

RESEARCH PAPER



Machine-learning assisted discovery unveils novel interplay between gut microbiota and host metabolic disturbance in diabetic kidney disease

I-Wen Wu^{a,b,*}, Yu-Chieh Liao^{c,*}, Tsung-Hsien Tsai^d, Chieh-Hua Lin^c, Zhao-Qing Shen^e, Yun-Hsuan Chan^d, Chih-Wei Tu^d, Yi-Ju Chou^f, Chi-Jen Lo^g, Chi-Hsiao Yeh^{b,h,i}, Chun-Yu Chen^{a,b}, Heng-Chih Pan^{a,b}, Heng-Jung Hsu^{a,b}, Chin-Chan Lee^{a,b}, Mei-Ling Cheng^{g,j,k}, Wayne Huey-Herng Sheu^{f,l,m}, Chi-Chun Lai^{b,h,n}, Huey-Kang Sytwu^{o,p}, and Ting-Fen Tsai^{e,f,q}

^aDepartment of Nephrology, Chang Gung Memorial Hospital, Keelung, Taiwan; ^bCommunity Medicine Research Center, Chang Gung Memorial Hospital, Keelung, Taiwan; ^cInstitute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan; ^dAdvanced Tech BU, Acer Inc, New Taipei City, Taiwan; ^eDepartment of Life Sciences and Institute of Genome Sciences, National Yang Ming Chiao Tung University, Taipei, Taiwan; ^fInstitute of Molecular and Genomic Medicine, National Health Research Institutes, Miaoli, Taiwan; ^gMetabolomics Core Laboratory, Healthy Aging Research Center, Chang Gung University, Taoyuan, Taiwan; ^hCollege of Medicine, Chang Gung University, Taoyuan, Taiwan; ⁱDepartment of Thoracic and Cardiovascular Surgery, Chang Gung Memorial Hospital, Taoyuan, Taiwan; ^jClinical Metabolomics Core Laboratory, Chang Gung Memorial Hospital, Taoyuan, Taiwan; ^kDepartment of Biomedical Sciences, College of Medicine, Chang Gung University (MLC), Taoyuan, Taiwan; ^lDivision of Endocrinology and Metabolism, Department of Internal Medicine, Taipei Veterans General Hospital, Taipei, Taiwan; ^mSchool of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan; ⁿDepartment of Ophthalmology, Chang Gung Memorial Hospital, Keelung, Taiwan; ^oNational Institute of Infectious Diseases and Vaccinology, National Health Research Institutes, Miaoli, Taiwan; ^pDepartment & Graduate Institute of Microbiology and Immunology, National Defense Medical Center, Taipei, Taiwan; ^qCenter for Healthy Longevity and Aging Sciences, National Yang Ming Chiao Tung University, Taipei, Taiwan

ABSTRACT

Diabetic kidney disease (DKD) is a serious healthcare dilemma. Nonetheless, the interplay between the functional capacity of gut microbiota and their host remains elusive for DKD. This study aims to elucidate the functional capability of gut microbiota to affect kidney function of DKD patients. A total of 990 subjects were enrolled consisting of a control group ($n = 455$), a type 2 diabetes mellitus group (DM, $n = 204$), a DKD group ($n = 182$) and a chronic kidney disease group (CKD, $n = 149$). Full-length sequencing of 16S rRNA genes from stool DNA was conducted. Three findings are pinpointed. Firstly, new types of microbiota biomarkers have been created using a machine-learning (ML) method, namely relative abundance of a microbe, presence or absence of a microbe, and the hierarchy ratio between two different taxonomies. Four different panels of features were selected to be analyzed: (i) DM vs. Control, (ii) DKD vs. DM, (iii) DKD vs. CKD, and (iv) CKD vs. Control. These had accuracy rates between 0.72 and 0.78 and areas under curve between 0.79 and 0.86. Secondly, 13 gut microbiota biomarkers, which are strongly correlated with anthropometric, metabolic and/or renal indexes, concomitantly identified by the ML algorithm and the differential abundance method were highly discriminatory. Finally, the predicted functional capability of a DKD-specific biomarker, *Gemmiger* spp. is enriched in carbohydrate metabolism and branched-chain amino acid (BCAA) biosynthesis. Coincidentally, the circulating levels of various BCAAs (L-valine, L-leucine and L-isoleucine) and their precursor, L-glutamate, are significantly increased in DM and DKD patients, which suggests that, when hyperglycemia is present, there has been alterations in various interconnected pathways associated with glycolysis, pyruvate fermentation and BCAA biosynthesis. Our findings demonstrate that there is a link involving the gut-kidney axis in DKD patients. Furthermore, our findings highlight specific gut bacteria that can act as useful biomarkers; these could have mechanistic and diagnostic implications.

ARTICLE HISTORY

Received 14 October 2024
Revised 24 January 2025
Accepted 21 February 2025







KEYWORDS

Diabetic kidney disease;
microbiota; machine
learning; branched-chain
amino acids


Introduction

Diabetic kidney disease (DKD) is the leading cause of end-stage kidney disease worldwide.¹ Hyperglycemia causes a cascade of molecular changes that triggers

increases in oxidative stress, together with higher levels of inflammatory cytokines, various growth factors, and a number of profibrotic substances. An imbalance in these important factors results in

CONTACT Chi-Chun Lai  chichun.lai@gmail.com  Community Medicine Research Center, Chang Gung Memorial Hospital, No. 222, Maijin Road, Anle District, Keelung City 204, Taiwan; Huey-Kang Sytwu  sytwu@nhri.edu.tw  National Institute of Infectious Diseases and Vaccinology, National Health Research Institutes, No. 35, Keyan Road, Zhunan Township, Miaoli County 350, Taiwan; Ting-Fen Tsai  tftsai@nycu.edu.tw  Department of Life Sciences and Institute of Genome Sciences, National Yang Ming Chiao Tung University, No. 155, Sec. 2, Linong Street, Beitou District, Taipei City 112, Taiwan

*Contributed equally to this article as co-first authors.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19490976.2025.2473506>

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

structural and functional changes to various molecular and intracellular moieties. In turn, this leads to mesangial expansion, podocyte injury, tubulointerstitial fibrosis and glomerulosclerosis.² Eventually, this metabolic disarrangement further aggravates host homeostasis, which leads to kidney damage.³

The gut microbiota is an extensive endogenous repertoire of microorganisms that produce many gut metabolites that have physiological functions.^{4–6} The uremic milieu, disease-related dietary modification and pharmaceutical interventions in DKD patients profoundly affect the functioning of the intestinal barrier, the species that are present in the gut microbiota, and the production of metabolites within the gut.^{7–10} Significant microbiota architectural changes, as well as functional alterations, have been extensively reported to be present in chronic kidney disease (CKD) and DKD patients.^{11–13} Currently, there are unmet challenges regarding the inter-comparability of such studies and the generalizability of their findings; this has impeded progress in our understanding of the place of the microbiome in medicine.¹⁴ The complexity of the high-dimensional datasets generated by full-length bacterial 16S rRNA gene sequencing and the meaningful analysis of the resulting metagenome sequences remains challenging when addressed by computational methods. Compositional tables are usually used to identify the relative abundances of specific species, but each sample contains a huge number of features, many of which are sparse in terms of numbers; furthermore, there are excessive zero counts.¹⁵ Typically, the application of a prevalence percentage filter, the use of log-transformations, applying a staying-in-the-simplex approach, or using ratios calculations are the normal approaches to solving the above problems.¹⁶ In spite of technological advances and the use of broad metadata collections in published studies, further analytical refinement, together with improved study design and increased sample size, are warranted in order to facilitate standardization of the methods used and the translation of study findings into useful clinical findings.

Machine learning (ML) has emerged as one of the artificial intelligence (AI) methodologies that can be used during microbiome research. ML provides an innovative and integrative method that can assist traditional analysis procedures in order

to improve prediction. Rigorous feature selection and extraction procedures are able to help overcome the dimensional complexity. The use of different ML algorithms during the analysis of the microbiome composition can also enhance microbial biomarker classification, phenotype prediction, possible host interactions and potential endogenous component interactions.^{12,15}

In this study, we adopt an integrative approach involving the use of ML together with differential abundance methods in order to investigate the composition and diversity of gut microbiota by analyzing full-length 16S rRNA gene sequencing. The fecal samples came from a relatively large cohort of diabetic patients with diverse levels of renal function, as well as from controls subjects with normal renal function. In addition, we incorporate metabolite datasets from blood that targeted lipidomic profiles in order to get insights into the possible interplay between the host metabolism and the host gut microbes. Moreover, the functional capabilities of various metabolic pathways that are potentially associated with DKD-prone microbial metabolism were explored in order to provide a foundation for translating these research findings into personalized therapeutic strategies.

