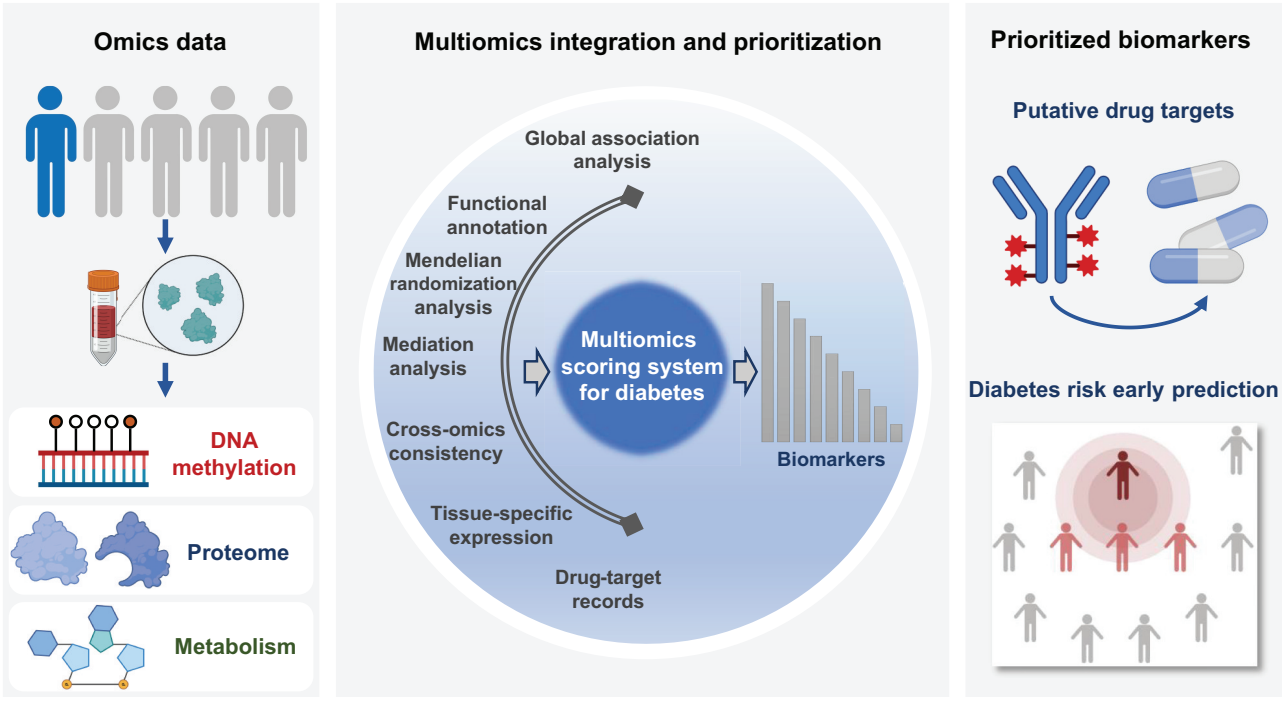
diabetes®

American Diabetes Association®

Multiomics Integration of Epigenetics, Proteomics, and Metabolomics Identifies Putative Drug Targets and Improves Early Prediction for Diabetes

Wenran Li, Yingyu Cheng, Aoyuan Cui, Mengyao Huang, Qingxia Huang, Qi Wang, Mingfeng Xia, Jiange Qiu, Qianqian Peng, Jiarui Li, Huating Li, Yong Wang, Geng Zong, Yan Zheng, Jiucun Wang, Xin Gao, Chen Ding, Huiru Tang, Bing-Hua Jiang, Li Jin, Yu Li, and Sijia Wang

Diabetes 2025;74(12):2418–2431 | https://doi.org/10.2337/db25-0354



Multiomics Integration of Epigenetics, Proteomics, and Metabolomics Identifies Putative Drug Targets and Improves Early Prediction for Diabetes

Wenran Li,[1] Yingyu Cheng,[1] Aoyuan Cui,[2] Mengyao Huang,[2] Qingxia Huang,[3] Qi Wang,[3] Mingfeng Xia,[4] Jiange Qiu,[5] Qianqian Peng,[1] Jiarui Li,[1] Huating Li,[6] Yong Wang,[7] Geng Zong,[2] Yan Zheng,[8] Jiucun Wang,[8,9] Xin Gao,[4,8] Chen Ding,[10] Huiru Tang,[3] Bing-Hua Jiang,[5] Li Jin,[8,9] Yu Li,[2] and Sijia Wang[1]

