nature portfolio

Corresponding author(s): Jun Li

Last updated by author(s): Oct 13, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	This study was conducted based on secondary data analysis of existing data from several well-established cohorts. We did not collect any new data.
Data analysis	Data processing, association analyses, and results presentations were mostly conducted in R 4.3.0 (https://www.r-project.org/) . Meta-analyses of results from metabolomic analyses across cohorts was conducted using R package metafor (https://cran.r-project.org/web/packages/metafor/index.html). Figures presenting associations results were generated using R packages ggplot2 (https://cran.r-project.org/web/packages/ggplot2/index.html), plotly (https://cran.r-project.org/web/packages/plotly/index.html), and networkD3 (https://cran.r-project.org/web/packages/networkD3/index.html). GWAS in WHI and CHS was conducted using RVTESTS (https://zhanxw.github.io/rvtests/), and in UKB, was conducted using BOLT-LMM (https://alkesgroup.broadinstitute.org/BOLT-LMM/BOLT-LMM_manual.html). Genome-wide meta-analysis were conducted using METAL (https://genome.sph.umich.edu/wiki/METAL). Plink (https://zzz.bwh.harvard.edu/plink/plink2.shtml) was used to identify independent genetic variants. SNPNexus was used for variants' gene annotation (https://www.snp-nexus.org/v4/). Manhattan plots were derived using R package CMplot (https://cran.r-project.org/web/packages/CMplot/index.html). Regional plots were draw with LocusZoom (https://my.locuszoom.org/). MetaCore was used for pathway analysis (https://portal.genego.com/). Genetic correlation was conducted using LDSC (https://github.com/bulik/ldsc). Colocalization analysis was conducted using R package coloc (https://cran.r-project.org/web/packages/coloc/index.html). Mendelian randomization analysis was conducted in R package MendelianRandomization (https://cran.r-project.org/web/packages/MendelianRandomization/index.html). Mediation analysis was conducted with R package CMAverse (https://bs1125.github.io/CMAverse/). Elastic net regression was conducted using R package glmnet (https://cran.r-project.org/web/packages/glmnet/index.html). The main code used to conduct this study is available on GitHub at https://github.com/JL-BWHlab/TOPMed_MWAS.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

