Online Supplements

Study Population, Data Collection, and Laboratory Assay

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL)

The HCHS/SOL is a prospective population-based study of 16,415 Hispanic/Latino adults aged 18 to 74 years at recruitment who were living in four US metropolitan areas (Bronx, NY; Chicago, IL; Miami, FL; and San Diego, CA).[1,2] A comprehensive battery of interviews and a clinical assessment with blood draw (both fasting and two hours after a 75 gram glucose load) were conducted at in-person clinic visits during 2008 to 2011 (Visit 1 baseline) and during 2014 to 2017 (Visit 2). Information on demographics, behaviors, health status, family and medical histories, and medication use was collected using structured questionnaires, and blood pressure and anthropometric traits were measured. Blood biomarkers including blood lipids and glycemic traits, were measured by standard methods.[3] Annual follow-up telephone interviews were conducted to ascertain information on health status. The study was approved by the institutional review boards at all participating institutions, and all participants gave written informed consent.

Diabetes Ascertainment. Participants who met at least one of the following criteria were defined as T2D cases**:** 1) fasting time >8 hours and fasting glucose ≥7.0 mmol/L (126 mg/dl); 2) fasting ≤8 hours and non-fasting glucose≥ 11.1 mmol/L (200 mg/dl); 3) post OGTT glucose ≥ 11.1 mmol/L (200 mg/dl); 4) HbA1c ≥6.5%; 5) current use of antidiabetic medications; or 6) self-reported physician diagnosed diabetes. Incident T2D was defined when a participant was free of diabetes at baseline but was identified as having T2D during annual follow-up interviews or visits. In the current analysis on the associations between serum metabolites and incident T2D, a total of 2,821 participants free of diabetes at baseline were included, with 367 incident T2D identified during approximately 6 years of follow-up.

Metabolite measurement. Serum metabolomic profiling was performed in a subsample of 3972 participants randomly selected from the whole cohort. On the basis of discoveryHD4 platform at Metabolon Inc. (Durham, NC), serum metabolites were measured using an untargeted liquid chromatography-mass spectrometry (LC-MS)-based metabolomic quantification protocol.[4,5] A total of 11 tryptophan related metabolites were captured by this platform, with a missing rate <20%, were included in this current study, including tryptophan, serotonin, 5 kynurenine metabolites (kynurenine, kynurenate, xanthurenate, quinolinate, picolinate) and 4 indole metabolites (indoleacetate, indolelactate, indolepropionate and indoxyl sulfate). Missing values of a metabolite were imputed using ½ of the minimum levels before analysis.

Genotyping and imputation. Genotyping was performed with an Illumina custom array (15041502 B3), which consists of the Illumina Omni 2.5M array (HumanOmni2.5-8v1-1) plus ~150k custom SNPs, with the quality control (QC) performed at HCHS/SOL Genetic Analysis Center.[6] Genome-wide imputation was carried out using the 1000 Genomes Project phase 3 worldwide reference panel, with SHAPEIT2 and IMPUTE2 software, as described previously. [6] An iterative procedure was used to simultaneously estimate principal components (PCs) reflecting population structure and Kinship coefficients measuring familial relatedness.[6] After standard quality control procedures, a total of 51,932,115 SNPs retained in the current genome-wide association study (GWAS) of metabolites among 3,933 SOL participants.

Diet assessment. Usual dietary intake was estimated using the National Cancer Institute methodology based on two 24-h dietary recalls and a food propensity questionnaire (FPQ), as described previously.[7] In brief, the first recall was administered through in-person interviews conducted at the time of the baseline visit, whereas the second was performed primarily via telephone approximately 30 days after the first interview. The FPQ, which was administered at the 1-year follow-up call, asked participants to report frequencies of foods eaten in the previous year. Foods and nutrients were analyzed using the multiple-pass methods of the Nutrition Data System for Research software (version 11) from the Nutrition Coordinating Center at University of Minnesota.