Materials and methods

Patient setting and clinical samples

We enrolled participants who were members of the Northeastern Taiwan Community Medicine Research Cohort (ClinicalTrials.gov: NCT04839796) into this study.^{17–19} Specifically, this study recruited community inhabitants aged greater than 30 years old between August 2013 and November 2022. Subjects who were pregnant, who were undergoing dialysis therapy, and/or who had undergone renal transplantation, were excluded. All participants gave signed written informed consent. This study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the Institutional Review Board of Chang Gung Medical Foundation (IRB No: 201800802B0, 202000077B0A3, 201800273B0C602, 202002535B0).

Disease definitions

After enrollment, the participants were classified into various different disease groups according to the following clinical definitions.^{17–19} Type 2 DM was defined as a fasting glucose of ≥ 126 mg/dL, a glycosylated hemoglobin ≥ 6.5 , or the use of hypoglycemic medication. CKD was defined using the National Kidney Foundation: Kidney Disease Outcomes Quality Initiative classification. These subjects had persistent proteinuria or an eGFR of less than 60 mL/min/1.73 m², as determined by the abbreviated Modification of Diet in Renal Disease equation.²⁰ Proteinuria was present if the protein to creatinine ratio was ≥ 150 mg/g or the urine albumin to creatinine ratio was ≥ 30 mg/g. DKD was diagnosed if subjects fulfilled both the DM criteria and the CKD criteria at the same time. Patients were not obese, did not have DM, CKD, DKD, or an acute illness were classified into the control group. Examination of a subject kidney histology by biopsy, in order to ascertain a clinical phenotype, was not performed; this was because this study involves research on a large population and such a biopsy-based approach would be highly invasive. Moreover, clinical indices that suggest superimposed glomerulonephritis, such as hematuria, red cell cast or nephrotic range proteinuria, were minimal among our subjects.

Stool DNA extraction and full-length 16S rRNA gene sequencing

Initially, 10 g of stool sample was harvested, placed in nucleic acid preservative (80% EtOH). The samples were then delivered to our research center within 24–48 h after collection. They were stored at -80°C until fecal DNA extraction took place. Fecal DNA was isolated using the QIAamp Fast DNA Stool Mini Kit (QIAGEN, Germany). Sequencing libraries were prepared by amplifying the full-length of the 16S rRNA genes using KAPA HiFi HotStart ReadyMix (Roche), which includes the 27F (AGRGTTYGATYMTGGCTCAG) and 1492 R (RGYTACCTTGTTACGACTT) barcoded primers. The following thermocycler conditions, 95°C for 3 min; 25 cycles of 95°C for 30 sec, 57°C for 30 sec, 72°C for 60 sec; and 72°C for 5 min, were used, and this was followed by the amplified

samples being stored at 4°C . Next, the amplicons were purified using AMPure PB beads and pooled in an equimolar manner. Libraries were constructed using the SMRTbell Express template Prep Kit 2.0 following the manufacturer's protocol. Sequencing was performed using the circular consensus sequence (CCS) mode on a PacBio Sequel IIe system; this generated HiFi reads with a predicted accuracy greater than Q30.

Processing and analysis of sequence data

Amplicon sequence variants (ASVs) were inferred using the DADA2 R package (1.18.0),²¹ and their taxonomic affiliation was assigned to the ASV with the help of BLAST and our unified 16S rRNA reference database (U16S) (<https://github.com/mammerlin/U16S-DPOT/tree/main/Curated%20DB>). An ASV count table was first aggregated into a taxonomy count table (1121 taxa), and then the taxonomic count table was rarefied for each sample to 7680, resulting in 1082 taxa across 990 samples. Taxa were removed if they were found in fewer than 10% of samples (a 10% prevalence filter),¹⁶ which resulted in 220 taxa. Downstream analyses were performed using R (version 4.2.1) and various Bioconductor packages including ggplot2,²² microViz,²³ phyloseq,²⁴ and vegan.²⁵ We determined α -diversity and β -diversity by means of Chao1 and PCoA with Bray-Curtis matrices, respectively. Differential abundance methods, including DESeq2,²⁶ LefSe,²⁷ limma voom²⁸ and MaAsLin2,²⁹ were used for discriminatory microbiota identification.

Measurement of lipophilic metabolite profiles

The comprehensive methodology used in our study has been described previously.^{17,18} Briefly, a commercially available kit (AbsoluteIDQ p180—BIOCRATES Life Sciences AG, Austria) was used to identify a total of 147 metabolites from five compound classes (15 acylcarnitines, 21 amino acids, 9 biogenic amines, 88 glycerophospholipids and 14 sphingolipids); this was done using ultra-high performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS). The analysis was performed in positive electrospray ionization mode using a Waters tandem mass spectrometer (TQS,

Waters MS Technologies, Manchester, UK). Chromatographic separation was performed on an Acquity BEH C8 column (75 mm × 2.1 mm, particle size of 1.7 µm; CWaters Crop., Milford, USA) at 50°C using a linear gradient that ranged from 0.2% formic acid in water to 0.2% formic acid in acetonitrile at a flow rate of 0.9 mL/min. All data was processed and analyzed using MetIQ software (Biocrates Life Science AG, Innsbruck, Austria). Metabolites with >10% missing values, as well as values below the limit of detection (LOD), were excluded from the analysis.

ML methods for predicting disease groups

We used gut microbiota biomarkers to carry out prediction on four groups, these were DM *vs.* Control, DKD *vs.* DM, DKD *vs.* CKD, and CKD *vs.* Control. We considered not only the contribution of single members of the gut microbiota but also the relationship between pairs within the microbiota clusters, as part of our models. For instance, at the genus level, we calculated the log ratio of *Anaerobutyricum* and *Marseillibacter* by dividing the abundance of *Anaerobutyricum* by that of *Marseillibacter*, followed by a log transformation to derive the “*Anaerobutyricum*_*Marseillibacter*_Genus_ratio”. The training process was as follows: the input dataset was divided into a training dataset and a test dataset with a split ratio of 80% and 20% for a 100-times bootstrap. The randomly selected samples in each bootstrap were then used to obtain the ranking of features using three ML methods, namely Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF) and Support Vector Machine (SVM). The model-building approaches used Logistic Regression, Random Forest and Extreme Gradient Boosting. The above methods were carried out using R (version 3.6.3) with the packages glmnet (4.1.4), random forest (4.6.14), xgboost (0.90.0.2) and e1071 (1.7.3).

Statistics analysis

Descriptive statistics are expressed as the mean, median or frequency. Normality of numerical variables was tested using the Kolmogorov–Simirnov method. Differences in clinical indices among groups were determined using Student’s t-test,

ANOVA or Kruskal–Wallis test. Overlapping significant taxa were determined from differential abundance and machine learning results. Taxonomy-informed functional predictions were obtained using PICRUSt2³⁰ by placing ASVs into a reference tree. The metabolic pathway enrichment analysis and functional prediction were performed using the MetaCyc database (<https://metacyc.org/>).³¹ Differentially abundant MetaCyc pathways of an overlapping taxon for a specified comparison were identified using Wilcoxon tests. Spearman correlations were used to determine the association of a differential taxon with a clinical index, and *p* values were adjusted using the Benjamini–Hochberg correction. All statistical tests are two-tailed, and a *p* < 0.05 is considered statistically significant. Data were analyzed using R (version 4.2.1).

Results

Clinical characteristics

The present study enrolled 990 subjects. Of these subjects, 455 subjects (45.9%) were controls, 204 subjects (20.6%) had type 2 DM, 182 subjects (18.4%) had DKD, and 149 subjects (15.1%) had non-diabetic CKD. Table 1 summarizes the baseline characteristics of entire cohort. The mean age of the study population was 62 years and 48.4% of patients were men. The median serum creatinine was 0.8 mg/dL. The median estimated glomerular filtration rate (eGFR) was 84 mL/min/1.73 m². The median urine albumin/creatinine ratio was 8.6 mg/g. The median fasting sugar was 101 mg/dL. The median glycosylated hemoglobin was 6%. DKD patients had lower eGFR levels and were older in age, while a greater proportion of the male patients suffered from metabolic syndrome and proteinuria than other groups.

Microbial composition and diversity in different disease groups

A total of 23,157,213 sequencing reads, ranging from 6190 to 74,715 per sample, was generated. The sequencing reads were resolved to 86,696 ASVs using DADA2, and then aggregated to 14

Table 1. Baseline characteristics of the study population.