This study did not generate any new/raw data but used existing data from multiple population-based cohorts, including the NHS, NHSII, HPFS, SOL, WHI, ARIC, FHS, MESA, and BPRHS cohorts, and the PREDIMED trial. Because of participant confidentiality and privacy concerns, these datasets are each governed by an approved data access policy, and are available upon request with formal applications submitted to the respective cohort committees, to adhere to data security and ethical considerations. Data for NHS/NHSII (detailed policies and access procedures https://nurseshealthstudy.org/; email to nhsaccess@channing.harvard.edu) HPFS (https://www.hsph.harvard.edu/hpfs/) are available upon written request; applications to use resources will be reviewed by an External Collaborators Committee for evaluation of the fit of the data for the proposed methodology, and verification that the proposed use meets the guidelines of the Ethics and Governance Framework and the consent that was provided by the participants. HCHS/SOL has established a process for the scientific community to apply for access to participant data and materials, with requests reviewed by the SOL Steering Committee (https://sites.cscc.unc.edu/hchs/). WHI metabolomic, genomic, and clinical data are available upon reasonable request to the WHI Publications and Presentations (P&P) Committee. Upon approval, requesters will be provided with details to access to the data (https://www.whi.org/propose-a-paper). Data access for FHS (also see detailed data policy at https://www.framinghamheartstudy.org/), MESA (https://www.mesa-nhlbi.org/), and ARIC (https://aric.cscc.unc.edu/aric9/) in the current study was approved by the TOPMed Publications and Presentations Steering Committees with data access provided by an approved project (#10065). GWAS summary statistics for metabolites from NHS/HPFS (doi: 10.1016/j.xcrm.2023.101085), SOL and ARIC (doi: 10.1016/j.ajhg.2020.09.003) and FHS (doi: 10.1016/j.cmet.2013.06.013) were each acquired from prior publications. For the PREDIMED trial (http://www.predimed.es/), due to the restrictions imposed by the Informed Consent and the Institutional Review Board, bona fide investigators interested in analyzing the PREDIMED dataset used for the present article may submit a brief proposal and statistical analysis plan to the corresponding author. Upon approval from the PREDIMED Steering Committee and Institutional Review Boards, the data will be made available to them using an onsite secure access data enclave. BPRHS data are available upon reasonable request, and information on data request can be found at https://www.uml.edu/research/uml-cph/research/bprhs/. The corresponding authors (Drs. Jun Li and Qibin Qi) will assist any person who is interested in these data to get data access from these cohorts.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study included male and female participants and thus sex was considered in the study design. Sex of participants was determined based on self-reported information. We did not stratify our analyses by sex given our focus on the associations in the entire study population. We controlled for sex as a covariate in the analyses whenever appropriate.
Reporting on race, ethnicity, or other socially relevant groupings	Race and ethnicity in our study was based on self-reported information. In each participating cohort, analyses were stratified by major racial and ethnic groups (non-Hispanic White, Hispanic/Latino, or African American individuals) if the sample size in each group was sufficient for analysis (e.g., >95% NHS/HPFS participants were non-Hispanic White therefore all NHS/HPFS participants were analyzed as one group; in MESA, analyses were done in two sets, non-Hispanic White individuals, and all participants who were not non-Hispanic White, to ensure sufficient power). Results from all study cohorts by major racial and ethnic groups were then combined, both for overall study samples and stratified by major racial and ethnic groups, using meta-analysis, and the heterogeneity in the associations was reported. Education and family income were adjusted for as covariates in association analyses in cohorts with available data. We did not consider other social factors. No stratified analysis was conducted on social factors. Detailed information on racial and ethnic groups considered in the analyses and analytical models in each cohort is presented in Supplementary Table S2.
Population characteristics	The detailed characteristics of study population from all participating cohorts is provided in Supplementary Table S1. Basic characteristics for participants included in the association analysis are: HCHS/SOL: N=2,821; 57.1% female; baseline mean age =45.0 years; baseline mean BMI =28.4kg/m2. NHS: N=6,890; 100% female; baseline mean age =56.5 years; baseline mean BMI =24.7kg/m2. NHSII: N=3,692; 100% female; baseline mean age =44.4 years; baseline mean BMI =24.3kg/m2. HPFS: N=2529; 0% female; baseline mean age =61.7 years; baseline mean BMI =25.4kg/m2. WHI: N=1,392; 100% female; baseline mean age =67.1 years; baseline mean BMI =26.8kg/m2. ARIC: N=2,721; 59.7% female; baseline mean age =52.0 years; baseline mean BMI =27.2kg/m2. FHS: N=1,424; 52.0% female; baseline mean age =55.0 years; baseline mean BMI =27.3kg/m2. MESA: N=902; 53.8% female; baseline mean age =60.1 years; baseline mean BMI =28.5kg/m2. BPRHS: N=378; 75.9% female; baseline mean age =55.7 years; baseline mean BMI =30.3kg/m2. PREDIMED: N=885; 61.0% female; baseline mean age =66.5 years; baseline mean BMI =29.7kg/m2.
Recruitment	The NHS, NHSII, and HPFS: The NHS is an ongoing prospective cohort study in the US initiated in 1976 and enrolled 121,701 female registered nurses aged 30-55 years; the NHSII was initiated in 1989 and recruited 116,429 female nurses aged 25-42 years; and the HPFS was established in 1986 and similarly enrolled 51,529 men in health professions nation-wide aged 40-75 years. Participants in all 3 cohorts completed a baseline questionnaire and have been followed up biennially. Diet was assessed every 4 years using a validated semi-quantitative food frequency questionnaire (FFQ). Blood samples were collected from sub-populations of the NHS (n=32,826) during 1989-1990, NHSII (n=29,611) during 1996-1999, and HPFS (n=18,018) during 1993-1995 (with >95% follow-up rate). Plasma metabolome was measured using samples from the first blood collection by multiple sub-studies. Our prospective analysis of plasma metabolites and incident T2D included 6,890 women from NHS, 3,692 women from NHSII, and 2,529 men from HPFS with qualified metabolomics data, and were free of diabetes, cardiovascular disease (CVD), and cancer at the time of blood draw.

HCHS/SOL is a prospective population-based study of 16,415 Hispanic/Latino adults aged 18-74 years living in 4 US metropolitan areas (Bronx, NY; Chicago, IL; Miami, FL; and San Diego, CA). Comprehensive interviews and clinical assessments were conducted at in-person clinic visits during 2008-2011 (Visit 1) and 2014-2017 (Visit 2), collecting information on demographics, behavioral factors, health status, and medical histories, as well as blood pressure, anthropometric traits, and standard clinical biomarkers. Annual follow-up telephone interviews were conducted to collect health information. Our analysis of incident T2D included 2,821 individuals with metabolomics data and were free of CVD, cancer and diabetes at Visit 1.

