and unbiased case–control comparisons of metagenome (e.g. increased *Streptococcus vestibularis* in SCZ and disrupted diversity in T2D). OMARU improves accessibility and reproducibility in the microbiome research community. Robust and multifaceted results of OMARU reflect the dynamics of the microbiome authentically relevant to disease pathophysiology.

GRAPHICAL ABSTRACT



INTRODUCTION

Microbiome is one of the major research areas in human diseases towards implementation of personalized medicine based on multi-layer omics data. Recent interests are on multidimensional integration of metagenome data with other omics layers such as host genome and metabolome, as well as deep analysis within the single metagenomic layer (1,2). Analytical approaches of microbiome are shifting from amplicon sequencing of 16S ribosomal RNA genes to whole-genome shotgun sequencing. However, we face a challenge of low reproducibility in findings of microbiome

*To whom correspondence should be addressed. Tel: +81 6 6879 3971; Email: yokada@sg.med.osaka-u.ac.jp
Correspondence may also be addressed to Toshihiro Kishikawa. Tel: +81 6 6879 3971; Email: t.kishikawa@aichi-cc.jp

studies. Differences in physiological variables and lifestyles of the samples also have been reported as a factor yielding this problem (1,3). In addition, we still lack a gold standard analytical pipeline which can overcome the problem of low reproducibility (3,4).

Here, we introduce OMARU (**O**mnibus **m**etagenome-wide **a**ssociation study [MWAS] with **r**ob**u**stness), a new end-to-end metagenome analysis workflow (Figure 1). Through implementation of rigorous quality control (QC) of shotgun sequence reads, samples, clades, and genes, OMARU constructs phylogenetic and functional profiling of the metagenome, the two main analytical pipelines. Three major components of the case–control association tests of MWAS (i.e. phylogenetic, gene, and biological pathway analyses) are subsequently conducted with rigorous handling of false positives in statistical analysis (5–7). In addition to solving the low reproducibility of metagenomic study, OMARU provides integrative analyses. As an example, OMARU can evaluate pathway-level links between the metagenome and the germline genome-wide association studies (GWAS) of the host genome. Furthermore, OMARU identifies the links between taxa and genes in the metagenome utilizing the results of phylogenetic and gene analyses. OMARU also visualizes attractive figures which enable comprehensive summary of the association test results. The referenced databases, which substantially affect the analytic results, is currently being rapidly expanded (8,9). OMARU is a flexible and extensible workflow that can be customized, such as adding an up-to-date database.

MATERIALS AND METHODS

Quality control

OMARU handles the shotgun sequencing data in the FASTQ format as input (currently, 16S rRNA data is not supported). QC of the sequencing reads is applied to maximize the quality of datasets as follows: (i) trimming of low-quality bases using Trimmomatic (10), (ii) identification and masking of human reads using bowtie2 (11) and BMTagger (12) and (iii) removal of duplicated reads using PRINSEQ-lite (13). As for QC of samples, there exist three factors for selecting samples to be excluded as follows; (i) overall quality of sequencing reads, (ii) status of phylogenetic abundance and mapping rates, (iii) status of contigs and open reading frames (ORFs) in assembly-based approach and mapping rates in mapping-based approach, and (iv) principal component analyses (PCA) in the phylogenetic data and gene abundance data. OMARU sequentially outputs graphical figures and tables representing statistical matrixes of each procedure, helping users select samples to be excluded at each step (Figures 1 and 2A). Clades and genes detected in less than the pre-defined threshold of the samples (e.g. 20%), or in no sample in either cases or controls, are removed. Besides, clades with an average relative abundance less than the pre-defined threshold of total abundance are removed (default: 0.001%).

Case-control association test for phylogenetic data

OMARU adopts a mapping-based approach to utilize the advantages of paired-end reads and reduce mapping errors. Users can flexibly customize the reference data in a FASTA format to the appropriate one: Default is modified DNA sequences of the Unified Human Gastrointestinal Genome (8). After read-mapping using bowtie2 (11), relative abundance of each clade is quantified for each sample up to the six taxonomic levels (L2: phyla, L3: classes, L4: orders, L5: families, L6: genera and L7: species). Subsequently, the relative abundance profiles are normalized using log transformation. Case-control association tests are performed using the lm function implemented in the *R* statistical software. Users can incorporate covariates for adjustment, such as sex and age. OMARU generally requires a sufficient number of principal components as covariates to robustly adjust the effect of hidden confounding factors and suppress *P*-value inflation (Figure 2B).