Diabetes 2025;74:2418–2431 | https://doi.org/10.2337/db25-0354

Diabetes holds significant social importance due to its high incidence rate and multitude of associated complications. The identification of diabetes biomarkers and the understanding of the intricate biological mechanisms underlying diabetes are crucial for the early diagnosis and treatment of diabetes. In this study, we conducted comprehensive omics profiling of CpGs, plasma proteins, and serum metabolites in an National Survey of Physical Traits (NSPT) cohort of 3,451 individuals, among whom 293 were patients with diabetes. Global association analysis identified 175 CpGs, 29 proteins, and 93 metabolites significantly linked to diabetes, among which 43 CpGs and 25 metabolites were validated in an independent cohort comprising 532 individuals. Mendelian randomization and mediation analysis identified 20 causal biomarkers and 190 signaling pathways linking biomarkers from different layers. By integrating the cross-omics evidence, we provide a list of putative causal biomarkers of diabetes to serve as a valuable resource for the diabetes community. Cross-omics integration prioritized biomarkers for therapeutic targeting, highlighting COLEC11 as an example of a potential target and whose function was further validated in vitro. The early-prediction model using the prioritized biomarkers improved the area under the receiver operating characteristic curve by 27.5% compared with the baseline model, using clinical features alone. Our findings provide a comprehensive list of prioritized multiomics biomarkers and elucidate specific signaling pathways in diabetes, contributing significantly to the selection of therapeutic target and the understanding of diabetes pathophysiology.

ARTICLE HIGHLIGHTS

- A total of 175 CpGs, 29 proteins, and 93 metabolites were identified as associated with diabetes, among which 43 CpGs and 25 metabolites were validated in an independent cohort.
- Causal and mediation analyses revealed 20 biomarkers and 190 signaling pathways involved in diabetes development.
- The integrative multiomics prioritization provides the community with an ordered list of diabetes biomarkers. We experimentally validated one of the prioritized proteins, COLEC11, and demonstrated its involvement in lipid metabolism.
- Our findings prioritize potential therapeutic targets and demonstrate that integrating multiomics biomarkers improves diabetes risk prediction beyond traditional clinical models.

Diabetes is recognized as one of the most significant chronic diseases with numerous complications, including stroke,

[1]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China
[2]CAS Key Laboratory of Nutrition, Metabolism and Food Safety, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China
[3]State Key Laboratory of Genetic Engineering, School of Life Sciences, Human Phenome Institute, Zhangjiang Fudan International Innovation Center, Metabonomics and Systems Biology Laboratory at Shanghai International Centre for Molecular Phenomics, Zhongshan Hospital, Fudan University, Shanghai, China
[4]Department of Endocrinology and Metabolism, Zhongshan Hospital and Fudan Institute for Metabolic Diseases, Fudan University, Shanghai, China
[5]Academy of Medical Science, Zhengzhou University, Zhengzhou, Henan, China
[6]Department of Endocrinology and Metabolism, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China
[7]Center for Excellence in Mathematics and Systems Science, National Center for Mathematics and Interdisciplinary Sciences, Hua Loo-Keng Center for Mathematical Sciences, Key Laboratory of Management, Decision and Information Systems, Chinese Academy of Sciences, Beijing, China

neuropathy, retinopathy, kidney injuries, and cardiovascular disease (1). Understanding the mechanisms of diabetes is essential to facilitate individualized prevention and treatment strategies, thereby reducing the incidence and societal burden of the disease. Despite the extensive research on diabetes (2), many studies have been limited to already identified biomarkers, without discovering new potential biomarkers and exploring the relationships between different omics layers in the regulatory mechanism of diabetes. This limitation hampers progress in diabetes research and prediction.

Advancements in sequencing and mass spectrometry (MS) technologies have enabled genome-wide identification of omics biomarkers for diabetes. Epigenome-wide association studies (EWAS) in diverse populations have identified numerous CpG sites as epigenomic markers of diabetes (3–6). Proteome-wide association studies (PWAS) of cohorts from Sweden and South Korea have highlighted proteins involved in lipid transport and regulation of lipoprotein levels (7–10). Metabolome-wide association studies (MWAS) have revealed significant associations between diabetes and metabolites such as hexose, lipids, and amino acids (11,12). However, relying solely on single omics data sets to identify biomarkers may not suffice to fully elucidate the signaling pathways and underlying mechanisms of diabetes pathophysiology.

