Supplementary appendix

Supplemental text

Metadata in GNHS and the hip fracture case-control study

Metadata included in this study was further categorized into 4 groups:

1) 5 demographic factors: age, sex, household income, marital status and self-reported educational level.

2) 10 lifestyle and dietary factors: physical activity, total energy intake, alcohol drinking, smoking, tea drinking, vegetable intake, fruit intake, fish intake, red and processed meat intake, and yogurt intake.

3) 5 blood test factors: Fasting glucose, HDL, LDL, TC, and TG.

4) 8 anthropometry factors: height, weight, hip circumference, waist circumference, neck circumference, BMI, DBP, SBP.

Description of each factor in different cohorts is listed in Table 1.

Demographic, lifestyle and dietary factors were all collected by questionnaire during on-site face-to-face interviews. Habitual dietary intakes over the past 12 months were assessed by a food frequency questionnaire, as previously described (1). Physical activity was assessed as a total metabolic equivalent for task (MET) hours per day on the basis of a validated questionnaire for physical activity (2). Anthropometric factors were measured by trained nurses on site during the baseline interview. Fasting venous blood samples were taken at each recruitment or follow-up visit. Serum low-density lipoprotein cholesterol and glucose were measured by coloimetric methods using a

Roche Cobas 8000 c702 automated analyzer (Roche Diagnostics GmbH, Shanghai, China). Intra-assay coefficients of variation (CV) was 2.5% for glucose. Insulin was measured by electrochemiluminescence immunoassay (ECLIA) methods using a Roche cobas 8000 e602 automated analyzer (Roche Diagnostics GmbH, Shanghai, China). High-performance liquid chromatography was used to measure glycated hemoglobin (HbA1c) using the Bole D-10 Hemoglobin A1c Program on a Bole D-10 Hemoglobin Testing System, and the intraassay CV was 0.75%. The whole-body composition was measured by dual-energy x-ray absorptiometry (DXA) (Discovery W; Hologic Inc.). We analyzed the lean mass, fat mass and bon mass of the whole body, arms, and legs using the Hologic Discovery software version 3.2 (3).

Stool sample collection and DNA extraction

The stool samples were collected at a local study site within the School of Public Health at Sun Yat-sen University, and were transferred to a -80°C facility within 4 hours after collection. Total bacterial DNA was extracted using the QIAamp® DNA Stool Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. DNA concentrations were measured using the Qubit quantification system (Thermo Scientific, Wilmington, DE, US). The extracted DNA was then stored at -20 °C.

16S gene amplicon sequencing

The 16S rRNA gene amplification procedure was divided into two PCR steps, in the first PCR reaction, the V3-V4 hypervariable region of the 16S rRNA gene was amplified from genomic DNA using primers 341F(CCTACGGGNGGCWGCAG) and 805R(GACTACHVGGGTATCTAATCC). Amplification was performed in 96-well microtiter plates with a reaction mixture consisting of 1X KAPA HiFi Hot start Ready Mix, 0.1μM primer 341 F, 0.1 μM primer 805 R, and 12.5 ng template DNA giving a total volume of 50 μL per sample. Reactions were run in a T100 PCR thermocycle (BIO-RAD) according to the following cycling program: 3 min of denaturation at 94 °C, followed by 18 cycles of 30 s at 94 °C (denaturing), 30 s at 55 °C (annealing), and 30 s at 72 °C (elongation), with a final extension at 72 °C for 5 min. Subsequently, the amplified products were checked by 2% agarose gel electrophoresis and ethidium bromide staining. Amplicons were quantified using the Qubit quantification system (Thermo Scientific, Wilmington, DE, US) following the manufacturers' instructions. Sequencing primers and adaptors were added to the amplicon products in the second PCR step as follows 2 μL of the diluted amplicons were mixed with a reaction solution consisting of 1×KAPA HiFi Hotstart ReadyMix, 0.5μM fusion forward and 0.5μM fusion reverse primer, 30 ng Meta-gDNA(total volume 50 μL). The PCR was run according to the cycling program above except with cycling number of 12. The amplification products were purified with Agencourt AMPure XP Beads (Beckman Coulter Genomics, MA, USA) according to the manufacturer's instructions and quantified as described above. Equimolar amounts of the amplification products were pooled together in a single tube. The concentration of the pooled libraries was determined by the Qubit quantification system. Amplicon sequencing was performed on the Illumina MiSeq System (Illumina Inc., CA, USA). The MiSeq Reagent Kits v2

(Illumina Inc.) was used. Automated cluster generation and 2 × 250 bp paired-end sequencing with dual-index reads were performed.

16S rRNA gene sequence data processing

Fastq-files were demultiplexed by the MiSeq Controller Software (Illumina Inc.). The sequence was trimmed for amplification primers, diversity spacers, and sequencing adapters, merge-paired and quality filtered by USEARCH. UPARSE was used for OTU clustering equaling or above 97%. Taxonomy of the OTUs was assigned and sequences were aligned with RDP classifier. The OTUs were analyzed by phylogenetic and operational taxonomic unit (OTU) methods in the Quantitative Insights into Microbial Ecology (QIIME) software version 1.9.0 (4). α-diversity (Observed OTU number, Shannon index, Simpson index, Chao1 index, Goods coverage index) and β-diversity (Unweight UniFrac distances and Weight UniFrac distances) measures were calculated based on the rarefied OTU counts.

Type 2 diabetes risk variants and genetic risk score

We used 28 significant variants identified in a meta-analysis of CKB and AGEN-type 2 diabetes studies (5) to construct a type 2 diabetes genetic risk score(GRS) as

$$GRS_i = \sum_{j=1}^{m} x_{ij} b_j$$

Where, GRS_i is a genetic risk score for individual *i, m* is the number of SNPs in the score, x_{ij} represented the number of the risk allele on two chromosomes for ith individual and jth SNP, $x_{ij} \in \{0,1,2\}$,b_j represent the natural logarithm of the published odds ratio.

Metagenomic sequencing

Samples were metagenomically sequenced as one library each multiplexed through Illumina HiSeq machines and sequenced using the 2 × 100 bp paired-end read protocol. PRINSEQ v0.20.4 (6) was employed to sample dereplication and low complexity filtering. The length of each reads was trimmed with FASTX from the 5′ e and 3′ end using a quality threshold of 20. Read pairs with either reads was shorter than 60 bp or contained "N" were removed. 3) deduplicate the reads. Bowtie2 v2.2.5 (7) (using --reorder --no-contain --dovetail) was used to map reads to the human genome for decontamination.

Taxonomy analysis

Taxonomic profiling of the metagenomic samples was performed using MetaPhlAn2 v2.6.02, which uses a library of clade-specific markers to provide pan-microbial (bacterial, archaeal, viral and eukaryotic) quantification at the species level. MetaPhlAn2 (8) was run using default settings.

Metabolomics profiling of human serum samples

For the discovery cohort and external validation cohort1, targeted identification and quantification of serum metabolites was performed using an ultra-performance liquid chromatography coupled to tandem mass spectrometry (UPLC-MS/MS) system. This platform provides measures of 199 serum metabolome traits, including 12 subclasses.

All of the standards of targeted metabolites were commercially purchased from Sigma-Aldrich (St. Louis, MO, USA), Steraloids Inc. (Newport, RI, USA) and TRC Chemicals (Toronto, ON, Canada). All the standards were prepared in water, methanol, sodium hydroxide solution, or hydrochloric acid solution to obtain individual stock solution at a concentration of 5.0 mg/mL. Appropriate amount of each stock solution was mixed to create stock calibration solutions.

Samples were thawed on ice-bath to diminish sample degradation and prepared as follows: 25μL of plasma was added to a 96-well plate and then the plate was transferred to the Biomek 4000 workstation (Biomek 4000, Beckman Coulter, Inc., Brea, California, USA). Three types of quality control samples i.e., test mixtures, internal standards, and pooled biological samples are routinely used in metabolomics platform. In addition to the quality controls, conditioning samples, and solvent blank samples are also required for obtaining optimal instrument performance. 100μL ice cold methanol with partial internal standards was automatically added to each sample and vortexed vigorously for 5 minutes. The plate was centrifuged at 4000g for 30 minutes (Allegra X-15R, Beckman Coulter, Inc., Indianapolis, IN, USA). Then the plate was returned back to the workstation. 30μL of supernatant was transferred to a clean 96-well plate, and 20μL of freshly prepared derivative reagents was added to each well. The plate was sealed and the derivatization was carried out at 30°C for 60 min. After derivatization, 350μL of ice-cold 50% methanol solution was added to dilute the sample. Then the plate was stored at -20°C for 20 minutes and followed by 4000g centrifugation at 4 °C for 30 minutes. 135μL of supernatant was transferred to a new 96-well plate with 15μL internal standards in each well. Serial dilutions of derivatized stock standards were added to the left wells. Finally, the plate was sealed for LC-MS analysis. The raw data files from UPLC-MS/MS were processed using the QuanMET software (v2.0, Metabo-Profile, Shanghai, China) to perform peak integration, calibration, and quantitation for each metabolite.

Classification Analysis

To train and validate our model, we divided the discovery cohort into three parts randomly at a ratio of 6:2:2, which were allocated at the training cohort, internal validation cohort, and internal test cohort, respectively. The hyperparameters of the model were tuned on the internal validation cohort.