Fecal Metagenomic sequencing. Fecal samples were collected using a home-based self-collection kit from 3035 participants who enrolled in a gut microbiome ancillary study during 2015 to 2017.[8] Two fecal samples were self-collected by each participant, with one sample collected by a Whatman FTA card (GE Healthcare, Chicago, IL) and another one placed in a tube of RNA*later* (Invitrogen, Carlsbad, CA). Detailed procedures have been described elsewhere.[8] Metagenomics sequencing was performed on DNA extracted from fecal samples collected by FTA card using a novel shallow-coverage method of shotgun sequencing-based on Illumina platforms, boosted using inference from reference genomes of likely genomic content. Shallow shotgun data was left trimmed to remove low quality bases that had a PHRED quality score of 25 or less using prinseq-lite 0.20.4.[9] The quality controlled paired end data was then concatenated and aligned against the NCBI RefSeq representative prokaryotic genome collection (release 82[10]) using default SHOGUn[11] settings. Bowtie2[12] was used to build the SHOGUn assemblies. For the final analytical set, two Illumina sequencing runs were pooled. ICC function within the *psych* package[13] was used with R version 3.6.0 to assess the consistency of the runs prior to pooling. Samples (n=36) with a coverage depth less than 100k reads per sample were excluded. The coverage depth ranged from 100k to 8,945k reads per sample and median was 730k reads persample. To account for the variability in sequencing depth, centered log-ratio transformation was applied on the abundance data using R/microbiome. A total of 92 bacterial genera with average relative abundance≥0.01% were included in the current analysis.

The Atherosclerosis Risk in Communities (ARIC) Study

The ARIC study is a community-based longitudinal study that recruited 15,792 mostly black and white US participants aged 45-64 years from 4 study centers (suburban Minneapolis, MN; Washington County, MD; Jackson, MS; and Forsyth County, NC). The initial enrollment was carried out in 1987-1989 (Visit 1), and six subsequent follow-up studies have been completed, in 1990–1992 (visit 2), 1993–1995 (visit 3), 1996–1998 (visit 4), 2011–2013 (visit 5), 2016-2017 (visit 6) and 2018-2019 (visit 7). At baseline and follow-up visits, trained staff collected information on demographics, lifestyles, family and medical histories, and medication use, through in-person interviewers using structured questionnaires, and measured blood pressure and anthropometric traits. Annual (semi-annual since 2012) follow-up telephone interviews were conducted by trained interviewers to ascertain information on medication use and health status. Blood glucose was assayed using a modified hexokinase/glucose-6-phosphate dehydrogenase method. Other measured biomarkers were described in detail in previous publications.[14] All study participants signed informed consent, the protocol was approved by the institutional review boards at all participating centers, and procedures were followed according to the Declaration of Helsinki.

Diabetes Ascertainment. Participants who met at least one of the following criteria were defined as T2D cases**:** 1) fasting time >8 hours and fasting glucose ≥7.0 mmol/L (126 mg/dl); 2) fasting ≤8 hours and non-fasting glucose≥ 11.1 mmol/L (200 mg/dl); 3) current use of antidiabetic medications; or 4) self-reported physician diagnosed diabetes.[15] The incident T2D was defined as newly onset T2D from baseline through 31th December 2015. In the association analysis of metabolites with incident T2D, a total of 2721 participants (1433 black and 1288 non-Hispanic white participants) free of diabetes at baseline were included, with 1036 incident T2D identified during approximately 25 years of follow up.

Genotyping and imputation. Genotyping was performed using the Affymetrix 6.0 array and genome-wide imputation was performed with the 1000 Genomes Project Phase 3 reference panel using IMPUTE2.[16] Population structure was estimated using PC analysis. After standard quality control procedures, 10,714,779 SNPs in 1509 whites, and 18,704,142 SNPs in 1772 blacks were retained in GWAS of metabolites.

Diet assessment. Habitual dietary intake was assessed by a modified 66-item interviewer-administered food-frequency questionnaire (FFQ) at baseline exam.[17] Participants reported the frequency with which they consumed specific foods and beverages according to 9 predefined frequency categories, ranging from never or less than 1 time per month to 6 or more times per day (standard portion sizes were given as a reference for intake estimation; pictures and food models were shown to the participants by the interviewers).[18]

Metabolite measurement. Metabolomics profiling were measured from stored serum samples collected at baseline, by Metabolon (Durham, NC, USA), using an untargeted approach with an ultra-performance liquid chromatography system and a high-resolution mass spectrometer analyzer. Detailed information on the metabolomic assays in the ARIC were provided elsewhere.[15] In the present study, 7 tryptophan metabolites, including tryptophan, serotonin, kynurenine, indoleacetate, indolelactate, indolepropionate, and indoxyl sulfate, with a missing rate <20%, were included. Missing values of metabolites were imputed to the ½ of the lowest detectable values within each batch.