Empirical null distributions of the minimum *P*-values ($= P_{\min}$) are calculated based on a phenotype permutation procedure ($\times$ 10,000 iterations) to control the type I error rates (14). The empirical Bonferroni significance threshold is defined at a significance level of 0.05, as the 95th percentile of $P_{\min}$ ($= P_{\text{sig}}$). The 95% confidence interval for $P_{\min}$ is calculated by a bootstrapping method of the Harrell-Davis distribution-free quantile estimator (Figure 2C). In addition to the standard figures to visualize distribution of statistics such as quantile-quantile and volcano plots (Supplementary Figure S1), OMARU illustrates a phylogenetic tree indicating the case–control association results of multilayered taxonomic levels (Figure 2D).

Case–control association test for functional data (gene and pathways)

Gene abundance data of metagenome are constructed by the assembly-based approach as follows; (i) *de novo* assembly of the sequencing reads into contigs using MEGAHIT (15), (ii) prediction of open reading frames (ORFs) on the contigs with Prokka (16), (iii) alignment of ORF against an appropriate database (default: UniRef90 (17)) with DIAMOND (18) and (iv) quantification of gene abundance by mapping the sequencing reads to the assembled contigs using bowtie2 (11). Normalization of gene abundance is conducted by the two steps. First, the ORF abundance is defined as the depth of each ORF's region of the ORF catalog according to the mapping result to avoid the bias of the gene lengths. Second, the gene abundance is adjusted by the sum of the ORF abundance for each sample to correct potential bias of heterogeneity in the total amount of sequence reads among the samples. Next, a rank-based inverse normal transformation is applied to correct the heterogeneity of each gene's abundance and distribution. Association tests are in the same way as phylogenetic analysis, including covariates and empirical threshold (Figure 3A).

As for the pathway analysis, OMARU adopts a gene set enrichment analysis using the ranking of the genes by *z*-values in case–control gene association tests. The pathway database could be flexibly customized (Default is Gene Ontology (19)).

Links between the microbe MWAS and the germline GWAS of host

OMARU identifies disease-specific biological pathway links between the microbe MWAS and the germline GWAS of