Parameters	All N = 990	Normal control N = 455	Diabetes N = 204	DKD N = 182	CKD N = 149	p-value
Age, years	62.0 ± 11.7	58.3 ± 12.5	62.5 ± 9.4	67.3 ± 8.9	66.5 ± 10.9	<0.001
Male, No. (%)	479 (48.4%)	199 (43.7%)	107 (52.5%)	116 (63.7%)	57 (38.3%)	<0.001
Comorbidities						
Metabolic Syndrome, No. (%)	476 (48.1%)	145 (31.9%)	140 (68.6%)	129 (70.9%)	62 (41.6%)	<0.001
Proteinuria, No. (%)	253 (25.6%)	0	0	145 (79.7%)	108 (72.5%)	<0.001
Anthropometrics						
Body mass index, kg/m ²	26.8 ± 4.5	26.5 ± 4.5	27.6 ± 4.8	27.3 ± 4.4	25.8 ± 3.9	0.319
Systolic BP, mmHg	134.6 ± 37.9	130.1 ± 17.8	139.6 ± 74.7	136.9 ± 18.5	138.8 ± 20.6	0.011
Diastolic BP, mmHg	78.2 ± 11.7	77.4 ± 11.5	78.5 ± 10.3	77.6 ± 11.7	80.8 ± 13.5	0.190
Laboratory						
eGFR, mL/min per 1.73 m ²	84 (12, 243)	89 (61, 160)	87.0 (61, 162)	60 (12, 129)	69 (16, 243)	<0.001
Serum creatinine, mg/dL	0.8 (0.3, 3.7)	0.8 (0.4, 1.2)	0.8 (0.4, 1.2)	1.1 (0.5, 3.7)	0.9 (0.3, 3.6)	<0.001
Cholesterol, mg/dL	182 (20, 418)	194 (20, 300)	168 (102, 270)	168 (76, 326)	186 (105, 418)	<0.001
Triglycerides, mg/dL	116 (25, 1418)	106 (26, 1418)	129.5 (28, 985)	131 (29, 430)	108 (25, 660)	0.029
Urine albumin/creatinine ratio, mg/g	8.6 (0.9, 6611.3)	5.1 (1.3, 29.8)	7.3 (0.9, 29.6)	52.4 (1.5, 6611.3)	47.8 (1.6, 1445.2)	<0.001
LDL-C/HDL-C, mg/dL	2.2 (0.6, 25)	2.3 (0.7, 25)	2.0 (0.8, 6.0)	2.0 (0.6, 5.1)	2.2 (0.7, 5.7)	0.006
Glycated Hemoglobin, %	6.0 (3.1, 13.7)	5.7 (3.1, 6.4)	6.8 (5.2, 13.0)	6.8 (5.1, 13.7)	5.8 (4.6, 6.4)	<0.001
Glucose, mg/dL	101 (67, 393)	95 (67, 124)	121 (74, 317)	130 (79, 393)	97.0 (77, 124)	<0.001

The values are expressed as means (SD) or median (Min, Max) or n (%).

Abbreviations: DM, diabetes mellitus; DKD, diabetic kidney disease; CKD, chronic kidney disease; BP, blood pressure; eGFR, estimated glomerular filtration rate; LDL-C, Low-density lipoprotein cholesterol; HDL-C, High-density lipoprotein cholesterol.

phyla, 27 classes, 54 orders, 111 families, 336 genera and 1121 species. In addition to the 1121 × 990 count table used in ML methods, the taxonomic count table was rarefied and prevalence-filtered to give a 220 × 990 table. This table was then used for the downstream microbiome analyses. Although subtle differences among groups regarding the taxonomic distributions of the top 30 genera (Supplementary Figure S1a) and species (Supplementary Figure S1b) were observed, significant differences in bacterial species richness (α -diversity, Chao1) were also obtained when comparing the Control vs. DM group and the Control vs. DKD group (Supplementary Figure S1c). Furthermore, an analysis of sample-to-sample dissimilarities in bacterial community structures (β -diversity) revealed that the gut microbiomes present in the DM and DKD groups were highly distinct from that of the control group (Supplementary Figure S1d). A significant difference in the Firmicutes to Bacteroidetes (F/B) ratio was detected between DM vs. Control (1.23 vs. 1.1; $p = 0.033$) and CKD vs. Control (1.27 vs. 1.1; $p = 0.0017$) (Supplementary Figure S1e). Specifically, the relative abundance of the Bacteroidetes phylum in the male population was found to be decreased in DM and DKD patients compared to the control subjects (Supplementary Figure S1f).

Identification of gut microbiota signatures in order to differentiate DKD from DM and CKD

A total of 14 DKD-associated bacterial biomarkers at the genus level were identified when the DKD groups were compared with the other groups using the linear discriminant analysis of the effect size (LEfSe) method (Figure 1). Interestingly, two genera, namely *Haemophilus* and *Acidaminococcus*, were enriched in DKD group compared with the Control or CKD groups. Conversely, the genus *Gemmiger* was significantly decreased in the DKD group compared with the Control and DM groups (Figure 1(a)). Using the full-length 16S rRNA datasets, *Haemophilus* and *Acidaminococcus* were further characterized at the species-level resolution to be *Haemophilus parainfluenzae* and *Gemmiger formicilis*. (Figure 1(b)). In addition to LEfSe method, three commonly used differential abundance methods, namely limma, MaAsLin2 and DESeq2, were applied to identify any potential microbial biomarkers that are associated with DKD. The results revealed that there is a discrepancy in both number and content when discriminatory microbes are identified by the various differential abundance methods (Supplementary Figure S2). Interestingly, in the comparison between the DKD vs. DM groups, the genus *Gemmiger* was identified as a potential biomarker both by the LEfSe method and the ML method (described below). Moreover, in the

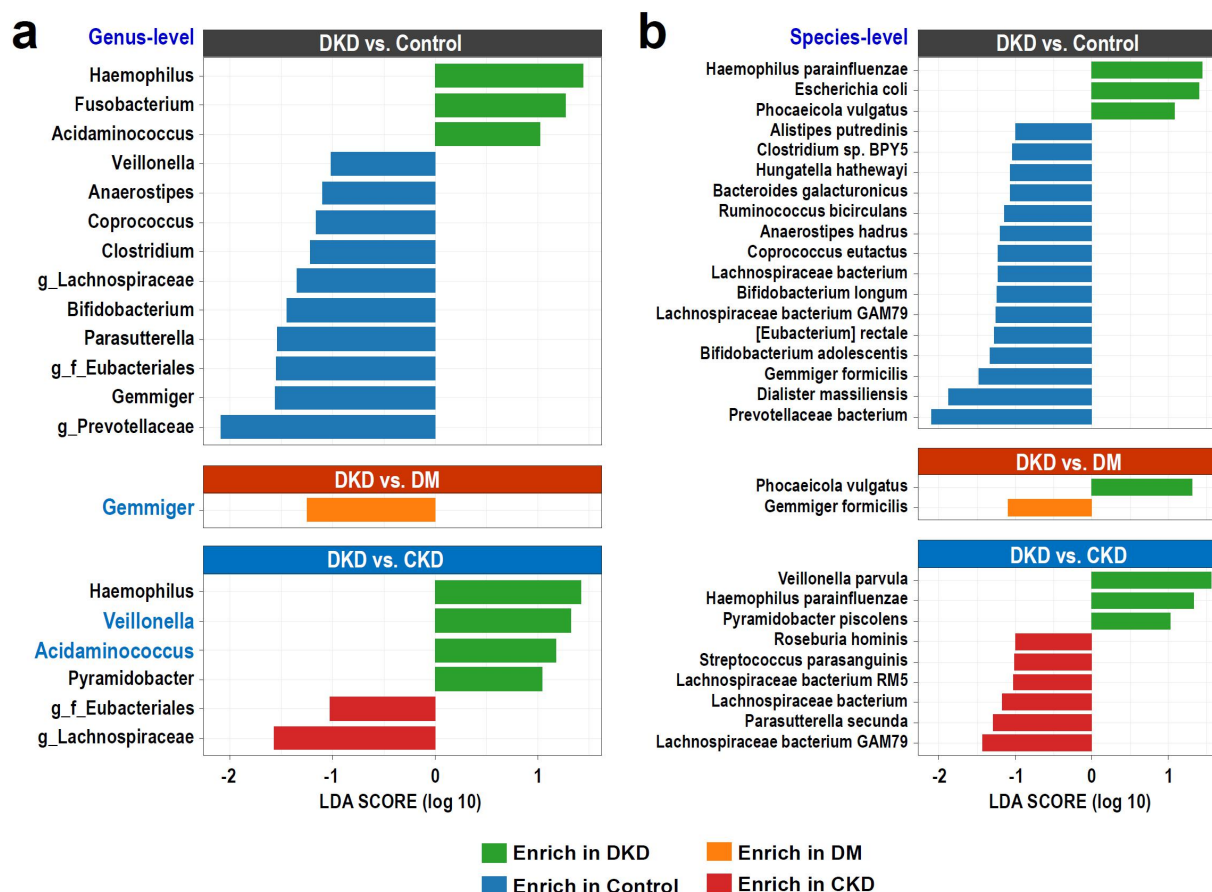


Figure 1. Determination of the bacterial biomarkers enriched in the DKD, DM, CKD and Control groups at genus-level (a) and species-level (b). The results were obtained using the linear discriminant analysis of effect size (LEfSe) method. In the comparison of DKD vs. DM, the genus *Gemmiger* (marked with blue) was concomitantly selected by both the LEfSe and ML methods. In the comparison of DKD vs. CKD, two genera, namely *Veillonella* and *Acidaminococcus* (marked with blue), were selected by both the LEfSe method and the ML methods.

comparison between the DKD vs. CKD groups, two genera, namely *Acidaminococcus* and *Veillonella*, were identified by both the LEfSe method and the ML method. Finally, *Romboutsia* was selected by three methods, the limma, MaAsLin2 and ML.