The Framingham Heart Study Offspring cohort (FHS)

The FHS is a community-based prospective cohort comprises 5,124 participants, and the current analyses include 3,799 participants who attended the 5th quadrennial examination cycle 1991-1995 (considered as the baseline examination).[19] At the 5th and subsequent quadrennial examination cycles, participant were inquired on medical histories, demographic and clinical characteristics, and lifestyle factors, underwent physician-administered physical examinations, and provided biological samples for routine laboratory tests. Blood biomarkers including blood lipids and glycemic traits, were measured by standard methods using blood samples collected at baseline.[20] All participants provided written informed consent and the study protocol was approved by the Boston University Medical Center Institutional Review Board.

T2D Ascertainment. Participants who met at least one of the following criteria were defined as T2D cases**:** 1) fasting time >8 hours and fasting glucose ≥7.0 mmol/L (126 mg/dl); 2) fasting ≤8 hours and non-fasting glucose≥ 11.1 mmol/L (200 mg/dl); or 3) current use of antidiabetic medications. Incident T2D was ascertained at every quadrennial examination. In the association analysis of circulating metabolites with incident, a total 1,424 participants free of diabetes at

baseline (exam 5) were included, with 218 incident T2D identified during approximately 20 years of follow up.

Metabolite measurement. Targeted metabolite profiling on baseline plasma samples was performed using liquid chromatography with tandem mass spectrometry (LC-MS/MS) at the Broad Institute (Cambridge, MA), which has been descried previously.[21,22] We excluded metabolites with a missing rate >20%. A total of 8 metabolites in the tryptophan metabolism pathway, including tryptophan, serotonin, kynurenine, kynurenate, xanthurenate, quinolinate, indolepropionate and indoxyl sulfate, were included in the current study. Missing values of a metabolite were imputed using ½ of the minimum levels before analysis.

Genotyping and imputation. The FHS samples were genotyped through the use of an Affymetrix 500K and a 50K Human Gene Focused Panel, followed by imputation according to HapMap CEU population release 22 using MACH.[23]

The Prevención con Dieta Mediterránea Study (PREDIMED)

The PREDIMED is a multi-center trial examining the efficacy of two Mediterranean diets over a control diet, for the primary prevention for cardiovascular disease. Starting from 2003, the PREDIMED study enrolled 7,447 participants aged 55-80 years and at high risk for cardiovascular disease (CVD) (i.e., presented at least three CVD risk factors or having diabetes), and allocate them into either a Mediterranean diet supplemented with extra virgin olive oil, a Mediterranean diet supplemented with mixed nuts, or a low-fat control diet arm, for an median of 4.8 years. The study populations were continued to be followed-up after the end of the trial upon consent. Institutional Review Boards at each recruitment center approved the study protocol, and all the participants provided written informed consent.

T2D ascertainment and case-cohort study design. T2D was a pre-defined secondary endpoint of the PREDIMED trial among participants without diabetes at baseline. Incident T2D was adjudicated through blind assessment by the Clinical Endpoint and Adjudication of Events Committee of PREDIMED, and an ad hoc panel of medical doctors, based on the criteria from the American Diabetes Association (including fasting plasma glucose ≥7.0 mmol/l or 2-h plasma glucose ≥11.1 mmol/l after a 75-g oral glucose load). Among the total 3,541 participants who were free of diabetes at baseline, 273 developed incident T2D during a median of 3.8 years of follow up. Among non-diabetic participants with qualified blood samples, a nested case-cohort study was designed for metabolomic profiling, which comprised a randomly selected sub-cohort with 694 participants (approximately 20% from all non-diabetic participants in the PREDIMED at baseline) and 251 incident T2D cases (53 overleaping participants between sub-cohort and cases). The detailed descriptions on the disease confirmation and the case-cohort design have been provided elsewhere.[24] After further excluding individuals missing >20% of named metabolites, 885 participants (including 248 incident T2D cases) were included in the current analysis.

Metabolite measurement. Metabolomics profiling of baseline plasma samples were measured by 3 complimentary liquid chromatography-tandem mass spectroscopy (LC-MS) methods at the Broad Institute (Cambridge, MA), according to protocol described previously.[24,25] Metabolites

with a missing rate of >20% were excluded and missing values of the remained metabolites were imputed by ½ minimum values. Five tryptophan metabolites, including tryptophan, kynurenine, quinolinic acid, indolepropionate and indoxyl sulfate, were included in this analysis.