In the discovery cohort and external validation cohort 1, we calculated the area under the receiver operating curve (AUC) for type 2 diabetes prediction for the identified microbiota features, host genetics (type 2 diabetes genetic risk score), and the traditional type 2 diabetes risk factors including the Framingham-Offspring Risk Score (FORS) components(age, sex, parental history of diabetes, BMI, systolic blood pressure, high-density lipoprotein cholesterol, triglycerides, and waist circumference), lifestyle and dietary factors (current smoking status, current tea-drinking, current alcohol drinking, physical activity, total energy intake, vegetable intake, fish intake, red and processed meat intake, fruit intake and yogurt intake).

Microbiome risk score (MRS) formula

$$MRS_i = \sum_{j=1}^{n} s_{ij}$$

Where, MRS_i is a MRS for individual *i,* $s_{ij} = \begin{cases} 0\,, if\ \ x_{shap,ij} < 0 \\ 1\,, if\ \ x_{shap,ij} > 0 \end{cases}$, s_{ij} is the microbiome risk score for the *jth* microbiome features in *ith* individual. *n* is the sum of the microbiome features, and $x_{shap,ij}$is the SHAP value for the *jth* microbiome features in *ith* individual.

Faecal suspension inoculum preparation and faecal microbiota transplantation

Nine participants were randomly selected as the representative donors according to the level of the MRS (ranges from 0-14):

(1) Low MRS group: 3 participants, MRS=0, or MRS=1.

(2) High MRS + non-type 2 diabetes group: 3 participants, MRS=11.

(3) High MRS + type 2 diabetes group: 3 participants, MRS=13, or MRS=14.

Each fecal sample (0.5 g) was diluted in 5 mL of a 0.09% (w/v) sterile normal saline in an anaerobic chamber (80% N_2:10% CO_2:10% H_2). The fecal material was suspended by thorough vortexing (5 min) and centrifuged at 4 °C 300 rpm/min for 5 min. The clarified supernatant was transferred to a clean tube and used immediately for gut microbiota transplantation. Surveillance for bacterial contamination was performed by periodic bacteriological examinations of feces, food and padding.

Normal saline was added into the samples with sufficient mixing. The mixtures were then cultured using the spread plate method on: 1) LB agar, Brain Heart Infusion agar and Thioglycolate agar under aerobic condition at 37°C for aerobic bacteria; 2) on Gifu anaerobic medium (GAM) agar under anaerobic condition at 37°C for anaerobic bacteria; and 3) on Modified Martin Agar and Tryptone Soya agar under aerobic condition at 25-28°C for fungi. All cultures were examined under optical microscope after 1, 2, 4, 7 and 14 days.

Weaned, germ-free male C57BL/6J mice (*n* = 40) were maintained in flexible-film plastic isolators under a regular 12-h light cycle (lights on at 06:00). The mice were fed a sterilized normal chow diet (10% energy from fat; 3.25 kcal/g; SLAC). At 4 weeks of age, the germ-free mice were housed in individual cages and randomly divided into four groups (each group was kept in an individual isolator). After 1 weeks of acclimatization, the CON group of mice (*n* = 10) were orally gavaged with 100 μL of normal saline, and the other three groups of mice (*n* = 10, per group) were orally gavaged with 100 μL of the fecal suspension inoculum (taken from the each of the above donor group, preparation methods see supplementary materials). All mice were fed a sterilized high-fat diet. On Day 0, 7 and 14, after 12 h of fasting, fasting glucose was measured through the tail vein (Sinocare, China).

Reference

1. Zhang CX HS. Validity and reproducibility of a food frequency Questionnaire among Chinese women in Guangdong province. Asia Pac J Clin Nutr. 2009;18:240–50.
2. Liu B, Woo J, Tang N, Ng K, Ip R, Yu A. Assessment of total energy expenditure in a Chinese population by a physical activity questionnaire : examination of validity. Int J Food Sci Nutr. 2001;52:269–82.
3. Chen Y, Liu Y, Liu Y, Wang X, Guan K, Zhu H. Higher serum concentrations of betaine rather than choline is associated with better profiles of DXA-derived body fat and fat distribution in Chinese adults. 2014;39(3):465–71.
4. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.
5. Gan W, Walters RG, Holmes M V., Bragg F, Millwood IY, Banasik K, et al. Evaluation of type 2 diabetes genetic risk variants in Chinese adults: findings from 93,000 individuals from the China Kadoorie Biobank. Diabetologia. 2016;59(7):1446–57.
6. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4.
7. Ben Langmead SLS. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
8. Senavirathne G, Liu J, Jr MAL, Hanne J, Martin-lopez J, Lee J, et al.

MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12(10):902–3.

Fig.S1. Study overview. (A) Identifying microbiome features, together with their optimal threshold and direction associated with type 2 diabetes. 1) Training and optimizing a machine-learning model to link the input factors with type 2 diabetes in a discovery cohort (n=1832, 270 cases); 2) Using SHAP method to explain the output of machine learning model and identify the microbiota pattern associated with type 2 diabetes risk; 3) Constructing a microbiome risk score (MRS) for type 2 diabetes based on the above-identified microbiota pattern. 4) Validating the MRS-type 2 diabetes association in two independent external validation cohorts: cohort 1 (n=203, 48 cases), cohort 2 (n=7009, 608 cases); 5) Validating the MRS-type 2 diabetes association by faecal microbiota transplantation (FMT). **(B)** Investigating the prospective association of baseline adiposity, dietary and lifestyle factors with the identified type 2 diabetes-related gut microbiota pattern (i.e., MRS), and the correlation of the MRS with host serum metabolome. Further, we investigated the role of body fat distribution linking the MRS and type 2 diabetes development in the discovery cohort and external validation cohort 1.

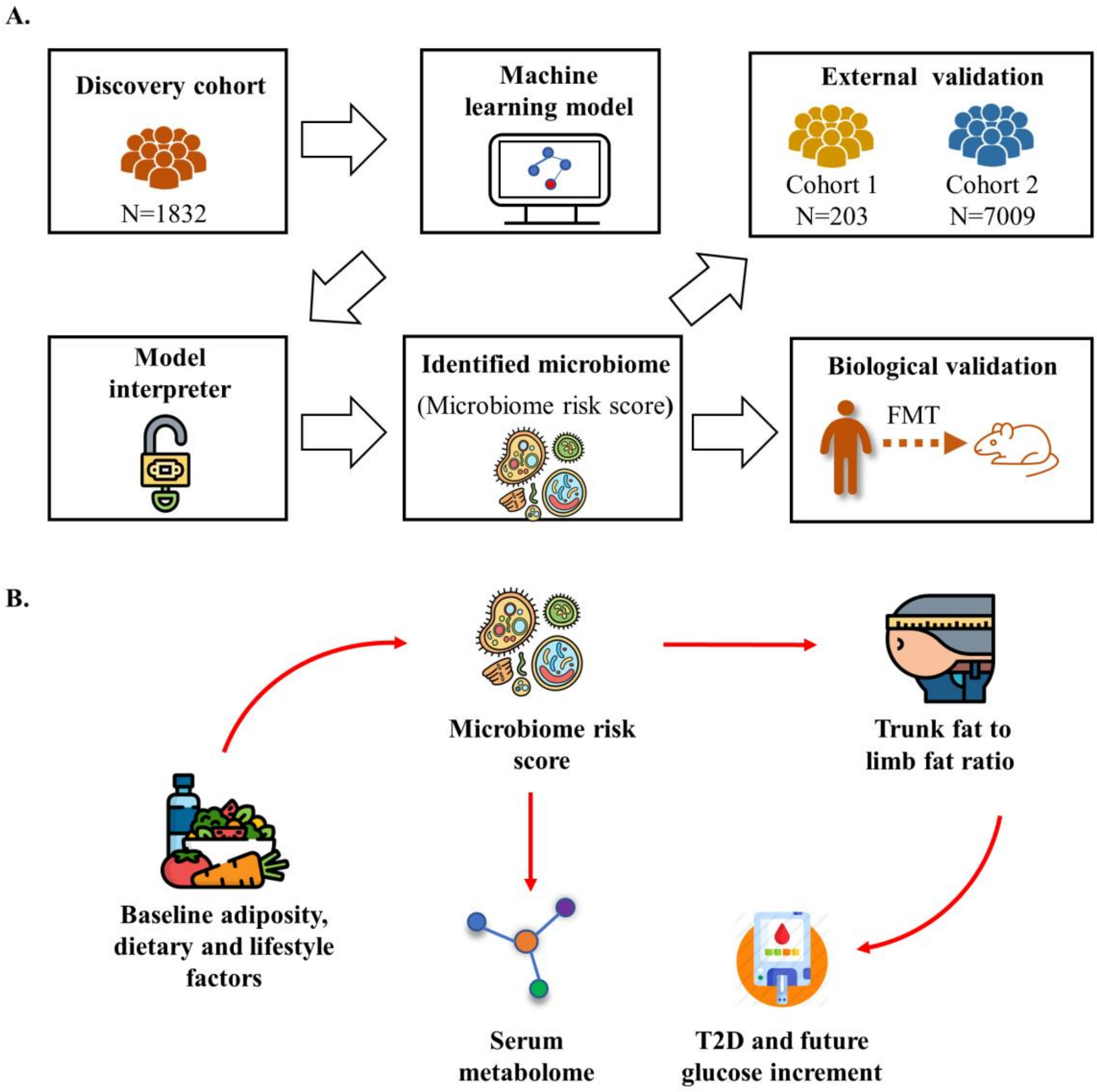

A.
Discovery cohort
N=1832
Machine learning model
External validation
Cohort 1
N=203
Cohort 2
N=7009
Model interpreter
Identified microbiome
(Microbiome risk score)
Biological validation
FMT
B.
Microbiome risk score
Trunk fat to limb fat ratio
Baseline adiposity, dietary and lifestyle factors
Serum metabolome
T2D and future glucose increment