ML creates new types of microbiota-based biomarkers that are able to distinguish DKD, DM and CKD patients

The data complexity allowed us to carry out a second analysis, namely to characterize the informative features using the ML methods and identifying the top features that are able to distinguish DKD patients, DM patients, CKD patients and Control subjects (Figure 2(a); Feature Selection). Intriguingly, in addition to the clinical characteristics (Table 1), new types of microbiota biomarkers were created by the ML methods, including

relative abundance (RA) of a microbe, the presence or absence (PA) of a microbe, and a hierarchy ratio, namely the ratio between two different genera, families, orders or classes. The concept of the hierarchical ratio was inspired by the widely recognized Firmicutes/Bacteroidetes ratio, which reflects the relative abundance of two prominent and representative phyla in the human gut microbiota. Building on this foundation, and utilizing advanced big data analytical approaches, we aimed to uncover novel targets by calculating hierarchical ratio across all taxonomic levels, namely from phylum to genus (Supplementary Figure S3). The optimal number of top features, that is the minimal number of features that are able to reach the best performance, was determined by Area Under Curve (AUC) method and by accuracy rate in the elbow plots (Figure 2(b)). Four different panels of features were selected by ML in order to

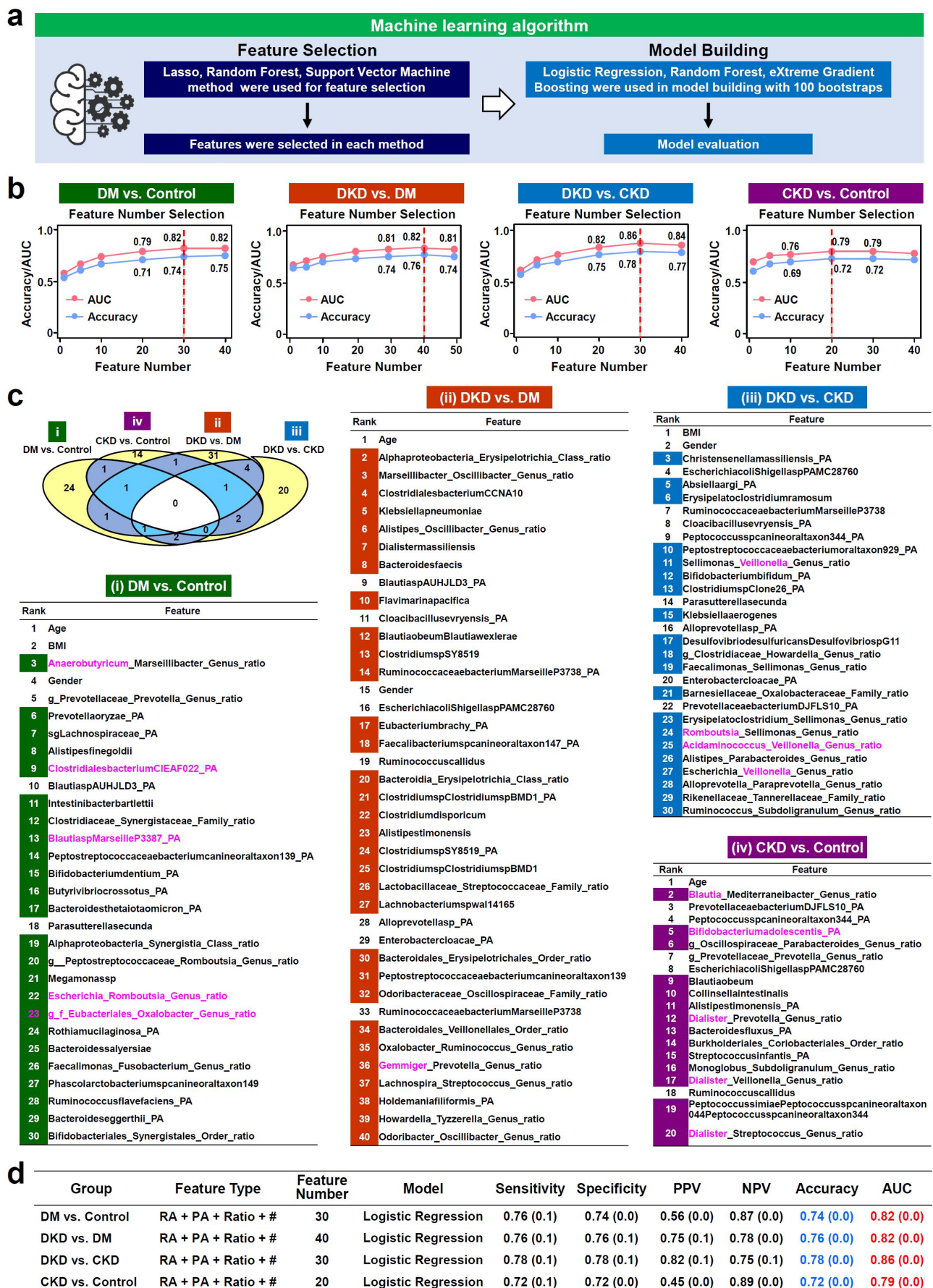


Figure 2. Feature selection and model prediction of ML methods in order to differentiate DKD, DM and CKD patients. (a) Workflow of the ML algorithm including feature selection and model building. (b) The number of features selected was determined by AUC and accuracy for the various disease group comparisons. (c) The four panels of top features selected by ML. A venn diagram is used to

differentiate between specific group pairings, namely, DM vs. Control, DKD vs. DM, DKD vs. CKD, and CKD vs. Control. Notably, in each panel, most of the features selected by ML are unique to that particular panel (Figure 2(c)).

To achieve model derivation and validation, the subjects were randomly split into training and testing datasets with a ratio of 80% to 20% and subjected to 100-times bootstrap (Figure 2(a)); Model Building). The performance was evaluated using the AUC and accuracy rate. Our analysis revealed that Logistic Regression gave the best performance (Supplementary Table S1). Using Logistic Regression as the model, we obtained the following results: (i) in the DM vs. Control comparison, the top 30 features yield an accuracy rate of 0.74 and an AUC of 0.82 when differentiating between DM patients and Control subjects; (ii) in the DKD vs. DM comparison, the top 40 features yield an accuracy rate of 0.76 and an AUC of 0.82 when differentiating between DKD patients and DM patients; (iii) in the DKD vs. CKD comparison, the top 30 features yield an accuracy rate of 0.78 and an AUC of 0.86 when differentiating between DKD patients and CKD patients; and (iv) in the CKD vs. Control comparison, the top 20 features yield an accuracy rate of 0.72 and an AUC of 0.79 when differentiating between CKD patients and Control subjects (Supplementary Table S2 ; Figure 2(d)).