The Women's Health Initiative (WHI)

The WHI, launched in 1993, is a long-term national study focused on the prevention of cardiovascular disease, breast and colorectal cancer, and osteoporotic fractures in postmenopausal women.[26] Postmenopausal women aged 50-79 years old were enrolled from 40 WHI clinical centers nationwide; 68,132 women entered one of a clinical trial including Hormone Therapy Trials (WHI-HT, including 27,347 women),[27] a Dietary Modification Trial, and a Calcium/Vitamin D Trial; 93,676 women who do not wanted to participant a clinical trial entered the Observational Study (WHI-OS), which is an prospective cohort study that aimed to examined relationship between lifestyle, health and risk factors and disease outcomes.[28] At the end of the initial study period in 2005, women who consented were continued to be followed up through WHI Extension Studies (2005-2010, 2010-2015, and 2015-2020). Data on health and lifestyle related information were collected at baseline in clinics by trained staff, and during follow up through telephone interviews and annual clinical visit. Blood and urine samples were collected at baseline and some follow-up visit. The study was approved by the institutional review boards and all participants provided written informed consent.

T2D ascertainment. T2D case was determined if a woman reported having a history of diabetes or using anti-diabetic medications (pills or shots) in visits/interviews. Incident T2D cases were identified among women free of diabetes at baseline during approximately 17 years of follow up. The primary study design for metabolomics profiling was based on a nested case-control study of coronary heart disease (CHD), which included a total of 472 incident CHD case and 472 matched controls from WHI-OS, and 315 incident CHD case and 315 matched controls from the WHI-HT placebo arms.[25] In the current study, we excluded women with diabetes at baseline or those with missing covariates, and a total of 1,392 participants (163 incident T2D cases) were included.

Metabolite measurement. Metabolomics was profiled using baseline plasma samples by the Broad Institute (Cambridge, MA), using 4 complimentary liquid chromatography-tandem mass spectroscopy (LC-MS) methods that has been previously described in detail.[25] We excluded samples and metabolites with a missing rate >20%, and 7 tryptophan metabolites, including tryptophan, serotonin, kynurenine, kynurenate, xanthurenate, indolepropionate and indoxyl sulfate, were included in the current analysis. Missing values of a metabolite were imputed using ½ of the minimum levels before analysis.

Statistical Analysis

Tryptophan metabolites and risk of T2D. Associations of 11 tryptophan metabolites and risk of T2D were first examined in the HCHS/SOL. Metabolites were inverse normal transformed before analysis (to ensure a standard normal distribution). Multivariable Cox regressions were used to estimate hazard ratios (HRs) and 95% confidence intervals (CIs) of incident T2D per standard deviation (SD) increment in metabolites, after adjusting for age, sex, smoking, alcohol

consumption, education, family income, fasting status, family history of diabetes, lipid-lowering medication use, anti-hypertensive medication use, and sampling weights (Model 1). In our exploratory analysis, we further adjusted for body mass index (BMI) and waist-hip-ratio (WHR) (Model 2), physical activity and Alternate Health Eating Index-2010 (AHEI) (Model 3), HDL-cholesterol, LDL-cholesterol and triglycerides (Model 4), or systolic and diastolic blood pressures (Model 5), based on Model 1.

We then sought for replications in the ARIC, WHI, FHS, and PREDIMED studies. Metabolites were inverse normal transformed before analysis in all replication studies. In the ARIC, WHI and FHS, similar multivariable Cox regressions were used to estimate HRs and 95% CIs of incident T2D per SD increment in metabolites, after adjusting for demographic, social, behavioral, and health-related factors, and other study-specific covariates. We also further adjusted for BMI and WHR (Model 2), physical activity and AHEI-2010 (Model 3), HDL-cholesterol, LDL-cholesterol and triglycerides (Model 4), or systolic and diastolic blood pressures (Model 5) in separated models. Due to a case-cohort study design in the PREDIMED, associations between metabolites with risk of T2D were examined using Cox regressions with Barlow weights[29] (i.e., inverse probability of the selection weights, to account for over selection of cases in a case-cohort study design and to mimic the association estimates from the whole cohort) with robust estimators. The basic model (Model 1) was stratified by study center and intervention arms, and were adjusted for age, sex, smoking, lipid-lowering medication use, anti-hypertensive medications, hypertension, dyslipidemia, and propensity scores (to account for probabilities for assigning to each intervention arms). In separate models, we further adjusted for BMI and WHR (Model 2), HDL-cholesterol, LDL-cholesterol and triglycerides (aforementioned Model 4), or systolic and diastolic blood pressures (aforementioned Model 5), and we did not adjusted for physical activity and AHEI-2010 as data were not available in the PREDIMED. Findings from the HCHS/SOL and replication studies were combined using a fixed-effect meta-analysis. Metabolites associated with risk of T2D at a consistent association direction across cohorts and with a combined *FDR*<0.05 were considered as statistically significant.