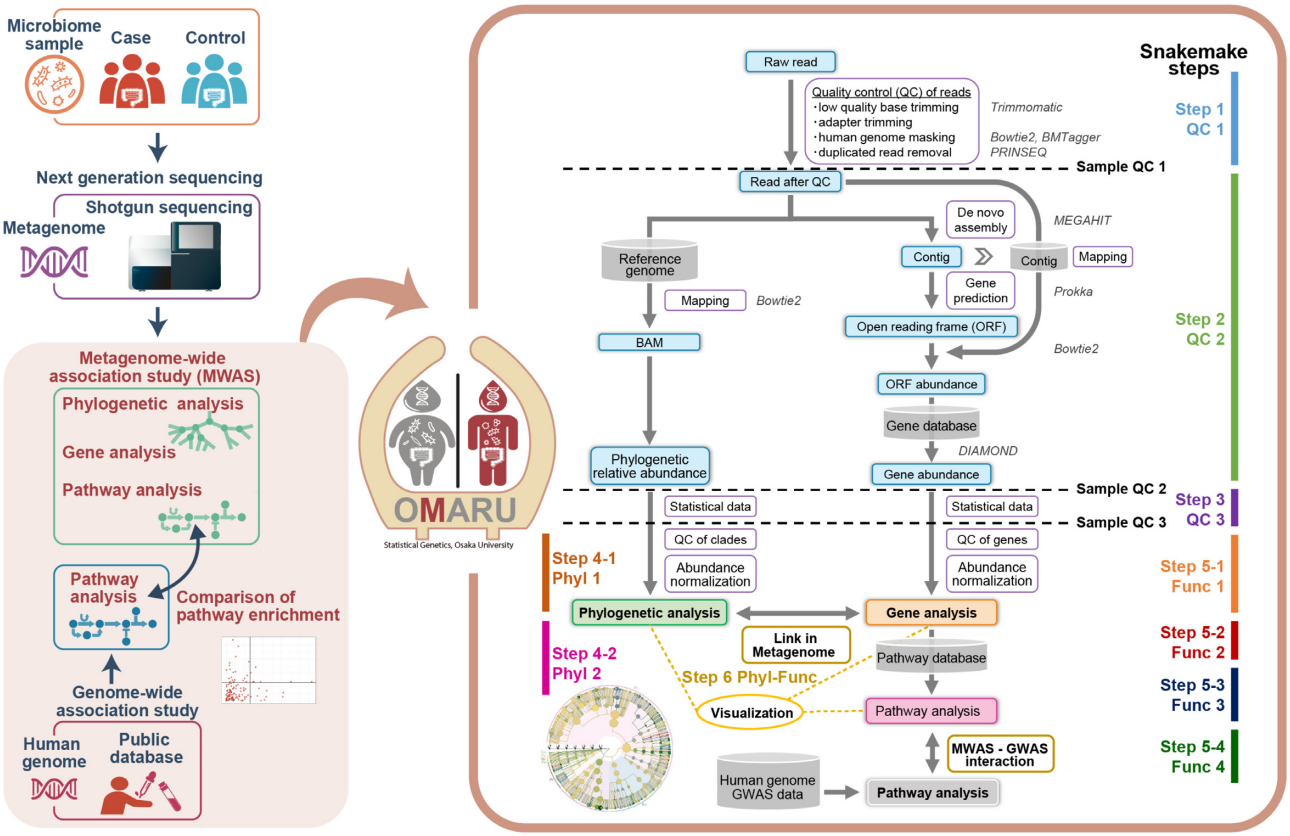


Figure 1. OMARU workflow and details as bioinformatics pipelines for the metagenome-wide association study. OMARU workflow. Using shotgun sequencing data, metagenome-wide association studies (phylogenetic, gene and pathway analyses) and additional analyses are performed, including comparing pathway analyses between genome-wide association studies (GWAS) and metagenome-wide association study (MWAS).

host (5–7). The result of pathway analysis using summary statistics of GWAS for the target disease is required as input. OMARU evaluates the overlap between the MWAS and GWAS in the pathway enrichment by Fisher's exact test, based on the classification of pathways with *P*-value threshold of 0.05 (Figure 3C).

Links between taxa and genes in the metagenome

Organisms of origin for each gene are an important factor to understand microbiome biology. While gene databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (20) and UniProt(21) collect organisms of origin, such information are based on the specific link between the registered gene and organisms, and may not reflect the real link in the target metagenomic sample. By tracing back to the level of sequencing reads, OMARU can directly estimate organisms of the origin for each gene in the target data (Figure 3B, Supplementary Figure S2).

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

For calculating diversities, all samples should be down-sampled at the appropriate same number of reads. OMARU calculates α-diversity (within-sample diversity) as a Shannon index based on the gene abundance and the six levels of phylogenetic relative abundance (L2–L7) for each sample. Case–control comparison are performed with predefined covariates and the effect size of disease state is evaluated. To evaluate β-diversity, multidimensional scaling (MDS) on the Bray-Curtis dissimilarity is used. For evaluating case–control differences in the dissimilarity, OMARU performs permutational multivariate analysis of variance (PERMANOVA) (22) using the adonis() function in R package vegan.

RESULTS

We adopted the two public fecal metagenomic data of schizophrenia (SCZ; 90 SCZ patients and 81 healthy controls) and type 2 diabetes (T2D; 170 T2D patients and 174 healthy controls) for a practical example of operation of OMARU (23,24).

In sample QC of the SCZ data, we excluded one SCZ sample that had singleton genes beyond four standard deviations and was an outlier of both phylogenetic and gene abundance data (Figure 2A). We used a phylogenetic reference, which was constructed by integrating those registered by Nishijima *et al.* (25) and those newly identified from the human gut bacteria projects (9,26,27), as previously described (5,6). We had 692 clades for the SCZ case–control association test, including 10 phyla (L2), 23 classes (L3), 34 orders (L4), 69 families (L5), 156 genera (L6) and 400 species (L7). We adopted sex, age, body mass index