WHI, launched in 1993, is a nationwide study for chronic disease prevention in postmenopausal women. The study enrolled women aged 50-79 years old, including 68,132 women in one of three WHI clinical trials, and 93,676 women in the WHI Observational Study. At the end of the initial study period in 2005, women who consented were continued to be followed up through WHI Extension Studies. Data on health and lifestyle related information were collected at baseline 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 visits. Our study utilized metabolomics data from a nested case-control study of coronary heart disease (CHD); 1,392 women free of CVD, cancer, and diabetes, and were not in the active arms of the hormonal therapy trials, were included in analyses.

ARIC 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 from 1987-1989, and six subsequent follow-up studies were completed from 1990 to 2019. At baseline and follow-up visits, trained staff collected information on demographics, socioeconomic status, health behaviors, and medical histories, and measured blood pressure and anthropometric traits. Blood samples were collected for clinical biomarker measurement and future assays. Annual (semi-annual since 2012) follow-up telephone interviews were conducted to collect health information. Our prospective analyses of T2D risk included 1,288 White and 1,433 Black individuals with metabolomics data and were free of CVD, cancer, and diabetes at study baseline.

FHS is a community-based prospective cohort comprising 5,124 participants, and our study was based on 3,799 adults who attended the 5th quadrennial examination cycle during 1991-1995 (considered as the baseline in this study). Participants were inquired on demographic characteristics, clinical and lifestyle factors, and medical histories; underwent physician-administered physical examinations to collect weight, height, blood pressures, among others; and provided biological samples for routine laboratory tests. Blood biomarkers were measured using standard methods. Specifically, our prospective analysis of T2D risk included 1,424 individuals with metabolomics data and free of CVD, cancer, and diabetes at study baseline.

MESA, initiated in 2000, enrolled 6,814 adults aged 45-84 years old who self-identified as white (38%), African American (28%), Hispanic (22%), or Chinese (12%), and were free of clinically diagnosed CVD from 6 study sites (New York, NY; Baltimore, MD; Forsyth County, NC; Chicago, IL; St Paul, MI; and Los Angeles, CA). Participants' demographic and psychosocial characteristics, family histories, medical histories, and anthropometric measures was collected at baseline (2000-2002), and time-varying information such as physical activity, alcohol consumption, and smoking, was collected at baseline and updated during follow-up through questionnaires. Four follow-up visits were conducted at an average of 1.6, 3.1, 4.8, and 9.5 years after baseline with high retention rate. A total of 902 participants free of T2D, CVD, and cancer at baseline and had metabolomic data were included in our analysis.

BPRHS is a prospective study that enrolled 1,500 self-identified Puerto Rican adults, aged 45–75 y, living in the greater Boston area, at baseline. At baseline, and ~2 and ~6-year visits, bilingual, in-person interviews were conducted to collect information on demographic and psychosocial characteristics, medical and family histories, and lifestyle; trained staff measured anthropometrics and blood pressures. Plasma clinical biomarkers were measured using standard assays. Habitual diet was assessed with a semiquantitative FFQ designed and validated for the Puerto Rican population. Physical activity was assessed using a modified Paffenbarger questionnaire from the Harvard Alumni Activity Survey. We included in our analysis, 378 participants who had baseline plasma metabolomics data, were free of diabetes and CVD at baseline, and attended 2-y or 5-y visits.

PREDIMED is a 5-year multi-center trial examining the efficacy of two Mediterranean diets over control diet, for the primary prevention for CVD. The study initiated in 2003 and enrolled 7,447 participants aged 55-80 years who were at high risk of CVD. At baseline and then annually, participants' medical histories and risk factors were collected through standard questionnaires; and anthropometric traits were measured by trained personnel. Two propensity scores were estimated based on 30 baseline variables to account for the probability of assignment to each of the intervention groups. Among participants who were non-diabetic at baseline and had qualified blood samples, a nested case-cohort study was designed for metabolomic profiling, which comprised a randomly selected sub-cohort with 694 participants and 251 incident T2D cases identified through a median of 3.8 years follow-up (53 overleaping between sub-cohort and cases). After further excluding individuals with disqualified metabolomics data, 885 participants (including 248 incident T2D cases) were included in our analysis.

CHS is a NHLBI-sponsored prospective cohort study, originally enrolling 5,201 participants during 1989-1990, focusing on risk factors of CVD among older adults; a second cohort initiated in 1992-1993 and enrolled 678 predominantly black participants. Our study only used GWAS summary statistics for plasma metabolites for 263 CHS participants of European ancestry whose metabolomics was measured using EDTA plasma samples from year 7.

Ethics oversight

NHS, NHSII, and HPFS: The Institutional Review Boards at Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health approved the study, with participants' consent implied by the return of the questionnaires.

HCHS/SOL: The study was approved by the institutional review boards at all participating institutions, and all participants gave written informed consent.