Tryptophan metabolites and cardiometabolic traits. In 2,821 HCHS/SOL participants free of diabetes at baseline, partial Spearman correlation analyses were used to examine correlation coefficients among 11 tryptophan metabolites, adjusting for age, sex and field center; and correlation coefficients of 11 tryptophan metabolites with age, fasting glucose, 2h glucose, HbA1c, fasting insulin, HOMA-IR, HOMA-B, LDL-cholesterol, HDL-cholesterol, total-cholesterol, triglycerides, systolic blood pressure, diastolic blood pressure, BMI and WHR, adjusting for age (if appropriate), sex, field center, lipid-lowering medication use, and anti-hypertensive medication use.

GWAS of tryptophan metabolites. GWAS of tryptophan metabolite levels (inverse-normalized to a standard distribution) in the HCHS/SOL were conducted using linear mixed effects regression models. Correlations between individuals were accounted for by incorporating covariance matrices corresponding to genetic relatedness (kinship), household, and census block group as random effects. The model also included field center, age, sex, the first five PCs to adjust for ancestry,[6] and sampling weights. [30] Analyses were performed using R 3.3.2. In the ARIC, serum metabolites were adjusted for age, sex, center and eGFR and then inverse-normalized to a standard distribution before analysis. GWAS of inverse-normalized tryptophan metabolite levels were conducted in African American and non-Hispanic participants separately, using liner regression models, adjusting for age, sex, center, eGFR, and the first 3 PCs. GWAS

of metabolites in the FHS has been published previously.[23] Briefly, circulating metabolite levels were adjusted for age and sex and normalized to a standard distribution. Association of genetic variants and metabolites was tested in 2,076 participants by mixed-effect linear regressions to accommodate pedigree data under an additive genetic model and were adjusted for the top PC. Analyses were performed using R and implemented by the *lmekin* function in the *kinship* package. We acquired the GWAS summary statistics from the previous publication[23] and converted the original genome build 36 to 37 to be consistent with other cohorts.

GWAS summery statistics from each cohort were first filtered, retaining SNPs with a minor allele frequency (MAF) ≥0.01 and imputation ratio ≥0.3. An inverse variance weighted fixed effect meta-analysis, implemented in METAL,[31] were used to combine GWAS results acquired from the HCHS/SOL, ARIC (African Americans and non-Hispanic whites separately), and FHS, for each of the metabolites. We used the PLINK clumping function (parameters: -clump-p1 5e-8 -clump-p2 1e-5 –clump-r2 0.1 –clump-kb 1000) to identify distinct loci and lead SNPs. Because our study included 11 tryptophan metabolites, genetic variants with a $P<4.5\times10^{-9}$ ($5.0\times10^{-8}/11$ metabolites) were considered as genome-wide significant.

Metabolites' genetic heritability causal relationship with T2D. GWAS summary statistics for T2D was acquired from the DIAGRAM consortium with 55,005 T2D cases and 400,308 controls of white populations.[32] Based on GWAS summary statistics for metabolites and T2D, we esitemated the SNP-based genetic heritability (h^2) at the observed scale for each of the tryptophan metabolites and their genetic correlations with T2D using linkage disequilibrium (LD) score regression.[33] We examined the potential causal relationship between metabolites and T2D using the latent causal variable (LCV) models.[34] LCV is a two-sample genetic causal inference methods based on GWAS summary statistics that could estimate causality while accounting for confounding from horizonal pleiotropy.[34] LCV esitimate the genetic causality proportion between 2 tratis when h^2 Z-score of of both traits were ≥7 (i.e., both traits have appreciable genetic heritability), otherwise the causal relationship would be overestimated. As such, we were able to test the causal relationships for serotonin, kynurenine, and indolepropionate with T2D. We were unable to test causal relationships with T2D for the 3 metabolites with $h^2=0$ (kynurenate, quinolinate, and indolelactate) and the other 5 metabolites with h^2 Z-score <7 in LCV (tryptophan, xanthurenate, picolinate, indoleacetate, indoxyl sulfate).