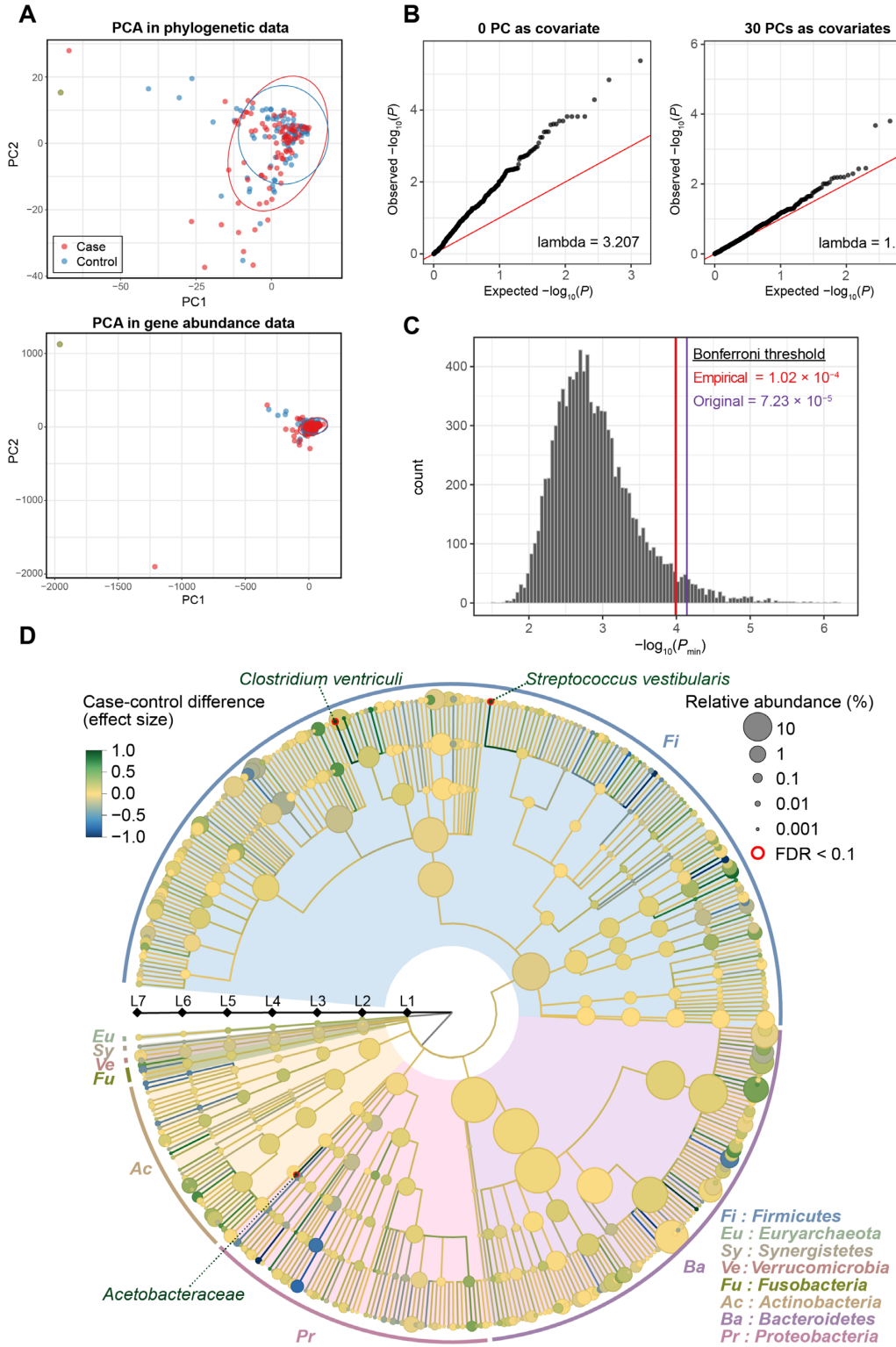


Figure 2. MWAS results of QC and phylogenetic analysis. (**A**) Principal component analysis (PCA) in phylogenetic and gene abundance of the schizophrenia (SCZ) data. The green dots represent the excluded sample as a result of quality control. (**B**) Quantile-quantile plots of the phylogenetic MWAS *P*-values of the clades in the SCZ data. The x-axis indicates log-transformed empirically estimated median *P*. The y-axis indicates observed $-\log_{10}(P)$. The diagonal dashed line represents $y = x$, which corresponds to the null hypothesis. The left and right figures show the results of including 0 principal component (PC) and 30 PCs as covariates, respectively, which indicates that PCs suppress the inflation of *P*-values. (**C**) A histogram of minimum *P*-values in the phenotype permutation procedure in the SCZ data. Vertical lines of red and purple indicate an empirical Bonferroni significance threshold at a significance level of 0.05 and a standard Bonferroni significance threshold in multiple comparison procedure ($0.05/692 = 7.23 \times 10^{-5}$), respectively. (**D**) A phylogenetic tree. Levels L2–L7 are from the inside layer to the outside layer in the SCZ data. The size and the color of the dots represent relative abundance and effect sizes, respectively. The three clades with significant case–control associations (false discovery rate < 0.05) are outlined in red.