Fig.S2. Overview of the discovery cohort: Guangzhou Nutrition and Health Study

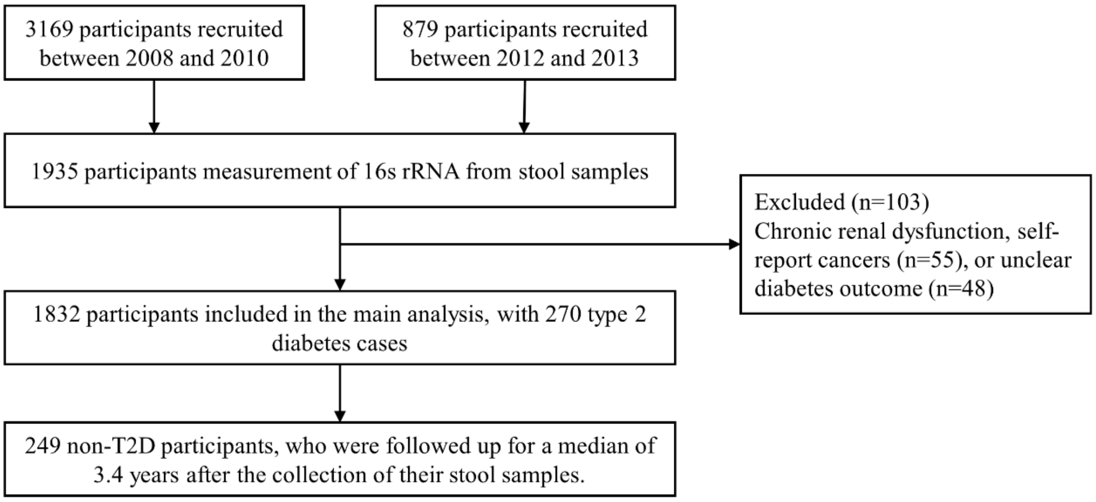

3169 participants recruited between 2008 and 2010
879 participants recruited between 2012 and 2013
1935 participants measurement of 16s rRNA from stool samples
Excluded (n=103)
Chronic renal dysfunction, self-report cancers (n=55), or unclear diabetes outcome (n=48)
1832 participants included in the main analysis, with 270 type 2 diabetes cases
249 non-T2D participants, who were followed up for a median of 3.4 years after the collection of their stool samples.

Fig.S3. The average impact of selected features on type 2 diabetes risk. The bars are colored according to data categories.

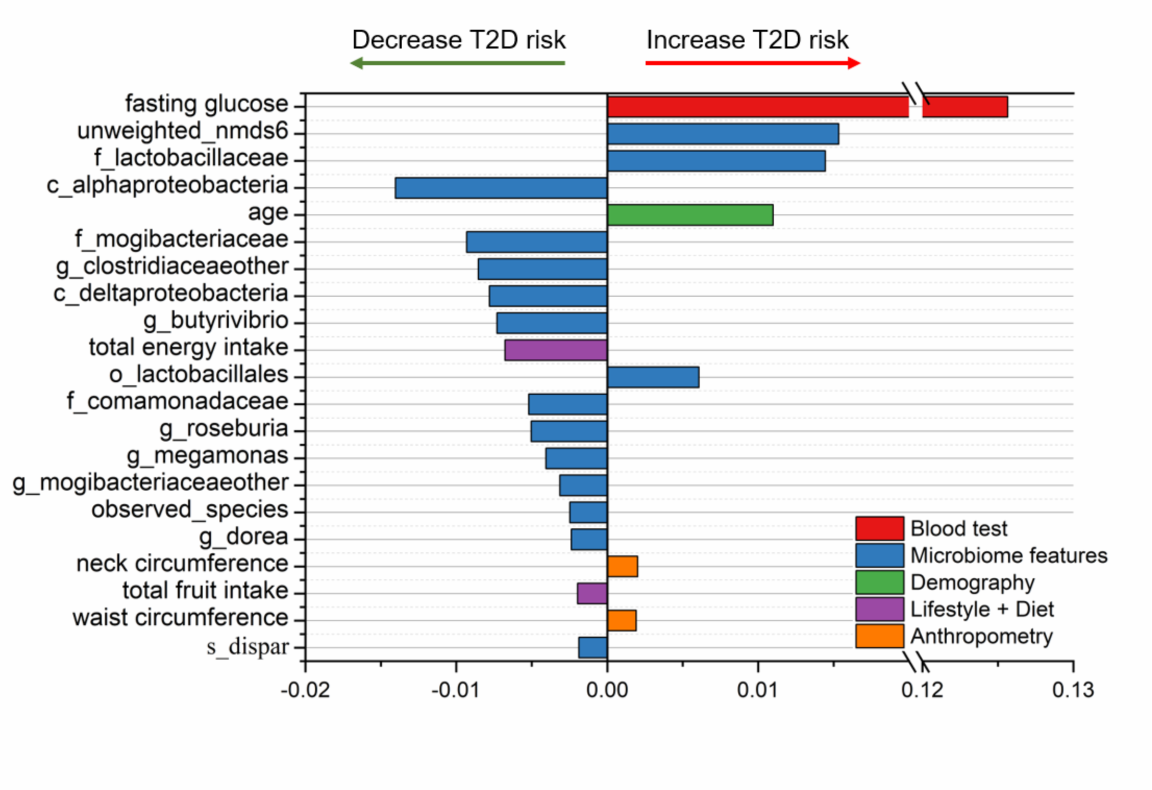
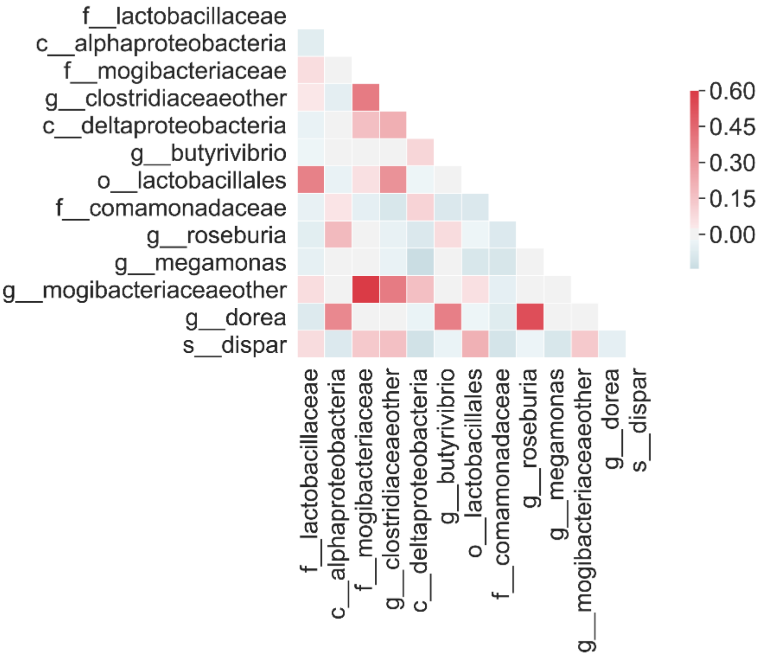


Fig.S4. The inter-correlation of selected taxa-related features in the discovery cohort (A) and external validation cohort 1 (B).

A.



B.

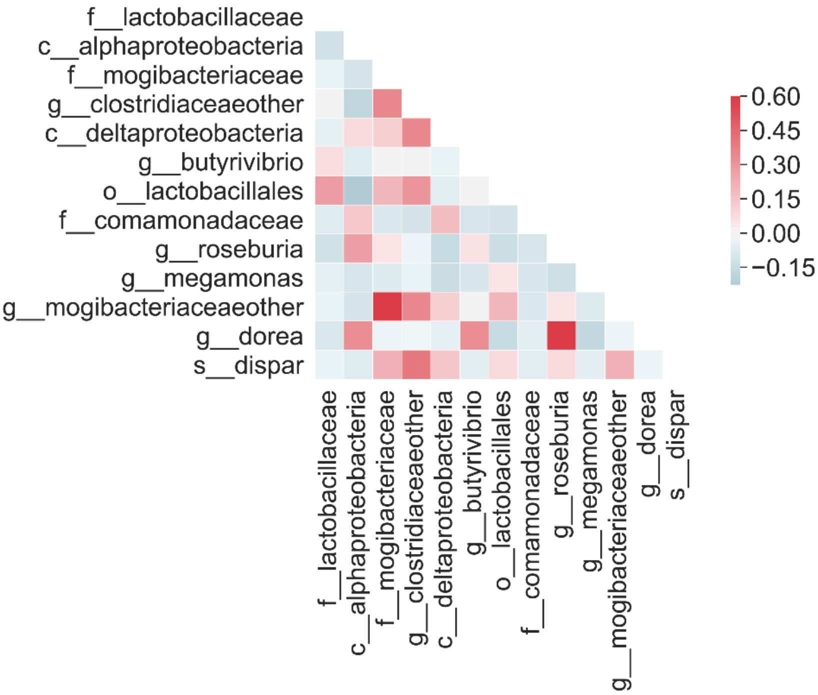


Fig.S5. Association of the microbiome risk score (MRS) with type 2 diabetes risk in different cohorts. Poisson regression was used to estimate the risk ratio (RR) and 95% confidence interval (CI) of type 2 diabetes per one unit change in the MRS, adjusting for demographic, dietary and lifestyle factors. The MRS was constructed based on the conventional method.