WHI: The study was approved by the institutional review board of Brigham and Women's Hospital, and all participants provided written informed consent.

ARIC: 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.

FHS: All participants provided written informed consent, and the study protocol was approved by the Boston University Medical Campus Institutional Review Board.

MESA: Written informed consent was obtained from the participants, and the study was approved by institutional review boards at each site.

BPRHS: The study was approved by the institutional review boards at Tufts University, Northeastern University and the University of Massachusetts Lowell. All participants provided written informed consent.

PREDIMED: Institutional Review Boards at each recruitment center approved the study protocol, and all the participants provided written informed consent.

CHS: The study was approved by institutional review committees at each field center and individuals in the present analysis had available DNA and gave informed consent including consent to use of genetic information for the study of CVD.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The NHS, NHSII, and HPFS: Our prospective analysis of plasma metabolites and incident T2D included 6,890 women from NHS, 3,692 women from NHSII, and 2,529 men from HPFS with qualified metabolomics data, and were free of diabetes, cardiovascular disease (CVD), and cancer at the time of blood draw. Our GWAS analysis included a median of 6,610 participants (min=971 and max=8,054) across different metabolites, who have both qualified metabolomics data and quality controlled GWAS data. HCHS/SOL: Our analysis of incident T2D included 2,821 individuals with metabolomics data (at the time of data analysis) and were free of CVD, cancer and diabetes at Visit 1. Our GWAS analysis included 3933 participants who have both qualified metabolomics data and quality controlled GWAS data. WHI: Our study utilized metabolomics data from a nested case-control study of coronary heart disease (CHD); 1,392 women free of CVD, cancer, and diabetes, and were not in the active arms of the hormonal therapy trials, were included in analyses. Our GWAS analysis included 1256 participants who have both qualified metabolomics data and quality controlled GWAS data. ARIC: Our prospective analyses of T2D risk included 1,288 White and 1,433 Black individuals with metabolomics data and were free of CVD, cancer, and diabetes at study baseline. Our GWAS analysis included 1509 White and 1772 Black participants who have both qualified metabolomics data and quality controlled GWAS data. FHS: Our prospective analysis of T2D risk included 1,424 individuals with metabolomics data and free of CVD, cancer, and diabetes at study baseline. The GWAS analysis summary data for metabolites were acquired from publicly available resources for 1802 participants. MESA: A total of 902 participants free of T2D at baseline and had metabolomic data were included in our prospective analysis of T2D risk. BPRHS: We included in our analysis, 378 participants who had baseline plasma metabolomics data, were free of diabetes and CVD at baseline, and attended 2-y or 5-y visits. PREDIMED: Among participants who were non-diabetic at baseline and had qualified blood samples, a nested case-cohort study was designed for metabolomic profiling, which comprised a randomly selected sub-cohort with 694 participants and 251 incident T2D cases identified through a median of 3.8 years follow-up (53 overleaping between sub-cohort and cases; details provided elsewhere). After further excluding individuals with disqualified metabolomics data, 885 participants (including 248 incident T2D cases) were included in our analysis. CHS: Our study only used GWAS summary statistics for plasma metabolites for 263 CHS participants of European ancestry whose metabolomics was measured using EDTA plasma samples from year 7.
Data exclusions	The exclusion criteria are pre-established in our analysis proposals prior to data analyses. In prospective analysis of metabolites and incident T2D, we excluded, in each cohort, participants without qualified metabolomics data (based on data available at the time of data analysis), or participants who reported having diabetes, cardiovascular disease, or cancer at study baseline. In WHI, we further excluded women who were randomized to active arms of the original hormonal therapy trials. In our GWAS analysis of metabolites, we excluded, participants without qualified metabolomics data or quality controlled imputed GWAS data (based on data available at the time of data analysis).
Replication	This is a consortium pooling study, and we conducted meta-analysis of metabolome-wide association studies cross 10 prospective cohorts, to

Replication	identify novel metabolites associated with T2D risk. The large sample size ensures the statistical power for new discovery. Given the diversity of the study cohorts included, metabolites showing significant associations in meta-analysis generally show consistent results across cohorts. We additionally provided statistics on heterogeneity tests from the meta-analysis to inform association heterogeneity across cohorts. We conducted analyses to compare results across metabolomic platforms and major racial/ethnic groups, which yielded generally consistent results.
Randomization	Randomization is not applicable to our study, as our study is a population-based observational study, using existing data from several established cohorts.
Blinding	This is a population-based observational study. Data collection is appropriately blocked, with disease status of participants being blocked to investigators, or lab technicians, or research staff when specimen collection and laboratory assays (e.g., genomics and metabolomics) were performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*