The correlation between microbiota and clinical indexes suggests that *Gemmiger* may have a reno-protective effect when DM is diagnosed

Interestingly, when we examined the comparison between the discriminatory microbes identified by the four methods of differential abundance (Supplementary Figure S2) and by the ML-selected biomarkers (Figure 2(c)), 13 bacterial biomarkers were consistently identified by both approaches. To explore the potential significance of these bacteria in a clinical situation and to identify their association

with disease progression, we analyzed the correlation between these bacteria and clinical indexes obtained from the subjects using Spearman analysis (Figure 3). Our analyses revealed the following. Firstly, in the DM vs. Control comparison, seven significant differential bacteria were identified (marked with a green background). Among these bacteria, *Romboutsia* and *Clostridiales bacterium CIEAF 022* were negatively correlated, and *Escherichia* was positively correlated with the levels of serum glucose and glycated Hemoglobin (HbA1c). Furthermore, *Romboutsia* and *Blautia sp. Marseille-P3387* were positively correlated, and *Escherichia* was negatively correlated with the levels of serum cholesterol and LDL. Secondly, in the CKD vs. Control comparison, three significant differential bacteria were identified (marked with a purple background). *Dialister* was negatively correlated with age, and positively correlated with weight. *Blautia* was positively correlated with BMI, waist and weight, while negatively correlated with uPCR and uACR. *Bifidobacterium adolescentis* was negatively correlated with age, and positively correlated with BMI, weight and height. Regarding the metabolic indexes, *Bifidobacterium adolescentis* was negatively correlated with the levels of serum glucose and HbA1c, and was positively correlated with the levels of serum cholesterol and LDL. Interestingly, regarding various renal indexes, *Bifidobacterium adolescentis* was positively correlated with eGFR and negatively correlated with uPCR and uACR, which suggests that this bacterium has a beneficial influence on kidney function. Thirdly, in the DKD vs. CKD comparison, two significant differential bacteria were identified (marked with a blue background). *Veillonella* was positively correlated with age, and negatively correlated with BMI, waist, weight and height. *Acidaminococcus* was positively correlated with serum triglyceride level. Finally, in the DKD vs. DM comparison, remarkably, *Gemmiger* (marked with a red background) appears to be specifically identified as a microbiota biomarker when DKD patients are compared with DM patients. Furthermore,

present the number of unique and overlapped features among the four panels of top features. The list and rank of features for distinguishing (i) DM vs. Control; (ii) DKD vs. DM; (iii) DKD vs. CKD; (iv) CKD vs. Control. The unique features are highlighted with a background color. Feature types include clinical variables (age, BMI and gender), and features of the microbiota, namely RA (relative abundance), PA (presence or absence, and ratio (hierarchy ratio)). Those microbes concomitantly selected by the ML method and at least one of the differential abundance methods (DESeq2, LefSe, limma voom or MaAsLin2) are marked with pink. (d) Prediction performance of the gut microbiota present in different disease groups using ML algorithms with bootstrapping 100 times. Feature type: # indicates age, BMI and gender. Data are presented as means (standard deviation).

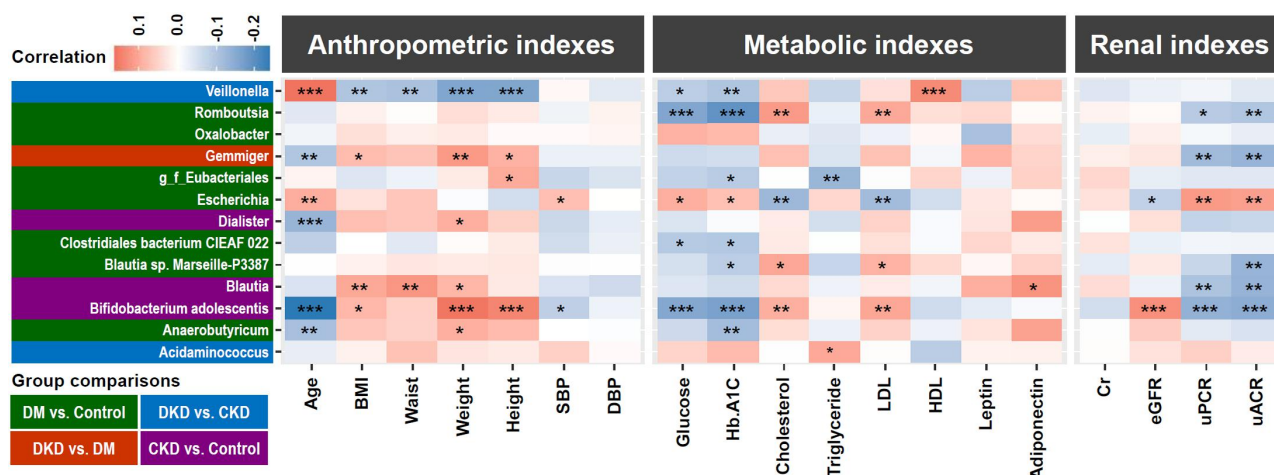


Figure 3. Correlation between bacterial biomarkers and clinical indexes using Spearman analysis. Thirteen gut bacterial biomarkers were concomitantly identified by the differential abundance method and ML algorithm across diverse group comparisons. The combinations of group comparisons are highlighted using a background color. * p value < 0.05, ** p value < 0.01, *** p value < 0.001.

Gemmiger was negatively correlated with age, and positively correlated with BMI, weight and height. Notably, *Gemmiger* was negatively correlated with uPCR and uACR, which suggests that it may have a beneficial effect and might be associated with reno-protection when DM is present.

Metabolic pathways associated with *Gemmiger*

Consistent with our results using linear discriminant analysis (Figure 1), the relative abundance of *Gemmiger* was significantly decreased in the DKD group (median 0.358%) compared with the DM group (median 0.568%, p value = 0.028) and the Control group (median 0.779%, p value = 0.00024) (Figure 4(a)). Since gut microbiota-derived metabolites may be associated with the pathophysiology and progression of DKD,³² we performed a metabolic pathway enrichment analysis and functional prediction using the MetaCyc database to decipher the potential role of *Gemmiger* in DKD patients. A total of 25 *Gemmiger*-associated metabolic pathways exhibited significant differences between the DKD and DM groups (Wilcoxon test p < 0.05). These metabolic pathways can be grouped into three major categories: (1) essential amino acids, (2) carbohydrate, and (3) ribonucleotide and nucleic acid (Figure 4(b)). Intriguingly, the biosynthesis of the branched-chain amino acids (BCAAs), namely L-valine, L-leucine and

L-isoleucine, was found to be the main pathway for the metabolism of essential amino acids. With regard to the metabolism of carbohydrate, a variety of pathways related to carbohydrate degradation (sucrose, galactose and starch degradation), glycogen biosynthesis, pyruvate fermentation and glycolysis were identified to be associated with *Gemmiger* (Figure 4(b)). Since *Gemmiger* was identified by both the LEfSe and ML methods during the DKD vs. DM comparison, these results suggest that an alteration involving these interconnected metabolic pathways may potentially contribute to the pathogenesis of DKD (Figure 4(c)).

Discussion

In this study, by thoroughly analyzing large fecal samples from subjects with DM, DKD, CKD and control groups, three findings can be pinpointed. **Firstly**, new types of microbiota biomarkers have been created by ML, namely the relative abundance (RA) of microbe, the presence or absence (PA) of a microbe, and the hierarchy ratio between two different taxonomies. Four different panels of features were selected to differentiate (i) DM vs. Control, (ii) DKD vs. DM, (iii) DKD vs. CKD, and (iv) CKD vs. Control, all with accuracy rates between 0.72 and 0.78 and all with areas under curve of between 0.79 and 0.86. **Secondly**, 13 gut microbiota biomarkers that