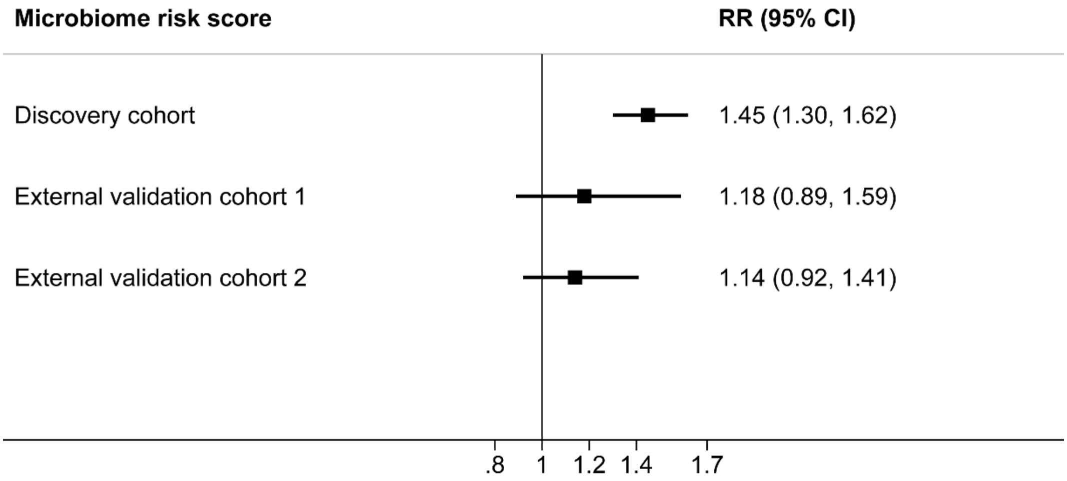
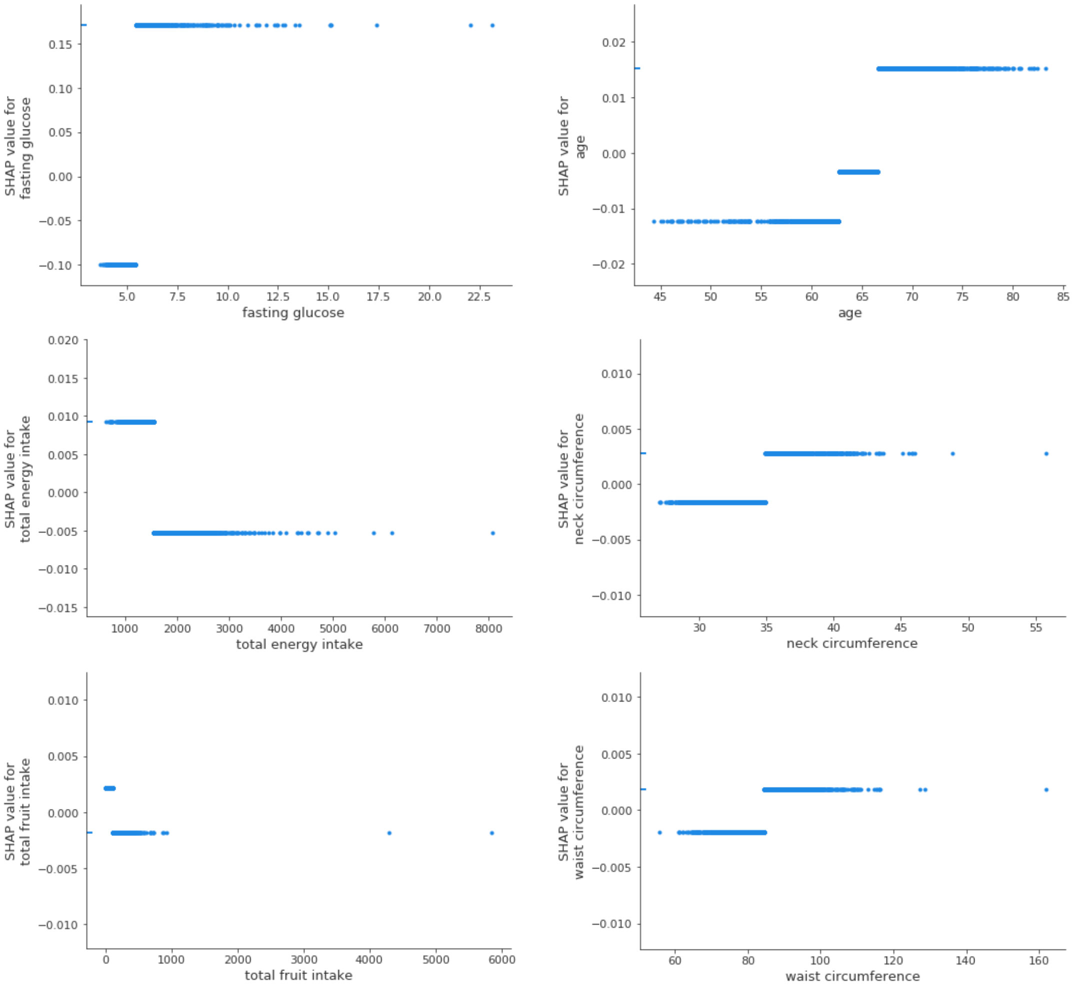
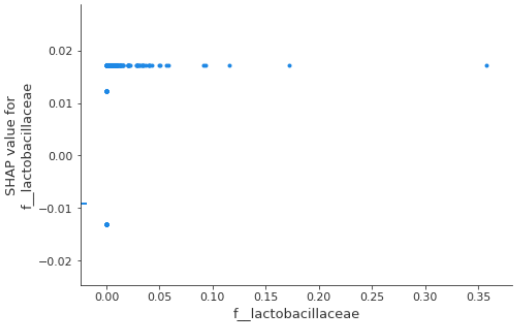
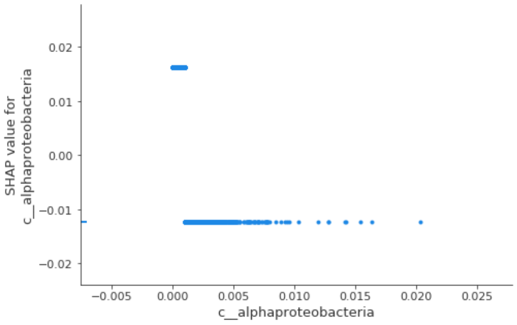
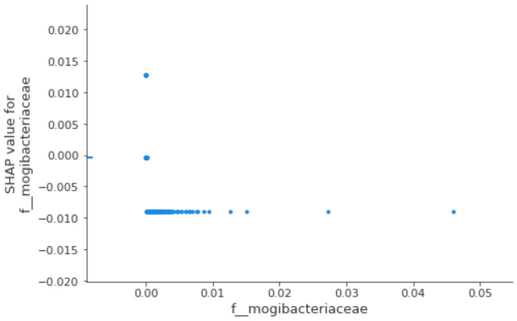


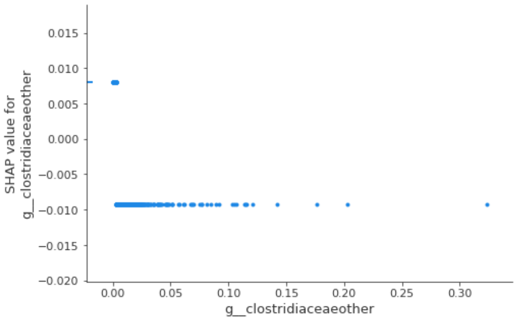
Fig.S6. The marginal effect of individual selected features on type 2 diabetes. We plot the SHAP values of every feature for each sample. X-axis represents the feature variable, while Y-axis represents the SHAP value for the feature variable. SHAP value greater than zero indicates that the feature may increase the type 2 diabetes risk for the given sample, otherwise, decrease the disease risk.