Dietary intake and serum metabolites. Associations of 10 major food groups with 11 tryptophan metabolites were examined among 3,938 HCHS/SOL participants using multivariable linear regressions, adjusting for age, sex, study center, smoking, education, family income, physical activity, and total energy intake. An *FDR*<0.05 was considered as statistically significant, controlling for multiple testing across 10 food groups and 11 metabolites (110 tests in total). We further included all 10 food groups in the multivariable linear regression model simultaneously to examine whether these food groups were associated with each of metabolites independent of each other. Subsequently, we examined associations of 3 macronutrients (carbohydrate, fat and protein) and other 2 nutrients of interest (fiber and tryptophan) with 11 tryptophan metabolites using multivariable linear regressions. In addition, we performed a medication analysis using multiple mediator models[35] was performed to examine the potential mediating effect of serum tryptophan metabolites on the association between the overall diet quality, measured by the

Alternate Healthy Eating Index 2010 (AHEI-2010),[36] and incident T2D in 2,821 HCHS/SOL participants free of diabetes at baseline.

Gut microbiota and indole metabolites. Linear regressions were applied to assess associations of 92 abundant gut bacterial genera (relative abundance≥0.01%) and 4 indole metabolites in 759 HCHS/SOL participants, adjusting for age, sex, study center, smoking, education, family income, physical activity, AHEI-2010, use of antibiotics, and use of probiotics. An *FDR*<0.05 was applied to define statistical significance, controlling for multiple testing across 92 bacterial genera and 4 metabolites (364 tests in total). The association results between 92 bacterial genera and indolepropionate were shown with taxanomic information in a circular format using GraPhlAn software.

Dietary fiber intake, gut microbiota and indolepropionate. Since dietary fiber intake was the top dietary factor positively associated with indolepropionate in our study and the identified indolepropionate-associated gut bacteria might be related to dietary fiber intake,[37,38] we then examined whether the association between dietary fiber intake and indolepropionate can be explained by gut bacteria. Among 2759 HCHS/SOL participants, we first examined associations of 21 indolepropionate-associated bacterial genera and dietary fiber intake using multivariate linear regressions, adjusting for the age, sex, study center, smoking, education, family income, physical activity, total energy intake, use of antibiotics, and use of probiotics. Subsequently, among 752 HCHS/SOL participants with gut microbiome, diet and metabolite data, we examined associations between dietary fiber intake and serum indolepropoinat levels using multivariate liner regressions, adjusting for the aforementioned covariates, and then further adjusting for all the 21 indolepropionate-associated bacterial genera or the 20 indolepropionate-associated bacterial genera (*Bifidobacterium* was not included because of its inverse association with dietary fiber intake). In addition, a mediation analysis using structural equation modelling[39] was conducted to examine whether these 20 indolepropionate-associated bacterial genera may partially explain the association between fiber intake and indolepropionate.

Host LCT genotype, gut Bifidobacterium and indolepropionate. Our analysis on dietary fiber intake, gut microbiome and indolepropionate suggested that gut *Bifidobacterium* might be associated with indolepropionate in a pathway not related to dietary fiber intake. Thus, we focused on *Bifidobacterium*, a common probiotic bacterial genus in the human gut, which has been related to a host functional *LCT* variant (SNP-rs4988235) and its interaction with milk intake.[40,41] This coincided with our novel GWAS findings at *LCT* associated with indolepropionate. We then examined associations of host *LCT* SNP rs4988235 genotype with milk intake in 12,531 HCHS/SOL participants, with gut *Bifidobacterium* abundance in 2,368 HCHS/SOL participants, and with serum indolepropionate levels in 3,933 HCHS/SOL participants, using linear regression models, adjusting for age, sex, center, smoking, education, family income, physical activity, AHEI-2010, and top 5 genomic PC. In 752 HCHS/SOL participants with genetic, diet, serum metabolite and gut microbiota data, we examined associations of *LCT* SNP rs4988235, milk intake and gut *Bifidobacterium* with serum indolepropionate mutually using a multivariable linear regression model. In 2,342 HCHS/SOL participants with genetic, diet and gut microbiome data, we examined the association between milk intake and gut *Bifidobacterium* abundance using multivariable linear regressions according to host *LCT* SNP rs4988235 genotype groups (AA+AG: lactase persistence; and GG, lactase non-persistence). In 3,899 HCHS/SOL participants with genetic, diet and serum metabolite data, we examined the association between milk intake and serum indolepropionate levels using