In this study, we systematically measured DNA methylation, protein expression, and metabolite levels, conducting a multiomics integration analysis for diabetes in the National Survey of Physical Traits (NSPT) cohort of 3,451 individuals from the Chinese population (Fig. 1). We aim to discover and prioritize novel molecular biomarkers of diabetes across different omics and investigate the regulatory processes among those biomarkers. As a result of our multiomics integration analysis, we provide a prioritized list of novel biomarkers and demonstrate that this resource can aid in the selection of therapeutic targets and enhance the accuracy of clinical diagnoses.

RESEARCH DESIGN AND METHODS

Study Design

The NSPT Cohort

Samples in the NSPT cohort included 3,557 Chinese individuals recruited as volunteers from three regional districts in China: Zhengzhou, Taizhou, and Nanning. The details of the preprocessing of genetics data and DNA methylation data have been described (13). To investigate the associations between diabetes and different omics layers, we limited the NSPT cohort to participants whose data sets contained *1*) all main omics measurements (i.e., DNA methylation, proteomics, metabolomics, and cardiovascular traits) from the same first blood sample collection; *2*) a fasting blood glucose (FBG) measurement; and *3*) genetic information (used as covariates). Diabetes was defined as individuals with self-reported diabetes or FBG >7 mmol/L. After excluding individuals without an FBG measurement or basic information like BMI, 3,451 individuals remained for analysis (n = 1,281 male and 2,170 female participants; age range, 18–83 years; mean age ± SD = 49.99 ± 12.74 years), among whom 293 had diabetes. Demographic and clinical characteristics of the participants are described in Supplementary Table 1.

The Changfeng Cohort

Samples in the Changfeng (CF) cohort included 532 Chinese individuals recruited from Zhongshan Hospital, Fudan University, Shanghai, China (n = 294 male and 238 female participants; age range, 47–80 years; mean age ± SD = 61.57 ± 7.64 years), 71 of whom were patients with diabetes (Supplementary Table 1). The details of the preprocessing of DNA methylation data and metabolites data have been published (14,15).

The Zhongyuan Cohort

Samples in the Zhongyuan cohort included 605 Chinese individuals recruited in the follow-up project of the NSPT cohort, among whom 14 developed new-onset diabetes (Supplementary Table 1). Fasting blood samples were collected after an overnight fast of at least 12 h, and all samples were stored at −80°C. Serum total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides, FBG, and 2-h postchallenge glucose levels were measured using an automated bioanalyzer following standard protocols.

Plasma Protein Profiling

Proteomic profiling of plasma samples was performed using liquid chromatography–MS with data-independent

[8]State Key Laboratory of Genetic Engineering, School of Life Sciences, Human Phenome Institute, Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai, China
[9]Research Unit of Dissecting the Population Genetics and Developing New Technologies for Treatment and Prevention of Skin Phenotypes and Dermatological Diseases, Chinese Academy of Medical Sciences, Shanghai, China
[10]Department of Urology, Fudan University Shanghai Cancer Center, State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Institute of Biomedical Sciences, and Human Phenome Institute, Fudan University, Shanghai, China

Corresponding author: Huiru Tang, huiru_tang@fudan.edu.cn, Bing-Hua Jiang, binghjiang@zzu.edu.cn, Li Jin, lijin@fudan.edu.cn, Yu Li, liyu@sinh.ac.cn, or Sijia Wang, wangsijia@sinh.ac.cn

Received 9 April 2025 and accepted 9 August 2025

This article contains supplementary material online at https://doi.org/10.2337/figshare.29881889.

W.L., Y.C., and A.C. contributed equally to this study.

See accompanying article, p. 2214.