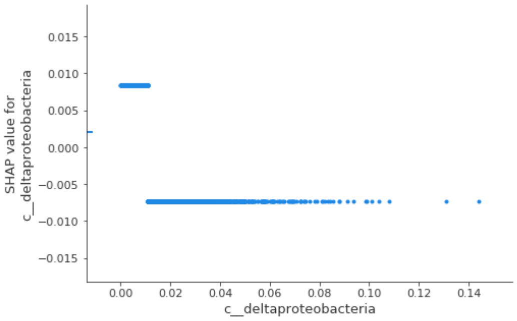


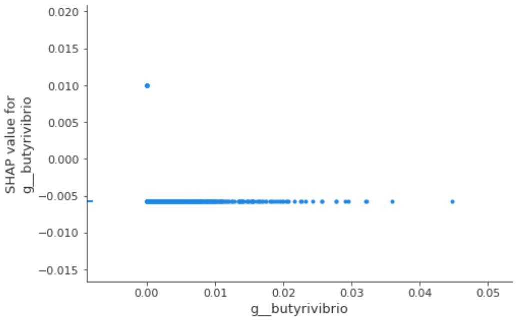

SHAP value for
f__lactobacillaceae
0.02
0.01
0.00
−0.01
−0.02
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
f__lactobacillaceae

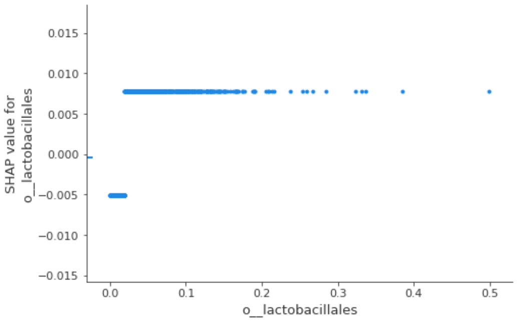

SHAP value for
c__alphaproteobacteria
0.02
0.01
0.00
−0.01
−0.02
−0.005
0.000
0.005
0.010
0.015
0.020
0.025
c__alphaproteobacteria

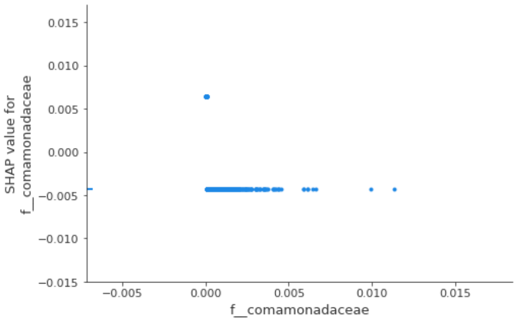

SHAP value for
f__mogibacteriaceae
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
0.00
0.01
0.02
0.03
0.04
0.05
f__mogibacteriaceae

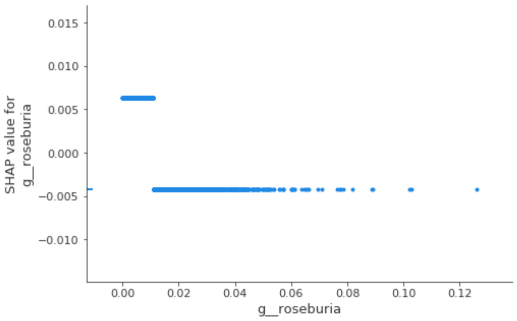

SHAP value for
g__clostridiaceaeother
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
0.00
0.05
0.10
0.15
0.20
0.25
0.30
g__clostridiaceaeother

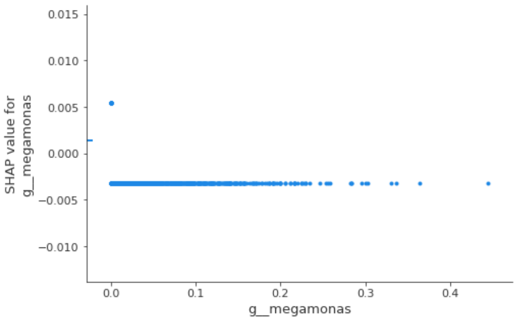

SHAP value for
c__deltaproteobacteria
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
c__deltaproteobacteria


SHAP value for
g__butyrivibrio
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
0.00
0.01
0.02
0.03
0.04
0.05
g__butyrivibrio


SHAP value for
o__lactobacillales
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
0.0
0.1
0.2
0.3
0.4
0.5
o__lactobacillales


SHAP value for
f__comamonadaceae
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.005
0.000
0.005
0.010
0.015
f__comamonadaceae


SHAP value for
g__roseburia
0.015
0.010
0.005
0.000
−0.005
−0.010
0.00
0.02
0.04
0.06
0.08
0.10
0.12
g__roseburia


SHAP value for
g__megamonas
0.015
0.010
0.005
0.000
−0.005
−0.010
0.0
0.1
0.2
0.3
0.4
g__megamonas

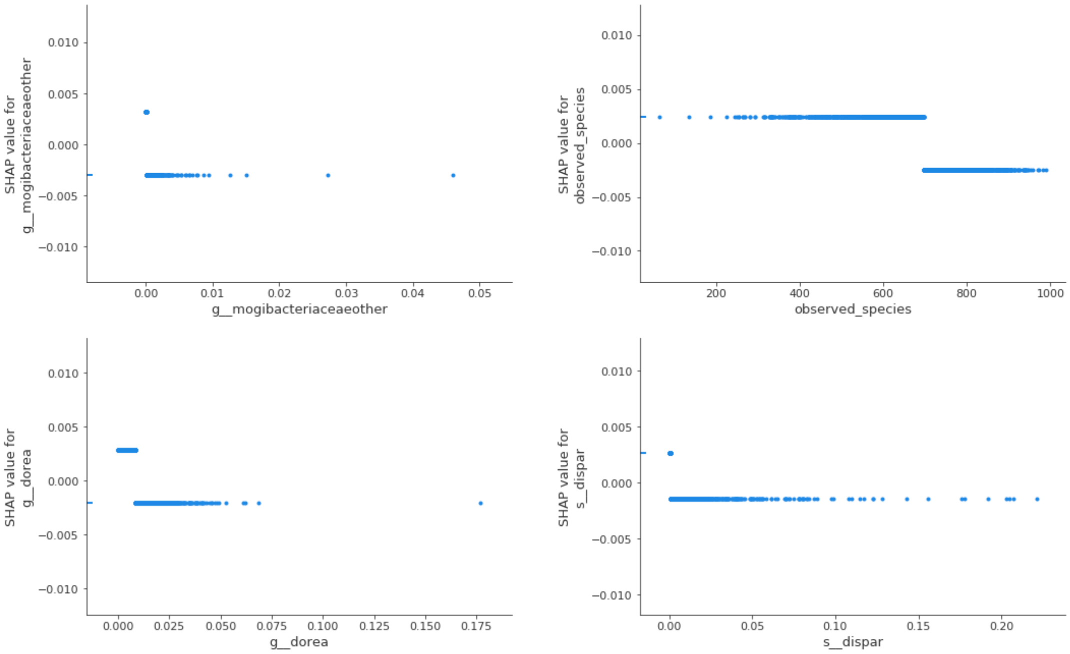
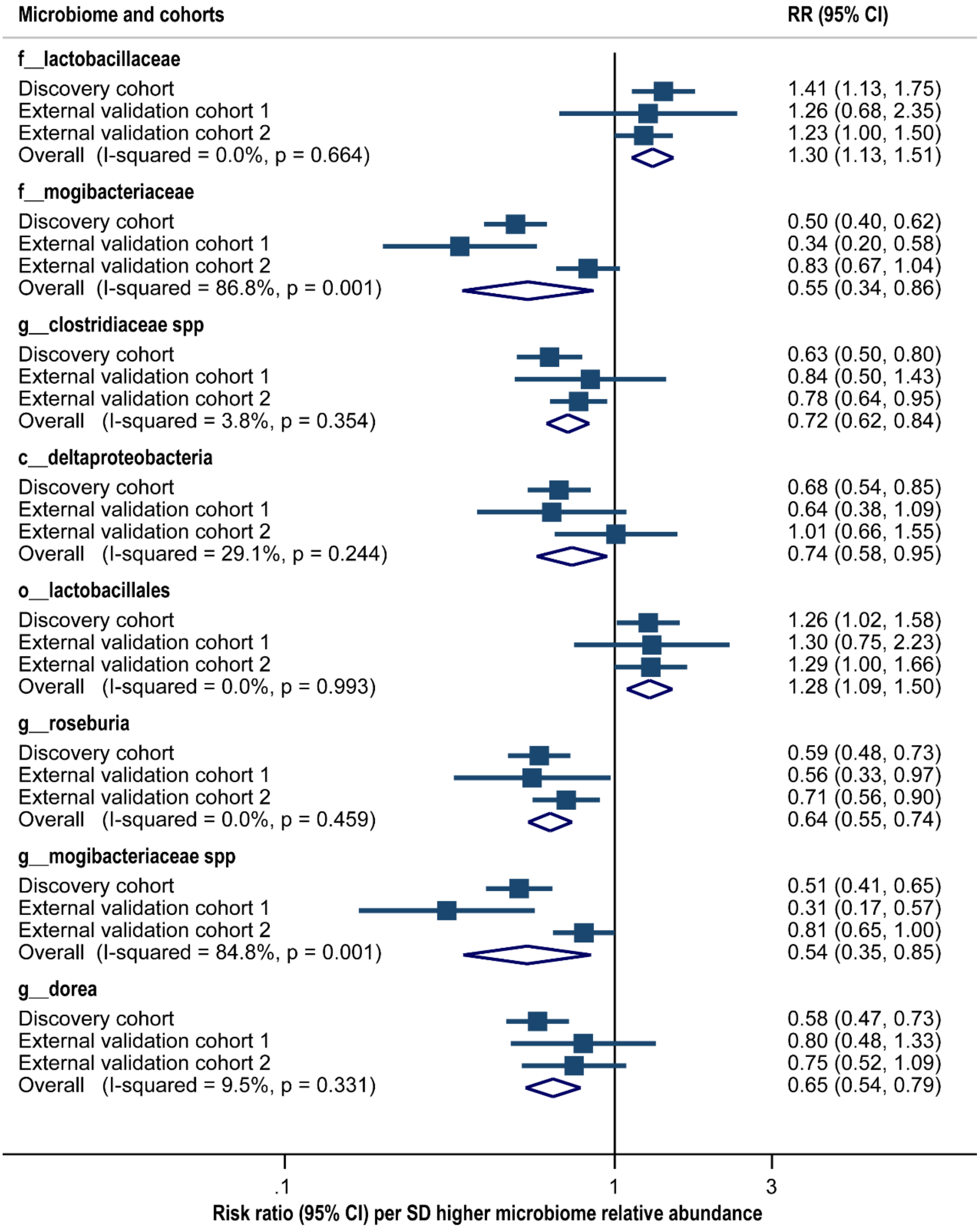