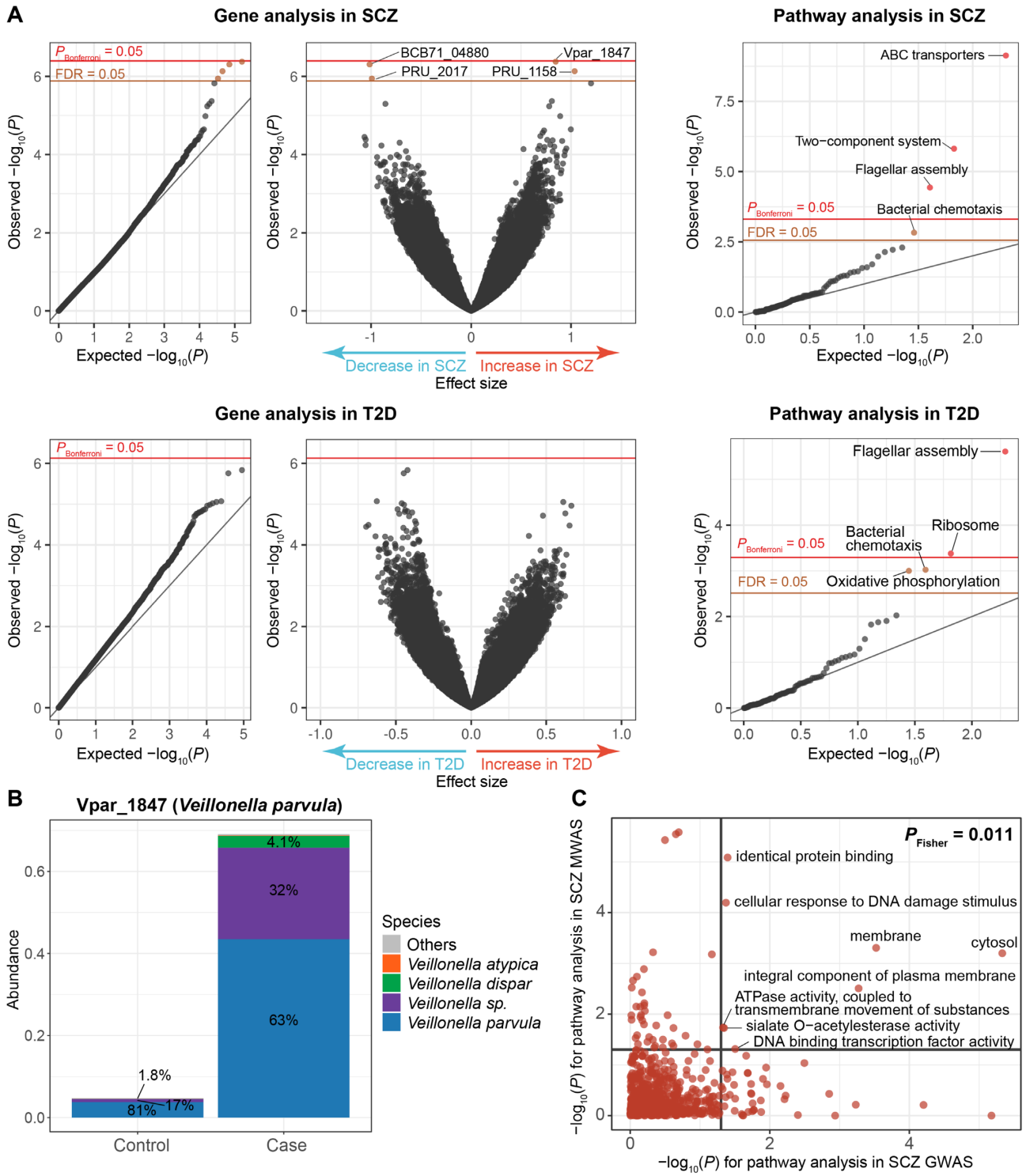


Figure 3. MWAS results of functional analysis. (**A**) Results of functional association tests in schizophrenia (SCZ) and type 2 diabetes (T2D). Left figures are quantile-quantile plots of the *P*-values in the gene association tests. The x-axes indicate empirically estimated median -$\log_{10}(P)$. The y-axes indicate observed -$\log_{10}(P)$. The diagonal grey lines represent $y = x$, which correspond to the null hypothesis. The horizontal red lines indicate the empirical Bonferroni-corrected threshold ($\alpha = 0.05$), and the brown line indicates the empirically estimated FDR threshold (FDR = 0.05). Center figures are volcano plots. The x-axes indicate effect sizes in linear regression. The y-axes, horizontal lines, and dot colors are the same as in the left quantile-quantile plots. Right figures are quantile-quantile plots of the *P*-values in the pathway association tests. Genes and pathways with *P*-values less than the Bonferroni threshold are plotted as red dots. Genes and pathways with FDR less than 0.05 are plotted as brown dots, and others are plotted as black dots. FDR; false discovery rate. (**B**) Links in the metagenome data between taxa and Vpar_1847, one of the schizophrenia-associated genes. Stacked bar graphs indicate the species of origin for each gene and their percentage, divided into cases and controls. The parentheses in each title represent the organism registered as the origin of the genes in the database. (**C**) Comparison of *P*-values of GO analyses between the SCZ MWAS and GWAS data. The x-axis indicates the *P*-value in the SCZ GWAS data. The y-axis indicates the *P*-value in the SCZ MWAS data. The horizontal and vertical black lines indicate *P* of 0.05. The overlap of the GO enrichment was evaluated by classifying the GO terms based on the significance threshold of $P < 0.05$ or $P \geq 0.05$ and using Fisher's exact test.