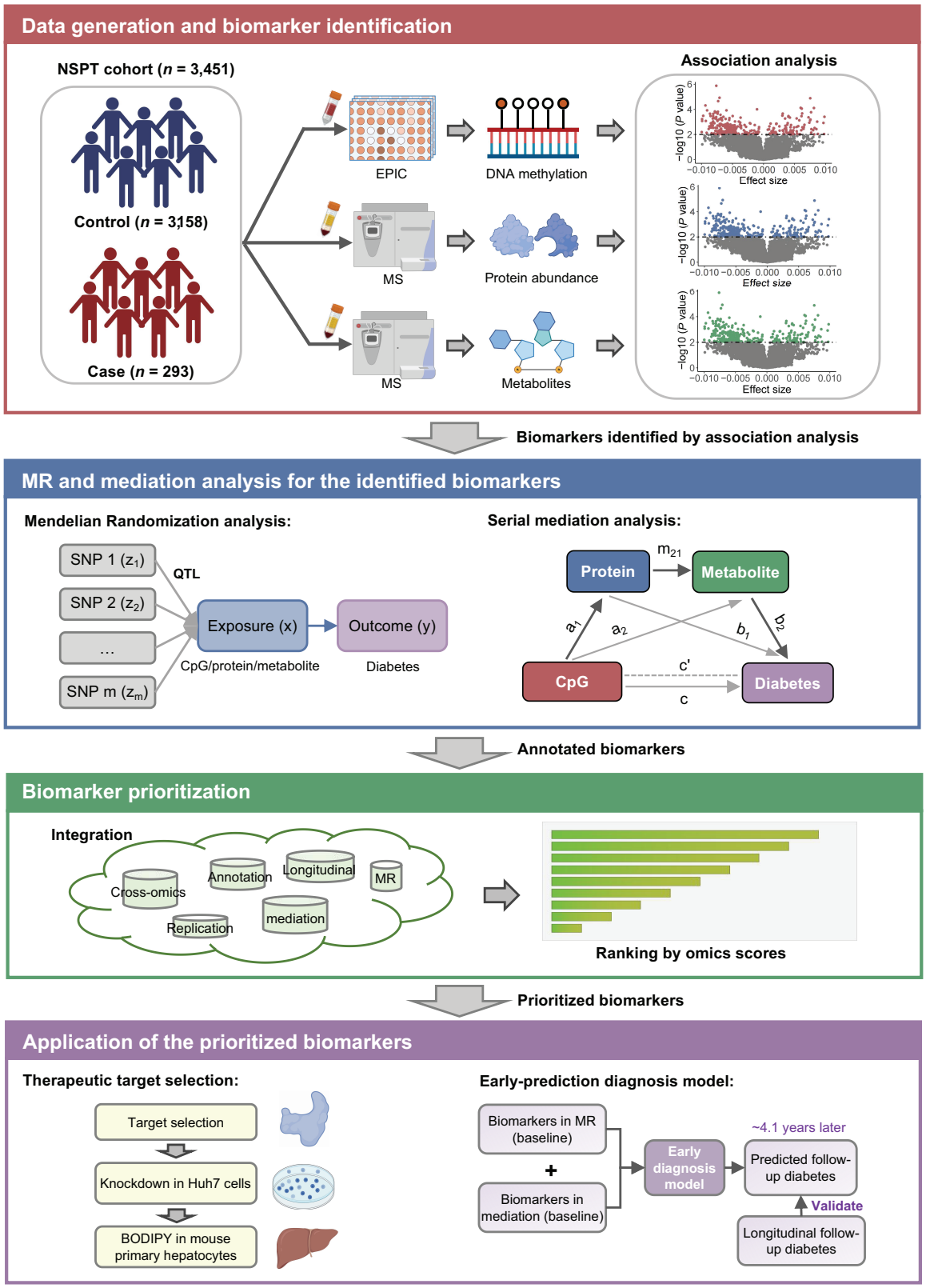


Figure 1—Overview of the study design. First, in the red box, the DNA methylation, protein abundance, and metabolites of individuals in the NSPT cohort were respectively measured by Illumina Methylation EPIC BeadChip (850 K) and MS. Global association analysis was conducted to identify diabetes-associated omics biomarkers. Then, as shown in the blue box, MR and mediation analyses were performed for the identified biomarkers to investigate causal biomarkers and signaling pathways across multiomics. Next (green box), the identified biomarkers were prioritized based on the cross-omics integration process, considering various evidences of each biomarker. Subsequently, as shown in the purple box, we validated the function of one of the putative therapeutic targets selected based on the signaling pathways and prioritization results (left). Finally, we showed the added performance of the prioritized biomarkers in the prediction of future diabetes (right).

acquisition (DIA) mode on a Q Exactive HF-X Orbitrap mass spectrometer. Peptides were separated using a C18 column and analyzed with a 30-segment DIA method. Data were processed via Firmiana, using FragPipe with MSFragger for DIA and Mascot for data-driven attribution (DDA), referencing the UniProt human database. Spectra libraries were built from DDA data using SpectraST. DIA data were quantified with DIA-NN and normalized using the fraction of total approach based on intensity-based absolute quantification values. Quality control involved pooled plasma samples analyzed every 20 runs. Details of the plasma protein profiling can be found in Supplementary Texts.