SHAP value for
g__mogibacteriaceaeother
0.010
0.005
0.000
−0.005
−0.010
0.00
0.01
0.02
0.03
0.04
0.05
g__mogibacteriaceaeother
SHAP value for
observed_species
0.010
0.005
0.000
−0.005
−0.010
200
400
600
800
1000
observed_species
SHAP value for
g__dorea
0.010
0.005
0.000
−0.005
−0.010
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
g__dorea
SHAP value for
s__dispar
0.010
0.005
0.000
−0.005
−0.010
0.00
0.05
0.10
0.15
0.20
s__dispar

Fig.S7. Associations of the selected microbiome features with risk of type 2 diabetes. In this graph, we only present the microbiome that was significantly associated with type 2 diabetes risk. (A) Multivariable Poisson regression model was used to examine the association with type 2 diabetes for each selected taxa-related feature at higher abundance (i.e., higher the optimal threshold) with those at lower abundance (i.e., lower the optimal threshold). Covariates included in the statistical models for the discovery cohort and external validation cohort 1 were as follows: age, sex, BMI, waist circumference, total energy intake, alcohol drinking, smoking, household income, marital status, and self-reported educational level. For external validation cohort 2, all aforementioned covariates but total energy intake (not collected in external validation cohort 2) were used in the statistical model. (B) Multivariable Poisson regression model was used to estimate type 2 diabetes risk per SD change in the selected taxa-related features, adjusted for the abovementioned covariates.

A.


Microbiome and cohorts
RR (95% CI)
f__lactobacillaceae
Discovery cohort
1.41 (1.13, 1.75)
External validation cohort 1
1.26 (0.68, 2.35)
External validation cohort 2
1.23 (1.00, 1.50)
Overall (I-squared = 0.0%, p = 0.664)
1.30 (1.13, 1.51)
f__mogibacteriaceae
Discovery cohort
0.50 (0.40, 0.62)
External validation cohort 1
0.34 (0.20, 0.58)
External validation cohort 2
0.83 (0.67, 1.04)
Overall (I-squared = 86.8%, p = 0.001)
0.55 (0.34, 0.86)
g__clostridiaceae spp
Discovery cohort
0.63 (0.50, 0.80)
External validation cohort 1
0.84 (0.50, 1.43)
External validation cohort 2
0.78 (0.64, 0.95)
Overall (I-squared = 3.8%, p = 0.354)
0.72 (0.62, 0.84)
c__deltaproteobacteria
Discovery cohort
0.68 (0.54, 0.85)
External validation cohort 1
0.64 (0.38, 1.09)
External validation cohort 2
1.01 (0.66, 1.55)
Overall (I-squared = 29.1%, p = 0.244)
0.74 (0.58, 0.95)
o__lactobacillales
Discovery cohort
1.26 (1.02, 1.58)
External validation cohort 1
1.30 (0.75, 2.23)
External validation cohort 2
1.29 (1.00, 1.66)
Overall (I-squared = 0.0%, p = 0.993)
1.28 (1.09, 1.50)
g__roseburia
Discovery cohort
0.59 (0.48, 0.73)
External validation cohort 1
0.56 (0.33, 0.97)
External validation cohort 2
0.71 (0.56, 0.90)
Overall (I-squared = 0.0%, p = 0.459)
0.64 (0.55, 0.74)
g__mogibacteriaceae spp
Discovery cohort
0.51 (0.41, 0.65)
External validation cohort 1
0.31 (0.17, 0.57)
External validation cohort 2
0.81 (0.65, 1.00)
Overall (I-squared = 84.8%, p = 0.001)
0.54 (0.35, 0.85)
g__dorea
Discovery cohort
0.58 (0.47, 0.73)
External validation cohort 1
0.80 (0.48, 1.33)
External validation cohort 2
0.75 (0.52, 1.09)
Overall (I-squared = 9.5%, p = 0.331)
0.65 (0.54, 0.79)
.1
1
3
Risk ratio (95% CI) per SD higher microbiome relative abundance

B.

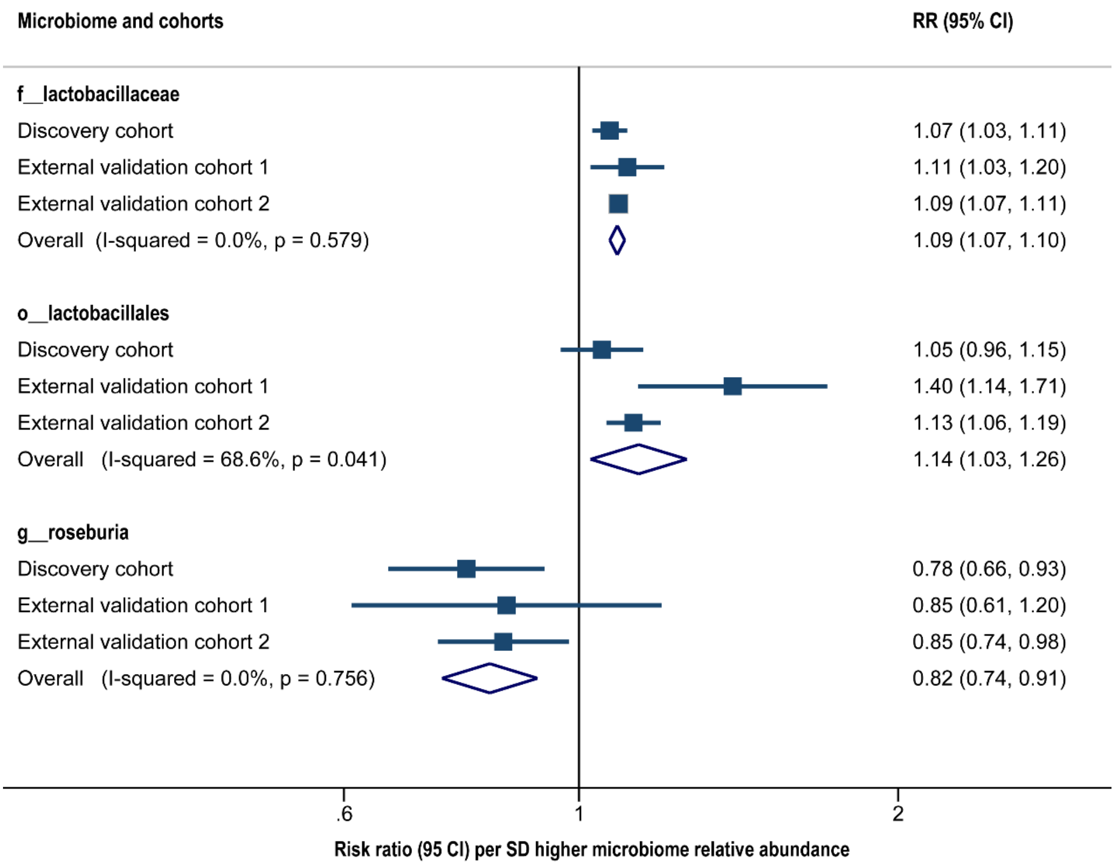

Microbiome and cohorts
RR (95% CI)
f__lactobacillaceae
Discovery cohort
1.07 (1.03, 1.11)
External validation cohort 1
1.11 (1.03, 1.20)
External validation cohort 2
1.09 (1.07, 1.11)
Overall (I-squared = 0.0%, p = 0.579)
1.09 (1.07, 1.10)
o__lactobacillales
Discovery cohort
1.05 (0.96, 1.15)
External validation cohort 1
1.40 (1.14, 1.71)
External validation cohort 2
1.13 (1.06, 1.19)
Overall (I-squared = 68.6%, p = 0.041)
1.14 (1.03, 1.26)
g__roseburia
Discovery cohort
0.78 (0.66, 0.93)
External validation cohort 1
0.85 (0.61, 1.20)
External validation cohort 2
0.85 (0.74, 0.98)
Overall (I-squared = 0.0%, p = 0.756)
0.82 (0.74, 0.91)
.6
1
2
Risk ratio (95 CI) per SD higher microbiome relative abundance

Fig.S8. Identified gut microbiota affects the type 2 diabetes development in germ-free mice. (A) Schematic diagram. (B) Fasting glucose curves. (C) Quantification of fasting glucose by AUC. * compared with CON group, # compared with Low MRS group, + compared with High MRS+non-type 2 diabetes group. (*, #, +) P< 0.05, (**, ##, ++) P< 0.01, (***, ###, +++) P< 0.001 by ANOVA. The P-values were adjusted using the Benjamini and Hochberg method.