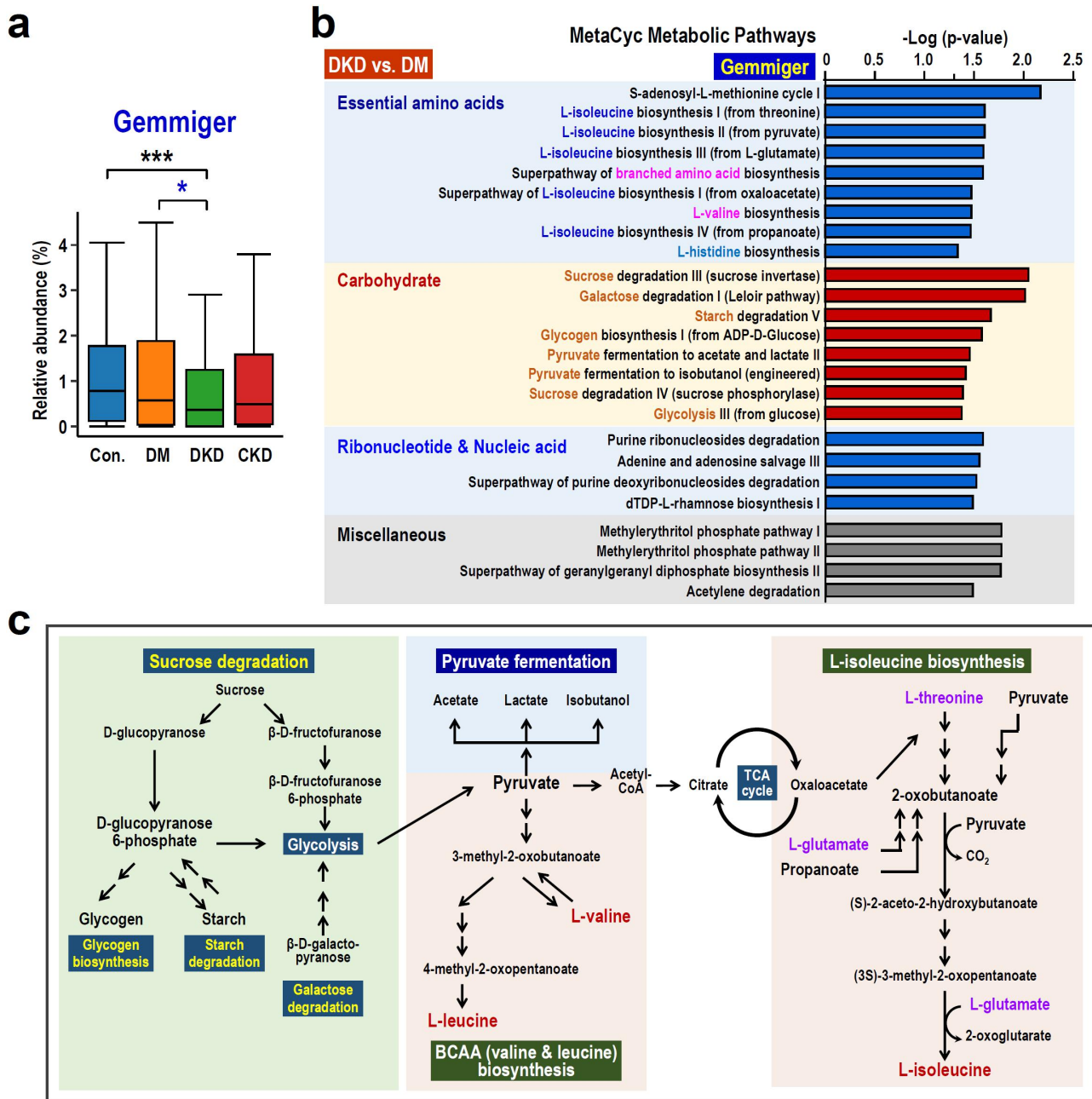


Figure 4. Pathway enrichment analysis and functional analysis of *Gemmiger*. (a) The relative abundance of *Gemmiger* in the control, DM, DKD and CKD groups. (b) MetaCycCyc metabolic pathway enrichment analysis of *Gemmiger*-associated pathways. (c) A graphic summary of the *Gemmiger*-related metabolic pathways. The interconnected metabolic pathways include the metabolism of several types of carbohydrates, pyruvate fermentation and BACC biosynthesis. The pathway annotation was carried out using MetaCyc metabolic pathway database (<https://metacyc.org>). Abbreviations: TCA, tricarboxylic acid cycle. The box plots display median abundances and the interquartile range (IQR) multiplied by 1.5. Significance between groups was assessed using the Wilcoxon rank-sum test. The asterisks denote significance levels: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

were concomitantly identified by the ML algorithm and the differential abundance method were found to be highly discriminatory across the various different group comparisons. Furthermore, these microbiota biomarkers were strongly correlated with a range of

anthropometric, metabolic and renal indices. **Finally**, the predicted functional capability of a DKD-specific biomarker, *Gemmiger*, was found to be enriched in various forms of carbohydrate metabolism and BCAA biosynthesis. Coincidentally, the circulating

levels of BCAAs (L-valine, L-leucine and L-isoleucine) and their precursor, L-glutamate, are significantly increased in DM and DKD patients, which suggests that there are alterations in the interconnected pathways of glycolysis, pyruvate fermentation and BCAA biosynthesis in the presence of hyperglycemia. Overall, our findings demonstrate links within the gut-microbiota-kidney axis when there is pathogenic renal impairment in DM patients. Furthermore, our findings highlight that specific gut bacteria are useful biomarkers and that these biomarkers have mechanistic and diagnostic implications.

Gut microbial biomarkers discovered in this and previous studies

Previously using 16S rRNA microbial profiling of fecal samples of biopsy-proven DKD patients, Tao et al. found that *g_Escherichia-Shigella* and genus *g_Prevotella_9* were highly discriminatory when separating DKD patients from DM patients.³³ Similarly, our study revealed a significantly higher abundance of *Escherichia coli* in the DKD group for the following comparisons: DKD vs. Control (Figure 1(b)), DKD vs. DM, and DKD vs. CKD (Figure 2(c); Supplementary Figure S2b). Moreover, our correlation analysis showed a positive association between the abundance of the *Escherichia* genus and the age, glucose and proteinuria of the subjects; conversely, there is a negative correlation between the abundance of the *Escherichia* genus and the lipid profile and eGFR of the subjects (Figure 3). Another study that integrated 16S pyrosequencing and metabolomic analyses identified 11 significantly different intestinal flora and 239 significantly different metabolites when a late-stage predialysis DKD group was compared with a dialysis DKD group. Furthermore, a number of gene functions associated with the phenylalanine and tryptophan metabolic pathways, together with altered levels of hippuric acid, indole-3-acetic acid, L-tryptophan and L-valine, were found to be most highly associated with DKD progression.³⁴ In addition, a metagenomic study involving a limited group of DKD patients found subtle differences in species and

functional pathways between non-DKD and DKD patients. Nine pathways related to *Lactobacillus crispatus* were differentially enriched in DKD patients and these were related to the sucrose degradation pathway and L-Lysine biosynthesis.⁶ Another metagenomic analysis has shown that the relative abundances of six bacterial species were elevated in the DKD patients. The gene functions of these bacteria were correlated with the BCAAs and methionine metabolic biosynthesis pathways.³⁵

It should be noted that there are inconsistencies regarding the DKD-specific microbes between previous reports and this study. This may be attributable to a number of possible factors. Firstly, there may be discrepancies in the severity of kidney disease, analytical methodology, geography and dietary patterns across different studies; all of these may contribute to variation in phenotypic outcomes. Secondly, a relatively small sample size of patient settings was used in many of the previously published studies. In the present study, our findings are based on full-length sequencing of the 16S rRNA gene using feces obtained from a relatively large study cohort that contains subjects with four distinct clinical phenotypes. Finally, our results are derived from the integrated analyses of traditional abundance methodologies (DESeq2, LEfSe, limma voom and MaAsLin2) together with the use of the ML approach. Our aim was to provide a possible biological explanation for the potential roles of microbiota identified in terms of the metabolic pathways and the crosstalk between gut microbiota and host metabolic disturbance in DKD.

Clinical implication of a decreased abundance of *Gemmiger* in the feces of DKD patients

The decrease in *Gemmiger* observed in our DKD patients is consistent with previous reports obtained from biopsy-proven diabetic nephropathy patients.³⁶ The strong association of the abundance of *Gemmiger* with various anthropometric measurements, such as body mass index and weight, is consistent with previous literature. In these studies, *Gemmiger* was found to be associated with sex-specific adipose distribution.³⁷ Furthermore, the relative abundance of *Gemmiger* was reduced when autoimmune disease is present³⁸ and similar findings were found for

women with rheumatoid arthritis.³⁹ *Gemmiger formicilis* is a chemoorganotrophic gram-negative bacterium that ferments sugars to produce formic, butyric and lactic acids, all of which exert anti-inflammatory effects.⁴⁰ Notably, *Gemmiger* utilizes the acetyl-CoA (ACoA) pathway to synthesize butyrate, a short-chain fatty acid.⁴¹ Sodium butyrate, a derivative of butyrate, has been identified as a potential therapeutic agent for the treatment of diabetic nephropathy⁴²; this supports the notion that *Gemmiger* may play a reno-protective role. On the other hand, *Gemmiger formicilis* has been positively correlated with both phenylacetylcarnitine and phenylacetylglutamine, which are derived from the gut fermentation of phenylalanine and acetylcholine. These metabolites have been correlated with microbial gene richness in obese subjects and have implications related to mitochondrial dysfunction and hepatocyte lipid accumulation in fatty liver disease.⁴³ In this cross-sectional study, we observed a significant association with a trend toward a decrease in the abundance of *Gemmiger* in DKD patients (Figure 4(a)) and from this we have inferred its potential reno-protective effects based on metabolic pathway enrichment analysis (Figure 4(c)). However, direct causal evidence linking *Gemmiger* to DKD is currently lacking. Therefore, further validation using animal and cell platforms is needed to establish a definitive relationship.