Serum Metabolites Profiling

Serum metabolomics profiling was performed using a 600 MHz AVANCE III nuclear magnetic resonance (NMR) spectrometer with broadband inverse probe. Samples were prepared by mixing serum with phosphate buffer and analyzed at 310 K. Two ^{1}H NMR spectra (NOESYGPPR1D and LEDBPGPPR2S1D) were acquired, and metabolites were quantified using Bruker's B.I.LISA and B.I.Quant-PS platforms. A total of 351 parameters were obtained, including 112 lipoprotein traits and 41 low-molecular-weight metabolites, along with derived functional measures such as lipid ratios and glycoproteins. Details of the serum metabolites profiling can be found in Supplementary Texts.

Statistical Analysis

EWAS Analysis

EWAS were performed in the NSPT cohort using the R package *limma* (16), applying linear regression with diabetes status as the independent variable and methylation level of 811,876 CpGs as the dependent variable. The model adjusted for confounding factors including sex, age, BMI, location, batch information, cell fractions, and the top 10 single nucleotide polymorphism (SNP) principal components. The cell fractions were predicted based on the methylation data using the R package *EpiDISH* (17). Significant CpG-diabetes associations were identified at $P < 1 \times 10^{-6}$, a commonly used threshold in EWAS. These associations were further replicated in the CF cohort ($P < 0.05$, consistent effect direction). Replicated CpGs were used for downstream functional analysis. Subgroup analyses by sex and BMI were performed to explore whether the associations between molecular biomarkers and diabetes differed across subpopulations (Supplementary Texts and Supplementary Fig. 1).

PWAS Analysis

A total of 12,882 plasma proteins were measured in the NSPT cohort using MS, with 6,528 proteins remaining after quality control. Proteins were standardized to *z* scores using the scale function. PWAS were conducted with the R package *limma*, using protein expression levels as the outcome and diabetes status as the independent variable, adjusting for sex, age, location, and BMI. Significant protein-diabetes associations were identified at a false discovery rate (FDR) <0.05, calculated using the Benjamini-Hochberg procedure to control for multiple testing. Enrichment analysis was performed using the R package *clusterProfiler*, version 4.8.116, checking for enrichment in the Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome pathways.

MWAS Analysis

A total of 351 metabolites were measured in the NSPT cohort, and 182 metabolites remained after quality control by removing ratio data and those with >10% missing values. MWAS were conducted using the R package *limma*, with adjustments for sex, age, location, and BMI. Significant associations with diabetes were identified using an FDR threshold of <0.05, with FDR calculated using the Benjamini-Hochberg method. Identified metabolites were replicated in the CF cohort, where those showing consistent effect directions and achieving $P < 0.05$ were considered successfully replicated.

Mendelian Randomization Analysis

The methylation quantitative trait loci (QTL) data were collected from Peng et al. (13). Protein QTL and metabolomic QLT were calculated using a linear mixed model under an additive genetic model in the GCTA tool (version 1.94.2) (18). Metabolite GWAS were performed using a linear mixed model under an additive genetic model in PLINK (version 2.0) (19). SNPs with minor allele frequency <5% and imputation quality score <0.6 were excluded. All analyses were adjusted for age, sex, and the top five genetic ancestry principal components. A total of 43 CpGs, 29 proteins, and 25 metabolites were included, with 6,783 methylation QTLs, 3,355 protein QTLs, and 12,248 metabolomic QTLs selected as instruments. Mendelian randomization (MR) analysis was performed with the R package *Two-Sample MR* (20), using the inverse variance weighted method to estimate causal effects. Heterogeneity was evaluated using the Cochran *Q* statistic, and outlier instruments were identified. Inverse variance weighted results with $P < 0.05$ were considered statistically significant. One-sample MR analysis was further performed to assess the robustness of causal inferences (Supplementary Texts and Supplementary Fig. 2).

Functional Enrichment Analysis

We first checked the enrichment of the identified CpGs in different traits recorded by EWAS Atlas. The CpGs were then mapped to their corresponding genes based on location information using the R package *annotatr* (21). And to explore potential functional roles, we performed pathway and biological process enrichment analyses on diabetes-associated CpGs and proteins using GO (22), KEGG (23), Reactome (24), and WiKi (25) via clusterProfiler, version 4.8.116 (26). *P* values for the enrichment analyses were calculated using two-tailed Fisher exact tests. Details of the enrichment analysis are provided in Supplementary Texts.