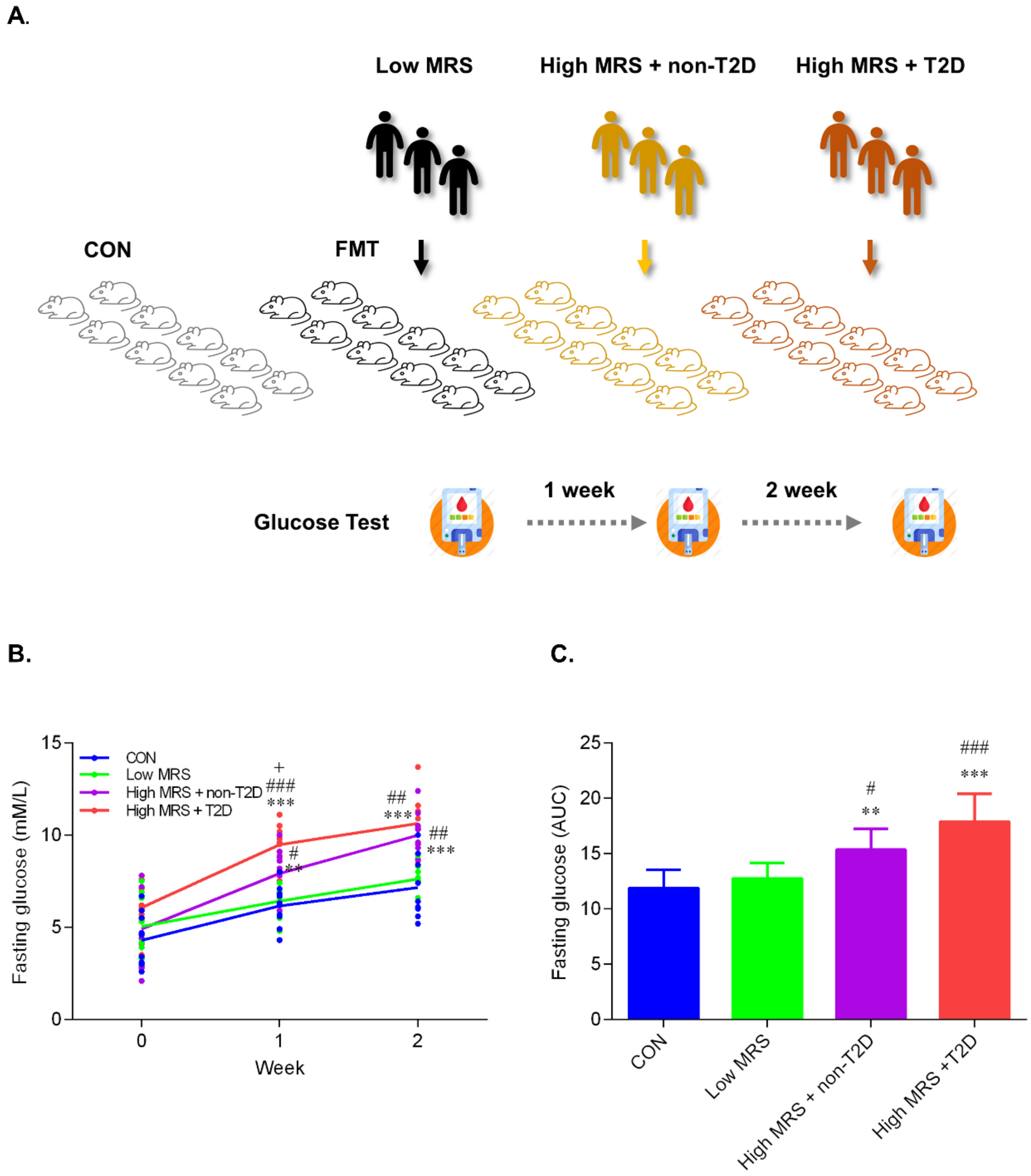


Table S1. Comparison of the prediction performance of LightGBM, random forest and logistic regression in the discovery cohort and validation cohort 1.

Algorithm	**Discovery cohort**			**Validation cohort 1**
	AUC (mean)	AUC (minimum)	AUC (maximum)	AUC
LightGBM	0.93	0.9	0.95	0.84
Random forest	0.84	0.79	0.88	0.53
Logistic regression	0.92	0.87	0.97	0.53

Table S2. Comparison of the prediction performance of all inputted and selected features in different cohorts.

Features	**Discovery cohort**			**Validation cohort 1**
	AUC (mean)	AUC (minimum)	AUC (maximum)	AUC
297 features	0.93	0.9	0.95	0.84
Identified 21 features	0.92	0.9	0.94	0.84

Table S3. Association of the gut microbiome risk score (MRS) with type 2 diabetes*

Cohorts	Median (MRS)	No. of cases / Total No.	Adjusted risk ratio (95% CI)	*P* value
Discovery cohort				
Quartile 1	3	33 / 569	1 (reference)	
Quartile 2	5	62 / 515	2.02 (1.35, 3.02)	<0.001
Quartile 3	7	70 / 419	2.73 (1.85, 4.04)	<0.001
Quartile 4	10	101 / 304	5.29 (3.66, 7.65)	<0.001
External validation cohort 1				
Quartile 1	4	7 / 65	1 (reference)	
Quartile 2	6	4 / 31	1.47 (0.49, 4.43)	0.49
Quartile 3	7	15 / 53	2.6 (1.17, 5.79)	0.019
Quartile 4	10	17 / 39	4.17 (1.96, 8.85)	<0.001
External validation cohort 2				
Quartile 1	6	236 / 3065	1 (reference)	
Quartile 2	7	147 / 1672	1.11 (0.91, 1.35)	0.31
Quartile 3	8	110 / 1104	1.27 (1.03, 1.57)	0.025
Quartile 4	9	104 / 946	1.36 (1.10, 1.68)	0.0051

*Poisson regression was used to estimate the risk ratio (RR) and 95% confidence interval (CI) of type 2 diabetes in each of the three cohorts, according to the gut microbiome risk score. In these comparisons, participants at low microbiome risk (Quartile 1) were treated as the reference group. The covariates for the discovery cohort and validation cohort 1 were total energy intake, age, waist circumference, sex, BMI, alcohol status, smoking status, education, marital status and income. For the validation cohort 2 (GGMP), covariates including age, waist circumference, sex, BMI, alcohol status, smoking status, education, marital status.

Table S4. Association of the gut microbiome risk score with type 2 diabetes stratified by age and sex in the discovery cohort *

	Mean (Microbiome risk score)	**No. of cases / Total No.**	**Adjusted risk ratio (95% CI)**	**P value**
Age				
< median	5.7	94 / 910	1.31 (1.21,1.41)	<0.001
⩾median	6.1	172 / 897	1.27 (1.21, 1.33)	<0.001
Sex				
Men	6	103 / 601	1.24 (1.17, 1.32)	<0.001
Women	5.9	163 / 1206	1.29 (1.23, 1.37)	<0.001

* Poisson regression was used to performed subgroup analysis for MRS-type 2 diabetes relationship stratified by age (<64.3 years vs. ≥64.3 years, with 64.3 years as the median age of this cohort) and sex in the discovery cohort. The covariates were total energy intake, age, waist circumference, sex, BMI, alcohol status, smoking status, education, marital status and income. The median age of the discovery cohort is 64.3 years.

Table S5. The optimal threshold of the selected microbiome features according to their SHAP dependence plot

Microbiome	Optimal threshold (relative abundance)	Taxa annotation
f__lactobacillaceae	0.0000877	p__Firmicutes; c__Bacilli; o__Lactobacillales; f__lactobacillaceae
c__alphaproteobacteria	0.00101	p__Proteobacteria; c__alphaproteobacteria
f__mogibacteriaceae	0.0000403	p__Firmicutes; c__Clostridia; o__Clostridiales; f__mogibacteriaceae
g__clostridiaceae spp	0.00313	p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__
c__deltaproteobacteria	0.0109	p__Proteobacteria; c__deltaproteobacteria
g__butyrivibrio	0.0000448	p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__butyrivibrio
o__lactobacillales	0.0193	p__Firmicutes; c__Bacilli; o__lactobacillales
f__comamonadaceae	0.0000645	p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__comamonadaceae
g__roseburia	0.011	p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__roseburia
g__megamonas	0.00054	p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__megamonas
g__mogibacteriaceae spp	0.0000855	p__Firmicutes; c__Clostridia; o__Clostridiales; f__mogibacteriaceae; g__
g__dorea	0.00861	p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__dorea
s__dispar	0.000757	p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Veillonella; s__dispar

Table S6. Associations of baseline adiposity and dietary factors with microbiome risk score*

	n	beta	95% CI	p
Age	1812	0.023	0.0026, 0.043	0.027
Energy intake	1812	0.059	-0.065, 0.18	0.35
MET	1812	-0.02	-0.12, 0.08	0.69
BMI	1812	0.1	0.023, 0.18	0.012
Educatioin	1812	0.2	0.042, 0.36	0.013
Hip circumference	1812	-0.039	-0.07, -0.007	0.017
Waist circumference	1812	-0.0041	-0.028, 0.02	0.74
Neck circumference	1812	-0.037	-0.099, 0.026	0.25
Income	1812	-0.12	-0.31, 0.06	0.19
Red and processed meat intake	1812	-0.051	-0.16, 0.59	0.37
Fruit intake	1812	-0.025	-0.14, 0.085	0.66
Fish intake	1812	0.061	-0.046, 0.17	0.26
Vegetable intake	1812	-0.08	-0.19, 0.03	0.15
Yogurt intake	1812	-0.027	-0.13, 0.076	0.6
Sex	1812	0.035	-0.38 0.45	0.87
Current alcohol drinking	1812	-0.33	-0.78, 0.12	0.15
Current tea drinking	1812	-0.25	-0.49, -0.018	0.035
Current smoke drinking	1812	0.09	-0.3, 0.48	0.65
Marital status	1812	0.144	-0.25, 0.54	0.47
Drug use	1812	2.56	2.18, 2.95	<0.001

*beta: correlation coefficient of baseline diet and basic attributes with microbiome features; CI: confidence interval.