(BMI) and the top 30 principal components as covariates. In multiple test correction, empirically estimated Bonferroni threshold was lower than the standard Bonferroni threshold (Figure 2C). It could reflect that microbiome composition within an individual was not independent between clades. We identified the three clades significantly increased in SCZ (FDR < 0.05; Figure 2D, Supplementary Figure S3, Supplementary Table S1). We had 789 clades for the T2D case–control association test and identified no clades with significant association. In both diseases, the numbers of disease-associated clades were considerably lower than those in the reference papers and other metagenome studies(23,24). Correction of hidden confounding factors mainly led to this result. The quantile-quantile plots of *P*-values in the SCZ data showed that the analysis without adopting no PCs as covariates demonstrated severe inflation of *P*-values and a large number of false positives (Figure 2B). *Streptococcus vestibularis*, one of the three SCZ-associated clades identified by OMARU, was reported to induce deficits in social behavior and alter neurotransmitter levels in peripheral tissues in recipient mice(23). Thus, OMARU is featured by its ability to specify robustly disease-associated clades by optimally adjusting confounding factors.

We selected KEGG database (20) as references of gene and biological pathway. After gene-level QC, we retained 185 663 and 104 487 genes for SCZ and T2D case–control comparison, respectively. In functional association tests, we obtained results with suppression of the inflation of *P*-values by adjusting covariates in the same way as the phylogenetic analyses. We identified four SCZ-associated genes, four SCZ-associated pathways, and four T2D-associated pathways (FDR < 0.05; Figure 3A, Supplementary Table S2 and S3). In the analysis of link between phylogenetic and gene data, Vpar_1847, one of the four SCZ-associated genes, was estimated to be derived from multiple *Veillonella* spp. (Figure 3B, Supplementary Figure S2). These clades were not significantly associated in our phylogenetic analyses, while their increase in SCZ was highlighted in the referenced paper (23). The cross-sectional assessment of OMARU could suggest that this gene may be an essential factor in the effect on the SCZ pathogenesis of *Veillonella* spp.