multivariable linear regressions according to host *LCT* SNP rs4988235 genotype groups (AA+AG: lactase persistence; and GG, lactase non-persistence). The effect modifications by host *LCT* genotype were examined by introducing an interaction term of SNP rs4988235 and milk intake in the multivariable linear regression models. Linear regression models were used to examine the association between milk intake and serum indolepropionate stratified by the host *LCT* SNP rs4988235 genotype and the interaction between milk intake and *LCT* SNP rs4988235 on serum indolepropionate among 3178 ARIC participants (1504 whites and 1674 African Americans), adjusting for age, sex, race and center.

Funding and Acknowledgment

The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago (HHSN268201300003I / N01-HC-65236 Northwestern University), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Neurological Disorders and Stroke, and NIH Institution-Office of Dietary Supplements. The authors thank the staff and participants of HCHS/SOL for their important contributions. A complete list of staff and investigators has been provided by Sorlie P., et al. in Ann Epidemiol. 2010 Aug;20: 642-649 and is also available on the study website http://www.cscc.unc.edu/hchs/.

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I), R01HL087641, R01HL059367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research. Metabolomics measurements were sponsored by the National Human Genome Research Institute (3U01HG004402-02S1).

The Framingham Heart Study Offspring cohort is supported by the National Heart, Lung and Blood Institute's FHS (contract no. N01-HC-25195 and HHSN268201500001I) and its contract with Affymetrix, Inc. for genotyping services (contract no. N02-HL-6-4278 and R01-HL081572).

The Women's Health Initiative program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C,

HHSN268201600003C, and HHSN268201600004C. The Metabolomic analysis in the WHI was funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, US Department of Health and Human Services through contract HHSN268201300008C. A list of WHI investigators is available online at https://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf.

The PREDIMED metabolomics studies were funded by National Institutes of Health (NIH) grants R01 HL118264 and R01 DK102896. The PREDIMED trial was funded by the Spanish Ministry of Health (Instituto de Salud Carlos III, The PREDIMED Network grant RD 06/0045, 2006– 13, coordinated by M.A. Martínez-González; and a previous network grant RTIC-G03/140, 2003–05, coordinated R. Estruch). Additional grants were received from the Ministerio de Economía y Competitividad-Fondo Europeo de Desarrollo Regional (Projects CNIC-06/2007, CIBER 06/03, PI06-1326, PI07-0954, PI11/02505, SAF2009-12304, and AGL2010– 22319-C03-03) and by the Generalitat Valenciana (PROMETEO17/2017, ACOMP2010-181, AP111/10, AP-042/11, ACOM2011/145, ACOMP/ 2012/190, ACOMP/2013/159 and ACOMP/213/165).

References

1. Lavange LM, Kalsbeek WD, Sorlie PD, et al. Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. Annals of epidemiology 2010;20:642-9.
2. Sorlie PD, Aviles-Santa LM, Wassertheil-Smoller S, et al. Design and implementation of the Hispanic Community Health Study/Study of Latinos. Annals of epidemiology 2010;20:629-41.
3. Qi Q, Strizich G, Merchant G, et al. Objectively Measured Sedentary Time and Cardiometabolic Biomarkers in US Hispanic/Latino Adults: The Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Circulation 2015;132:1560-9.
4. Zheng Y, Yu B, Alexander D, et al. Associations between metabolomic compounds and incident heart failure among African Americans: the ARIC Study. American journal of epidemiology 2013;178:534-42.
5. Chen GC, Chai JC, Yu B, et al. Serum sphingolipids and incident diabetes in a US population with high diabetes burden: the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Am J Clin Nutr 2020.
6. Conomos MP, Laurie CA, Stilp AM, et al. Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. Am J Hum Genet 2016;98:165-84.
7. Siega-Riz AM, Sotres-Alvarez D, Ayala GX, et al. Food-group and nutrient-density intakes by Hispanic and Latino backgrounds in the Hispanic Community Health Study/Study of Latinos. Am J Clin Nutr 2014;99:1487-98.
8. Kaplan RC, Wang Z, Usyk M, et al. Gut microbiome composition in the Hispanic Community Health Study/Study of Latinos is shaped by geographic relocation, environmental factors, and obesity. Genome Biol 2019;20:219.
9. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 2011;27:863-4.
10. Al-Ghalith G, Knights D. BURST enables optimal exhaustive DNA alignment for big data. DOI: doi org/105281/zenodo 2017;806850.
11. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, et al. Evaluating the information content of shallow shotgun metagenomics. Msystems 2018;3:e00069-18.
12. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods 2012;9:357.
13. Revelle W, Revelle MW. Package 'psych'. The Comprehensive R Archive Network 2015.
14. Rebholz CM, Yu B, Zheng Z, et al. Serum metabolomic profile of incident diabetes. Diabetologia 2018;61:1046-54.
15. Selvin E, Steffes MW, Zhu H, et al. Glycated Hemoglobin, Diabetes, and Cardiovascular Risk in Nondiabetic Adults. New England Journal of Medicine 2010;362:800-11.
16. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2009;2:73-80.
17. Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. Am J Epidemiol 1985;122:51-65.
18. Zheng Y, Yu B, Alexander D, Steffen LM, Boerwinkle E. Human metabolome associates with dietary intake habits among African Americans in the atherosclerosis risk in communities study. Am J Epidemiol 2014;179:1424-33.