Table S7. Associations of the microbiome risk score with body fat distribution in the discovery cohort*

Outcome	n	beta	95% CI	p
TOTAL_FAT	1750	-5.344	-27.28-16.59	0.63
TOTAL_MASS	1750	-10.166	-55.01-34.68	0.66
TOTAL_PFAT	1750	-0.032	-0.11-0.05	0.44
ANDROID_FAT	1750	2.577	-7.22-12.38	0.61
ANDROID_MASS	1750	5.064	-13.42-23.55	0.59
ANDROID_PFAT	1750	0.005	-0.1-0.11	0.93
GYNOID_FAT	1750	-7.921	-21.4-5.56	0.25
GYNOID_MASS	1750	-15.231	-42.97-12.51	0.28
GYNOID_PFAT	1750	-0.050	-0.13-0.03	0.22
TOTAL_PERCENT_FAT	1750	-0.004	-0.08-0.08	0.92
BODY_MASS_INDEX	1750	0.149	-0.45-0.74	0.62
ANDROID_GYNOID_RATIO	1750	0.002	-0.00084-0.0047	0.17
ANDROID_PERCENT_FAT	1750	0.005	-0.1-0.11	0.93
GYNOID_PERCENT_FAT	1750	-0.050	-0.13-0.03	0.22
FAT_MASS_RATIO	1750	0.005	0.0016-0.0074	0.00225
TRUNK_LIMB_FAT_MASS_RATIO	1750	0.007	0.0037-0.011	0.000117
FAT_MASS_HEIGHT_SQUARED	1750	0.033	-0.05-0.11	0.422
TOTAL_FAT_MASS	1750	2.746	-84.04-89.53	0.951
GLOBAL_FAT	1750	-3.066	-90.56-84.43	0.945
GLOBAL_MASS	1750	-34.092	-202.98-134.8	0.692
GLOBAL_PFAT	1750	-0.016	-0.1-0.07	0.705
HEAD_FAT	1750	-0.368	-2.12-1.38	0.681
HEAD_MASS	1750	-2.770	-10.15-4.62	0.462
HEAD_PFAT	1750	0.006	-0.0026-0.0014	0.183
LARM_FAT	1750	1.654	-4.96-8.27	0.624
LARM_MASS	1750	-2.245	-12.41-7.92	0.665
LARM_PFAT	1750	0.032	-0.09-0.16	0.606
RARM_FAT	1750	2.092	-4.3-8.48	0.521
RARM_MASS	1750	-1.769	-12.08-8.55	0.737
RARM_PFAT	1750	0.042	-0.08-0.16	0.490
TRUNK_FAT	1750	26.380	-24.08-76.84	0.306
TRUNK_MASS	1750	37.376	-55.31-130.06	0.429
TRUNK_PFAT	1750	0.030	-0.06-0.12	0.536
L_LEG_FAT	1750	-14.150	-29.21-0.91	0.066
L_LEG_MASS	1750	-28.815	-56.72--0.91	0.043
L_LEG_PFAT	1750	-0.079	-0.18-0.02	0.105
R_LEG_FAT	1750	-14.513	-30-0.97	0.066
R_LEG_MASS	1750	-26.408	-54.8-1.98	0.068
R_LEG_PFAT	1750	-0.093	-0.19-0.01	0.063
SUBTOT_FAT	1750	1.463	-85.46-88.39	0.974
SUBTOT_MASS	1750	-21.861	-182.34-138.62	0.789
SUBTOT_PFAT	1750	-0.007	-0.09-0.08	0.868
WBTOT_FAT	1750	1.095	-86.71-88.9	0.980
WBTOT_MASS	1750	-24.630	-189.48-140.21	0.770

WBTOT_PFAT	1750	-0.006	-0.09-0.08	0.888

*Linear regression was performed to examine the association of microbiome risk score with components of body fat distribution, adjusted for total energy intake, age, sex, alcohol status, smoking status, education, marital status and income

Microbiome risk score: components including index of α-diversity (observe species), and 13 taxa-related features (*f__lactobacillaceae, c__alphaproteobacteria, f__mogibacteriaceae, g__clostridiaceae spp, c__deltaproteobacteria, g__butyrivibrio, o__lactobacillales, f__comamonadaceae, g__roseburia, g__megamonas, g__mogibacteriaceae spp, g__dorea, s__dispar*).

Table S8. Associations of the microbiome risk score with body fat distribution in the external validation cohort 1*

Outcome	n	Beta	95% CI	p
TOTAL_FAT	185	-5.120	-75.29-55.53	0.884
TOTAL_MASS	185	19.324	-123.67-136.55	0.782
TOTAL_PFAT	185	-0.102	-0.35-0.15	0.449
ANDROID_FAT	185	15.973	-12.86-41.36	0.273
ANDROID_MASS	185	37.384	-18.1-83.13	0.169
ANDROID_PFAT	185	0.074	-0.24-0.42	0.678
GYNOID_FAT	185	-21.093	-65.86-17.6	0.348
GYNOID_MASS	185	-18.060	-109.09-56.94	0.686
GYNOID_PFAT	185	-0.191	-0.44-0.06	0.157
TOTAL_PERCENT_FAT	185	0.009	-0.23-0.26	0.943
BODY_MASS_INDEX	185	0.122	-0.08-0.3	0.231
ANDROID_GYNOID_RATIO	185	0.009	-0.00033-0.0178	0.059
ANDROID_PERCENT_FAT	185	0.074	-0.24-0.42	0.678
GYNOID_PERCENT_FAT	185	-0.191	-0.44-0.06	0.157
FAT_MASS_RATIO	185	0.007	0.0067-0.016	0.159
TRUNK_LIMB_FAT_MASS_RATIO	185	0.015	0.0023-0.03	0.020
FAT_MASS_HEIGHT_SQUARED	185	0.045	-0.06-0.15	0.438
TOTAL_FAT_MASS	185	102.950	-177.6-360.91	0.477
GLOBAL_FAT	185	102.918	-177.61-360.84	0.477
GLOBAL_MASS	185	213.248	-313.55-684.25	0.427
GLOBAL_PFAT	185	0.009	-0.23-0.26	0.944
HEAD_FAT	185	2.185	-3.92-6.54	0.437
HEAD_MASS	185	4.644	-20.15-23.75	0.694
HEAD_PFAT	185	0.021	-0.01-0.04	0.108
LARM_FAT	185	4.281	-15.06-23.13	0.677
LARM_MASS	185	9.323	-20.45-36.65	0.544
LARM_PFAT	185	-0.038	-0.39-0.34	0.844
RARM_FAT	185	6.775	-13.95-25.61	0.524
RARM_MASS	185	14.976	-19.18-43.11	0.371
RARM_PFAT	185	-0.028	-0.37-0.35	0.885
TRUNK_FAT	185	103.849	-39.88-242.11	0.171
TRUNK_MASS	185	202.665	-76.77-457.47	0.158
TRUNK_PFAT	185	0.090	-0.16-0.37	0.529
L_LEG_FAT	185	-4.545	-61.75-44.84	0.874
L_LEG_MASS	185	-10.916	-107.84-78.19	0.827
L_LEG_PFAT	185	-0.075	-0.42-0.23	0.669
R_LEG_FAT	185	-9.819	-65.59-40.44	0.731
R_LEG_MASS	185	-8.410	-102.94-75.64	0.861
R_LEG_PFAT	185	-0.141	-0.47-0.18	0.419
SUBTOT_FAT	185	100.541	-176.34-356.24	0.483
SUBTOT_MASS	185	207.638	-303.48-667.37	0.426
SUBTOT_PFAT	185	0.007	-0.25-0.28	0.963
WBTOT_FAT	185	102.726	-178.09-360.61	0.478
WBTOT_MASS	185	212.282	-315.69-683.18	0.429
WBTOT_PFAT	185	0.010	-0.23-0.26	0.942

*Linear regression was performed to examine the association of microbiome risk score with components of body fat distribution, adjusted for total energy intake, age, sex, alcohol status, smoking status, education, marital status and income

Microbiome risk score: components including index of α-diversity (observe species), and 13 taxa-related features (*f__lactobacillaceae, c__alphaproteobacteria, f__mogibacteriaceae, g__clostridiaceae spp, c__deltaproteobacteria, g__butyrivibrio, o__lactobacillales, f__comamonadaceae, g__roseburia, g__megamonas, g__mogibacteriaceae spp, g__dorea, s__dispar*).