As for the MWAS–GWAS interaction, we used PASCAL with summary statistics from the SCZ GWAS (22,778 cases and 35 362 controls) (28) and the T2D GWAS (77 418 cases and 433 5440 controls) (29) in order to determine GO term enrichment of the human genome. We compared the *P*-values of the each GO term shared between the metagenome data and GWAS data. We found significant overlaps between the pathways enriched in the MWAS and GWAS ($P_{\text{Fisher}} = 0.011$ and 0.008 in SCZ and T2D, respectively; Figure 3C). Our results suggested that there was disease-specific links between human genome and metagenome, namely MWAS- GWAS interaction, in the pathology of SCZ and T2D.

We performed case–control comparison of α-diversity and β-diversity in the phylogenetic data (L2–L7) and the gene abundance data based on KEGG database. In SCZ, no significant differences of α-diversity in the phylogenetic data ($P > 0.05/6 = 0.0083$) and the gene abundance data ($P = 0.134$) were observed, and neither was β-diversity (Supplementary Table S4). In T2D, α-diversity in the taxonomic level of L3 and L4 ($P < 8.3 \times 10^{-3}$) and the gene abundance data ($P = 5.1 \times 10^{-3}$) significantly increased, while significant differences of β-diversity in the taxonomic level of L5–L7 ($P < 8.3 \times 10^{-3}$) and the gene abundance data ($P < 1.0 \times 10^{-4}$) were observed (Supplementary Figure S4, Supplementary Table S4).

DISCUSSION

While several bioinformatic tools for microbiome has been developed recently (30–35), OMARU has a unique characteristic as highlighted in case–control MWAS analysis using shotgun sequencing data. In contrast to several existing tools which are limited to a single part of the analysis, such as phylogenetic or functional analysis, OMARU provides end-to-end analysis from the processing of sequencing data, such as QC of reads and samples, to the three major analyses and the assessment of diversities. It should be meaningful to perform those analyses in a single pipeline with integrative assessments of the results of each part of the analysis, providing deep interpretation of case–control differences in the microbiome. Further, evaluation of links between the metagenome and host genome is one of the novel features of OMARU.

We demonstrated that OMARU yields robust and multifaceted results by using public metagenome data. OMARU identified a sample in the SCZ data to be excluded. It's quite difficult to perform sample QC manually in metagenome analyses and comprehensive decision based on multiple assessments is required. OMARU can provide users with multifaceted data to help them make the decision. By statistical processing in OMARU including reduction of false positives, SCZ-associated clades were narrowed down to the clade with functional support, which demonstrates the robustness of OMARU in identifying crucial biomarkers. While hidden confounding factors would better to be adjusted by integration of the covariates into a case–control model, it is not currently implemented in OMARU and thus considered to be one of the limitations.

In addition, integrative analyses with multifaceted evaluation, such as the MWAS-GWAS interaction and the links between disease-associated genes and clades, provided a comprehensive understanding of the microbiome-associated pathology. The metagenome of SCZ had little difference of diversities while T2D had significant ones compared to healthy controls. Diversity analysis provides evidence of microbiome's role in disease pathology from a different aspect than other analyses. We note that the metagenome analysis is still highly dependent on reference databases and database development is a challenge for the future.

In conclusion, OMARU, as a well-organized and user-friendly workflow, can improve the accessibility and reproducibility of MWAS in the microbiome research community. Robust and multifaceted results of OMARU, including the association with the host genome, reflect the dynamics of the microbiome authentically relevant to disease pathophysiology, leading to the identification of potential biomarkers.

DATA AVAILABILITY

OMARU is publicly available at https://github.com/toshi-kishikawa/OMARU and can be downloaded in the format of a Conda package.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

Japan Society for the Promotion of Science (JSPS) KAKENHI [19H01021, 20K21834]; Japan Agency for Medical Research and Development (AMED) [JP21km0405211, JP21ek0109413, JP21gm4010006, JP21km0405217, JP21ek0410075]; JST Moonshot R&D [JPMJMS2021, JPMJMS2024]; Takeda Science Foundation, Bioinformatics Initiative of Osaka University Graduate School of Medicine, Grant Program for Next Generation Principal Investigators at Immunology Frontier Research Center (WPI-IFReC), Osaka University.
Conflict of interest statement. None declared.