19. Yu E, Ruiz-Canela M, Guasch-Ferre M, et al. Increases in Plasma Tryptophan Are Inversely Associated with Incident Cardiovascular Disease in the Prevencion con Dieta Mediterranea (PREDIMED) Study. J Nutr 2017;147:314-22.
20. Merino J, Leong A, Liu C-T, et al. Metabolomics insights into early type 2 diabetes pathogenesis and detection in individuals with normal fasting glucose. Diabetologia 2018;61:1315-24.
21. Rhee EP, Cheng S, Larson MG, et al. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. The Journal of Clinical Investigation 2011;121:1402-11.
22. Wang TJ, Larson MG, Vasan RS, et al. Metabolite profiles and the risk of developing diabetes. Nature Medicine 2011;17:448-53.
23. Rhee Eugene P, Ho Jennifer E, Chen M-H, et al. A Genome-wide Association Study of the Human Metabolome in a Community-Based Cohort. Cell Metabolism 2013;18:130-43.
24. Yu E, Papandreou C, Ruiz-Canela M, et al. Association of Tryptophan Metabolites with Incident Type 2 Diabetes in the PREDIMED Trial: A Case–Cohort Study. Clinical Chemistry 2018;64:1211-20.
25. Paynter Nina P, Balasubramanian R, Giulianini F, et al. Metabolic Predictors of Incident Coronary Heart Disease in Women. Circulation 2018;137:841-53.
26. Design of the Women's Health Initiative Clinical Trial and Observational Study. Controlled Clinical Trials 1998;19:61-109.
27. Manson JE, Chlebowski RT, Stefanick ML, et al. Menopausal Hormone Therapy and Health Outcomes During the Intervention and Extended Poststopping Phases of the Women's Health Initiative Randomized Trials. JAMA 2013;310:1353-68.
28. Langer RD, White E, Lewis CE, Kotchen JM, Hendrix SL, Trevisan M. The women's health initiative observational study: baseline characteristics of participants and reliability of baseline measures. Annals of Epidemiology 2003;13:S107-S21.
29. Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. J Clin Epidemiol 1999;52:1165-72.
30. Pfefferman D. The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review* 1993;Vol. 61:317-37
31. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;26:2190-1.
32. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature Genetics 2018;50:1505-13.
33. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015;47:291-5.
34. O'Connor LJ, Price AL. Distinguishing genetic correlation from causation across 52 diseases and complex traits. Nat Genet 2018;50:1728-34.
35. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods 2008;40:879-91.
36. Chiuve SE, Fung TT, Rimm EB, et al. Alternative dietary indices both strongly predict risk of chronic disease. J Nutr 2012;142:1009-18.
37. Makki K, Deehan EC, Walter J, Backhed F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. Cell Host Microbe 2018;23:705-15.

38. Deehan EC, Duar RM, Armet AM, Perez-Munoz ME, Jin M, Walter J. Modulation of the Gastrointestinal Microbiome with Nondigestible Fermentable Carbohydrates To Improve Human Health. Microbiol Spectr 2017;5.
39. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. J Stat Softw 2011;48.
40. Bonder MJ, Kurilshikov A, Tigchelaar EF, et al. The effect of host genetics on the gut microbiome. Nat Genet 2016;48:1407-12.
41. Goodrich JK, Davenport ER, Beaumont M, et al. Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe 2016;19:731-43.