Limitations and future perspectives

To define the potential functionality of gut microbiome associated with DKD, additional efforts are needed to address several limitations of the present study. **Firstly**, a replication cohort is needed, although many lines of our microbiota and metabolomic results are consistent with previous literature. Nevertheless, we acknowledge that the findings obtained from the cohort recruited from a specific community in northeast Taiwan may not be generalized to other populations. **Secondly**, we applied full-length sequencing of the 16S rRNA gene and predicted the potential function of the microbiota by inference from PICRUSt2 analysis.³⁰ This approach is a cost-effective way for a large sample research. However, in-depth metagenome information with a resolution at the species or strain level still

remains elusive. Therefore, further experiments using metagenomic sequencing or fecal metabolomics are warranted to elucidate the direct interaction between microbiota and host metabolism. **Thirdly**, our findings represent real-world population data in which the imbalances of disease distribution (especially in the DM and DKD groups) and baseline information are inherently related to the characteristics of the status of the various diseases (such as an older age and greater metabolic dysregulation among the DKD and CKD patients). It should be noted that the subgroup analyses of the relative abundance of most of the 14 microbial biomarkers were non-differential in terms of age, eGFR and glycosylated hemoglobin (Supplementary Table S3). To address these issues and to reduce sampling bias, we applied the bootstrap method and performed 100 resampling iterations. Each iteration involved random sampling, which helps to minimize sampling bias by averaging out variations across resamples. This approach allowed us to enhance the robustness and reliability of our results despite the limited sample size. We believe that this method effectively reduces the constraints of small sample sizes and improves the applicability of our study conclusions. **Fourthly**, with regard to dietary recall, medication usage and other unknown confounding factors, these information types were not collected by the community project that provided the dataset and thus may represent a potential drawback. However, all the participants came from similar geographic and cultural areas and this should have minimized pronounced variations in dietary patterns and other habits among individuals. Moreover, a dietary composition investigation in our previous microbiota study that recruited participants from the same geographic area has revealed that the daily serving portions (in terms of vegetable, meat, fruit and rice/noodle) had no overt difference between non-CKD vs. CKD patients.⁹ Additionally, subjects using antibiotics, prebiotics and probiotics were excluded from enrollment. Participants with an acute infection, including viral infection, were also excluded from recruitment in order to minimize as much as possible any distortion of host microbiota. **Finally**, several

studies have suggested that the DNA extraction reagent we used (QIAamp Fast DNA Stool Mini Kit) can extract fungal DNA,⁴⁴ and that subsequent internal transcribed spacer region sequencing can provide information on fungal species and their relative abundance. However, fungi and viruses are only sporadically abundant in the human gut microbiome compared to bacteria.⁴⁵ Our current methodology remains unable to extract information on intestinal viruses, fungi, bacteriophages and parasites; thus, such evaluations may be beyond the scope of this research. Overall, comprehensive full-length pyrosequencing datasets and ML-assisted analyses should help with the data-driven discovery of novel species–host interaction that will increase our understanding of DKD pathogenesis. Thus, further *in-vivo* experiments targeting different components of the gut microbiota are required to demonstrate the causality of our findings in the future.

Conclusions

Advances in our understanding of the molecular mechanisms and pathways involved in the pathogenesis of DKD remain an unfulfilled need, and the identification of biomarkers that can be used for early diagnosis or prognosis prediction will help our understanding of these disease states. The present study shows that the altered gut microbiota present in DKD patients is functionally associated with specific and distinct carbohydrate and BCAA metabolism pathways. Our study implies that specific gut microbes can be used as a potential biomarker during DKD diagnosis and may also serve as candidate therapeutic targets when gut-microbiota-kidney intervention in DM patients is carried out. Further clinical trials are warranted in order to investigate the effects of manipulating these specific microbes and this will allow the exploration of the changes that then occur in circulating BCAA levels. These findings can then be linked to the outcomes of DKD patients who are undergoing such therapy.

Acknowledgments

We acknowledge the participant recruitment and sample preservation carried out as part of the Northeastern Taiwan

Community Medicine Research Cohort Study. We thank National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) of Taiwan for providing computational resources.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded by grants from the Ministry of Health and Welfare (Smart Healthcare for Obesity Therapeutics, PD-109-GP-02, MG-110-GP-03 and MG-111-GP-03 to HKS and MG-112-GP-03 to WHHS; Development of Precision Prevention and Treatment Strategies for Metabolic and Related Chronic Diseases: Prediction and Implementation of an Intelligent Prediction System, MG-113-GP-03 to WHHS), and from Chang Gung Memorial Hospital (CRRPG2H0121-124 to IWW; CORPG3N1481-1482 and CMRPG3K2241-2243 to CHY; CMRPG2K0141-142 to CCL).

Authors' contributions

All the authors contributed to the manuscript preparation. Chi-Chun Lai, Huey-Kang Sytwu and Ting-Fen Tsai co-designed the research framework. I-Wen Wu, Chi-Hsiao Yeh, Chun-Yu Chen, Heng-Chih Pan, Heng-Jung Hsu and Chin-Chan Lee recruited subjects and defined the clinical stages. Yi-Ju Chou contributed to stool DNA extraction and full-length 16S rRNA sequencing. Yu-Chieh Liao and Chieh-Hua Lin contributed to processing and analysis of microbiome datasets. Chi-Jen Lo and Mei-Ling Cheng contributed to metabolite analysis. Zhao-Qing Shen contributed to metabolic pathway analysis. Tsung-Hsien Tsai designed and supervised the ML analysis. Yun-Hsuan Chan and Chih-Wei Tu contributed to ML analysis. Wayne Huey-Herng Sheu provided critical and inspired comments. I-Wen Wu, Yu-Chieh Liao and Chieh-Hua Lin co-designed the manuscript structure, prepared the figures and drafted the manuscript. Ting-Fen Tsai wrote the final version of manuscript. All authors have read and agreed to the published version of the manuscript.

Data availability statement

The data supporting the findings of this study are available from the corresponding authors on reasonable request.

References

1. U.S. Renal Data System. USRDS annual data report: epidemiology of kidney disease in the United States. Bethesda (MD): National Institutes of Health, National

- Institute of Diabetes and Digestive and Kidney Diseases; 2022.
2. DeFronzo RA, Reeves WB, Awad AS. Pathophysiology of diabetic kidney disease: impact of SGLT2 inhibitors. *Nat Rev Nephrol.* 2021;17(5):319–334. doi:10.1038/s41581-021-00393-8.
 3. Fiaccadori E, Cosola C, Sabatino A. Targeting the gut for early diagnosis, prevention, and cure of diabetic kidney disease: Is the phenyl sulfate story another step forward? *Am J Kidney Dis.* 2020;75(1):144–147. doi:10.1053/j.ajkd.2019.07.001.
 4. Cani PD. Microbiota and metabolites in metabolic diseases. *Nat Rev Endocrinol.* 2019;15(2):69–70. doi:10.1038/s41574-018-0143-9.
 5. Kikuchi K, Saigusa D, Kanemitsu Y, Matsumoto Y, Thanai P, Suzuki N, Mise K, Yamaguchi H, Nakamura T, Asaji K, et al. Gut microbiome-derived phenyl sulfate contributes to albuminuria in diabetic kidney disease. *Nat Commun.* 2019;10(1):1835. doi:10.1038/s41467-019-09735-4.
 6. Sato N, Kakuta M, Hasegawa T, Yamaguchi R, Uchino E, Murashita K, Nakaji S, Imoto S, Yanagita M, Okuno Y. Metagenomic profiling of gut microbiome in early chronic kidney disease. *Nephrol Dial Transpl.* 2021;36(9):1675–1684. doi:10.1093/ndt/gfaa122.
 7. Linh HT, Iwata Y, Senda Y, Sakai-Takemori Y, Nakade Y, Oshima M, Nakagawa-Yoneda S, Ogura H, Sato K, Minami T, et al. Intestinal bacterial translocation contributes to diabetic kidney disease. *J Am Soc Nephrol.* 2022;33(6):1105–1119. doi:10.1681/asn.2021060843.
 8. Hsu CK, Su SC, Chang LC, Shao SC, Yang KJ, Chen CY, Chen YT, Wu IW. Effects of low protein diet on modulating gut microbiota in patients with chronic kidney disease: a systematic review and meta-analysis of international studies. *Int J Med Sci.* 2021;18(16):3839–3850. doi:10.7150/ijms.66451.
 9. Wu IW, Lee CC, Hsu HJ, Sun CY, Chen YC, Yang KJ, Yang CW, Chung WH, Lai HC, Chang LC, et al. Compositional and functional adaptations of intestinal microbiota and related metabolites in CKD patients receiving dietary protein restriction. *Nutrients.* 2020;12(9):2799. doi:10.3390/nu12092799.
 10. Hsu CK, Su SC, Chang LC, Yang KJ, Lee CC, Hsu HJ, Chen YT, Sun CY, Wu IW. Oral absorbent AST-120 is associated with compositional and functional adaptations of gut microbiota and modification of serum short and medium-chain fatty acids in advanced CKD patients. *Biomedicines.* 2022;10(9):2234. doi:10.3390/biomedicines10092234.
 11. Wu IW, Lin CY, Chang LC, Lee CC, Chiu CY, Hsu HJ, Sun CY, Chen YC, Kuo YL, Yang CW, et al. Gut microbiota as diagnostic tools for mirroring disease progression and circulating Nephrotoxin levels in chronic kidney disease: discovery and validation study. *Int J Biol Sci.* 2020;16(3):420–434. doi:10.7150/ijbs.37421.
 12. Wu IW, Gao SS, Chou HC, Yang HY, Chang LC, Kuo YL, Dinh MCV, Chung WH, Yang CW, Lai HC, et al. Integrative metagenomic and metabolomic analyses reveal severity-specific signatures of gut microbiota in chronic kidney disease. *Theranostics.* 2020;10(12):5398–5411. doi:10.7150/thno.41725.
 13. Wang X, Yang S, Li S, Zhao L, Hao Y, Qin J, Zhang L, Zhang C, Bian W, Zuo L, et al. Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and rodents. *Gut.* 2020;69(12):2131–2142. doi:10.1136/gutjnl-2019-319766.
 14. Krukowski H, Valkenburg S, Madella AM, Garssen J, van Bergenhenegouwen J, Overbeek SA, Huys GRB, Raes J, Glorieux G. Gut microbiome studies in CKD: opportunities, pitfalls and therapeutic potential. *Nat Rev Nephrol.* 2023;19(2):87–101. doi:10.1038/s41581-022-00647-z.
 15. Hernández Medina R, Kutuzova S, Nielsen KN, Johansen J, Hansen LH, Nielsen M, Rasmussen S. Machine learning and deep learning applications in microbiome research. *ISME Commun.* 2022;2(1):98. doi:10.1038/s43705-022-00182-9.
 16. Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, Jones CMA, Wright RJ, Dhanani AS, Comeau AM, et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nat Commun.* 2022;13(1):342. doi:10.1038/s41467-022-28034-z.
 17. Wu IW, Tsai TH, Lo CJ, Chou YJ, Yeh CH, Cheng ML, Lai CC, Sytwu HK, Tsai TF. Discovery of a biomarker signature that reveals a molecular mechanism underlying diabetic kidney disease via organ cross talk. *Diabetes Care.* 2022;45(6):e102–e104. doi:10.2337/dc22-0145.
 18. Wu IW, Tsai TH, Lo CJ, Chou YJ, Yeh CH, Chan YH, Chen JH, Hsu PWC, Pan HC, Hsu H-J, et al. Discovering a trans-omics biomarker signature that predisposes high risk diabetic patients to diabetic kidney disease. *NPJ Digit Med.* 2022;5(1):166. doi:10.1038/s41746-022-00713-7.
 19. Yang NI, Yeh CH, Tsai TH, Chou YJ, Hsu PWC, Li CH, Chan YH, Kuo LT, Mao CT, Shyu YC, et al. Artificial intelligence-assisted identification of genetic factors predisposing high-risk individuals to asymptomatic Heart failure. *Cells.* 2021;10(9):2430. doi:10.3390/cells10092430.
 20. National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Am J Kidney Dis.* 2002;39(2 Suppl 1):S1–S266.

21. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, Sk M, Dougherty MK. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.* 2019;47(18):e103. doi:10.1093/nar/gkz569.
22. Wickham H. ggplot2: elegant graphics for data analysis. Springer International Publishing; 2016. p. 189–201. doi:10.1007/978-3-319-24277-4_9.
23. Barnett D, Arts I, Penders J. microViz: an R package for microbiome data visualization and statistics. *J Open Source Softw.* 2021;6(63):3201. doi:10.21105/joss.03201.
24. McMurdie PJ, Holmes S, Watson M. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE.* 2013;8(4):e61217. doi:10.1371/journal.pone.0061217.
25. Dixon P. VEGAN, a package of R functions for community ecology. *J Veg Sci.* 2003;14(6):927–930. doi:10.1111/j.1654-1103.2003.tb02228.x.
26. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with DESeq2. *Genome Biol.* 2014;15(12):550. doi:10.1186/s13059-014-0550-8.
27. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. *Genome Biol.* 2011;12(6):R60. doi:10.1186/gb-2011-12-6-r60.
28. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol.* 2014;15(2):R29. doi:10.1186/gb-2014-15-2-r29.
29. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, Tickle TL, Weingart G, Ren B, Schwager EH, et al. Multivariable association discovery in population-scale meta-omics studies. *PLOS Comput Biol.* 2021;17(11):e1009442. doi:10.1371/journal.pcbi.1009442.
30. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol.* 2020;38(6):685–688. doi:10.1038/s41587-020-0548-6.
31. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD, et al. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* 2020;48(D1):D445–D453. doi:10.1093/nar/gkz862.
32. Mao ZH, Gao ZX, Liu DW, Liu ZS, Wu P. Gut microbiota and its metabolites - molecular mechanisms and management strategies in diabetic kidney disease. *Front Immunol.* 2023;14:1124704. doi:10.3389/fimmu.2023.1124704.
33. Tao S, Li L, Li L, Liu Y, Ren Q, Shi M, Liu J, Jiang J, Ma H, Huang Z, et al. Understanding the gut–kidney axis among biopsy-proven diabetic nephropathy, type 2 diabetes mellitus and healthy controls: an analysis of the gut microbiota composition. *Acta Diabetol.* 2019;56(5):581–592. doi:10.1007/s00592-019-01316-7.
34. Zhang Q, Zhang Y, Zeng L, Chen G, Zhang L, Liu M, Sheng H, Hu X, Su J, Zhang D, et al. The role of gut microbiota and microbiota-related serum metabolites in the progression of diabetic kidney disease. *Front Pharmacol.* 2021;12:757508. doi:10.3389/fphar.2021.757508.
35. Kim JE, Nam H, Park JI, Cho H, Lee J, Kim HE, Kim DK, Joo KW, Kim YS, Kim BS, et al. Gut microbial genes and metabolism for methionine and branched-chain amino acids in diabetic Nephropathy. *Microbiol Spectr.* 2023;11(2):e0234422. doi:10.1128/spectrum.02344-22.
36. Lu X, Ma J, Li R. Alterations of gut microbiota in biopsy-proven diabetic nephropathy and a long history of diabetes without kidney damage. *Sci Rep.* 2023;13(1):12150. doi:10.1038/s41598-023-39444-4.
37. Min Y, Ma X, Sankaran K, Ru Y, Chen L, Baiocchi M, Zhu S. Sex-specific association between gut microbiome and fat distribution. *Nat Commun.* 2019;10(1):2408. doi:10.1038/s41467-019-10440-5.
38. Wang T, Sternes PR, Guo XK, Zhao H, Xu C, Xu H. Autoimmune diseases exhibit shared alterations in the gut microbiota. *Rheumatol (Oxford).* 2023;63(3):856–865. doi:10.1093/rheumatology/kead364.
39. Yun H, Wang X, Wei C, Liu Q, Li X, Li N, Zhang G, Cui D, Liu R. Alterations of the intestinal microbiome and metabolome in women with rheumatoid arthritis. *Clin Exp Med.* 2023;23(8):4695–4706. doi:10.1007/s10238-023-01161-7.
40. Salanitro JP, Muirhead PA, Goodman JR. Morphological and physiological characteristics of Gemmiger formicilis isolated from chicken ceca. *Appl Environ Microbiol.* 1976;32(4):623–632. doi:10.1128/aem.32.4.623-632.1976.
41. Kircher B, Woltemate S, Gutzki F, Schlüter D, Geffers R, Bähre H, Vital M. Predicting butyrate- and propionate-forming bacteria of gut microbiota from sequencing data. *Gut Microbes.* 2022;14(1):2149019. doi:10.1080/19490976.2022.2149019.
42. Ye K, Zhao Y, Huang W, Zhu Y. Sodium butyrate improves renal injury in diabetic nephropathy through AMPK/SIRT1/PGC-1 α signaling pathway. *Sci Rep.* 2024;14(1):17867. doi:10.1038/s41598-024-68227-8.
43. Stols-Gonçalves D, Mak AL, Madsen MS, van der Vossen EW, Bruinstroop E, Henneman P, Mol F, Scheithauer TPM, Smits L, Witjes J, et al. Faecal microbiota transplantation affects liver DNA methylation in

- non-alcoholic fatty liver disease: a multi-omics approach. *Gut Microbes*. 2023;15(1):2223330. doi:10.1080/19490976.2023.2223330.
44. Fernández-Pato A, Sinha T, Gacesa R, Andreu-Sánchez S, Gois MFB, Gelderloos-Arends J, Jansen DBH, Kruk M, Jaeger M, Joosten LAB, et al. Choice of DNA extraction method affects stool microbiome recovery and subsequent phenotypic association analyses. *Sci Rep*. 2024;14(1):3911. doi:10.1038/s41598-024-54353-w.
45. Ravikrishnan A, Wijaya I, Png E, Chng KR, Ho EXP, Ng AHQ, Naim ANM, Gounot JS, Guan SP, Hanqing JL, et al. Gut metagenomes of Asian octogenarians reveal metabolic potential expansion and distinct microbial species associated with aging phenotypes. *Nat Commun*. 2024;15(1):7751. doi:10.1038/s41467-024-52097-9.