REFERENCES

1. Asnicar,F., Berry,S.E., Valdes,A.M., Nguyen,L.H., Piccinno,G., Drew,D.A., Leeming,E., Gibson,R., Le Roy,C., Khatib,H.A. *et al.* (2021) Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.*, **27**, 321–332.
2. Liu,X., Tang,S., Zhong,H., Tong,X., Jie,Z., Ding,Q., Wang,D., Guo,R., Xiao,L., Xu,X. *et al.* (2021) A genome-wide association study for gut metagenome in chinese adults illuminates complex diseases. *Cell Discov.*, **7**, 9.
3. Vujkovic-Cvijin,I., Sklar,J., Jiang,L., Natarajan,L., Knight,R. and Belkaid,Y. (2020) Host variables confound gut microbiota studies of human disease. *Nature*, **587**, 448–454.
4. 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.
5. 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.
6. 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.
7. 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.
8. Almeida,A., Nayfach,S., Boland,M., Strozzi,F., Beracochea,M., Shi,Z.J., Pollard,K.S., Sakharova,E., Parks,D.H., Hugenholtz,P. *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, **39**, 105–114.
9. 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.
10. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
11. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**, 357–359.
12. Rotmistrovsky,K. and Agarwala,R. (2011) *BMTagger*. ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/(01 February 2022, date last accessed).
13. Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.
14. Kanai,M., Tanaka,T. and Okada,Y. (2016) Empirical estimation of genome-wide significance thresholds based on the 1000 genomes project data set. *J. Hum. Genet.*, **61**, 861.
15. 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.
16. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
17. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B. and Wu,C.H. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
18. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
19. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
20. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
21. The UniProt Consortium. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
22. Anderson,M.J. (2001) A new method for non-parametricmultivariate analysis of variance. *Austral Ecol.*, **26**, 3246.
23. Zhu,F., Ju,Y., Wang,W., Wang,Q., Guo,R., Ma,Q., Sun,Q., Fan,Y., Xie,Y., Yang,Z. *et al.* (2020) Metagenome-wide association of gut microbiome features for schizophrenia. *Nat. Commun.*, **11**, 1612.
24. 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.
25. 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.
26. 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.
27. 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.
28. Lam,M., Chen,C.-Y., Li,Z., Martin,A.R., Bryois,J., Ma,X., Gaspar,H., Ikeda,M., Benyamin,B., Brown,B.C. *et al.* (2019) Comparative genetic architectures of schizophrenia in east asian and European populations. *Nat. Genet.*, **51**, 1670–1678.
29. Spracklen,C.N., Horikoshi,M., Kim,Y.J., Lin,K., Bragg,F., Moon,S., Suzuki,K., Tam,C.H.T., Tabara,Y., Kwak,S.-H. *et al.* (2020) Identification of type 2 diabetes loci in 433,540 east asian individuals. *Nature*, **582**, 240–245.
30. Pasolli,E., Truong,D.T., Malik,F., Waldron,L. and Segata,N. (2016) Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.*, **12**, e1004977.
31. Wirbel,J., Zych,K., Essex,M., Karcher,N., Kartal,E., Salazar,G., Bork,P., Sunagawa,S. and Zeller,G. (2021) Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.*, **22**, 93.
32. Norouzi-Beirami,M.H., Marashi,S.-A., Banaei-Moghaddam,A.M. and Kavousi,K. (2021) CAMAMED: a pipeline for composition-aware mapping-based analysis of metagenomic data. *NAR Genomics Bioinformatics*, **3**, lqaa107.
33. Eng,A., Verster,A.J. and Borenstein,E. (2020) MetaLAFFA: a flexible, end-to-end, distributed computing-compatible metagenomic functional annotation pipeline. *BMC Bioinf.*, **21**, 471.
34. de la Cuesta-Zuluaga,J., Ley,R.E. and Youngblut,N.D. (2019) Struo: a pipeline for building custom databases for common metagenome profilers. *Bioinformatics*, **36**, 2314–2315.
35. Clarke,E.L., Taylor,L.J., Zhao,C., Connell,A., Lee,J.-J., Fett,B., Bushman,F.D. and Bittinger,K. (2019) Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome*, **7**, 46.