Mediation Analysis

Mediation analysis was performed using the R package *mediation* (27). CpGs were treated as independent variables, diabetes status as the dependent variable, and proteins and metabolites as mediators. The direct effect of CpGs on diabetes (c path), the effect of CpGs on the mediator (a path), and the effect of the mediator on diabetes controlling for CpGs (b path) were estimated. The indirect effect (a × b) and the adjusted direct effect (c′) were evaluated. Mediation was considered present if the indirect effect was significant and c′ was attenuated. Serial mediation analysis was conducted using the R package *lavaan* (28) to assess pathways from CpGs to diabetes through proteins and metabolites. Indirect and overall mediation effects were estimated, and bootstrapping was applied to generate CIs. Mediation results with $P < 0.05$ were included in the following analysis.

Functional Consistency of Biomarkers From Different Omics Layers

To demonstrate the functional consistency across different omics layers, we selected diabetes-related CpGs and proteins based on MR and mediation analyses for integrative functional analysis. Enrichment was performed using GO, KEGG, and Reactome to examine shared biological roles. This revealed 72 pathways jointly regulated by 40 CpGs and 10 proteins, highlighting coordinated mechanisms in diabetes pathogenesis. Key pathways were visualized to illustrate cross-omics functional consistency.

Tissue-Specific Gene Expression of the Identified Proteins

We collected the gene expression profiles across 60 tissues from the FANTOM5 consortium and calculated the tissue-specificity scores to represent the tissue-specific expression levels of diabetes-related proteins. The tissue-specificity score was defined as the difference between the tissue-specific expression and the average expression, formulated as:

$$s_{it} = \frac{x_{it} - \left(\sum_1^T x_{it}/T\right)}{\sum_1^T x_{it}/T}$$

where x_{it} is the expression of ith protein in tth tissue, and T is the total number of tissues related with diabetes ($T = 60$ in our case). For each gene or protein, we checked the expression level and tissue-specificity scores in diabetes-associated tissues, including adipose, cerebellum, liver, pancreas, and skeletal muscle, as reported by Roden and Shulman (29,30).

Drug Target Enrichment Analysis

We used data from the DrugBank Database and the Therapeutic Target Database, organized by GREP software, to identify the drugs corresponding to the diabetes-related CpGs and proteins we discovered. Then, we mapped the identified drugs to specific disease categories based on the ICD-10 in the Therapeutic Target Database. To illustrate the translational potential of these biomarkers, we finally constructed a gene–drug–disease network by linking CpGs, proteins, drugs, and their related disease categories, using a Sankey diagram.

Cross-Omics Integration Analysis

A systematic omics score was developed to assess the relevance of biomarkers to diabetes by integrating evidence across epigenomics, proteomics, and metabolomics. Scores were assigned by summing points based on seven criteria: *1*) significance in global association analysis, *2*) causal inference by MR analysis, *3*) validation through longitudinal data, *4*) mediation via signaling pathways, *5*) cross-omics functional consistency, *6*) tissue-specific expression in diabetes-relevant tissues, and *7*) drug targeting. The final score reflected the overall importance of each biomarker in diabetes.

In Vitro Experiments

We conducted in vitro experiments using human Huh7 cells and primary mouse hepatocytes to investigate the role of COLEC11. Huh7 cells were cultured in DMEM and transfected with siCOLEC11 or control siRNA for 24 h, followed by serum starvation and treatment with high concentrations of glucose (30 mmol/L) and insulin (100 nmol/L) for another 24 h. Primary mouse hepatocytes were isolated via collagenase digestion and cultured on collagen-coated plates. Various assays were performed, including boron dipyrromethene (BODIPY) staining to assess lipid accumulation, immunoblotting to analyze protein expression, and qRT-PCR to measure gene expression levels. Statistical analyses were conducted using the two-tailed Student *t* test, with significance defined as $P < 0.05$. The experimental details are presented in Supplementary Texts.

Construction of the Early-Prediction Model

A diabetes prediction model was developed using top-ranked omics biomarkers in the CF longitudinal cohort, excluding individuals with baseline diabetes or abnormal FBG values. A fivefold cross-validation with a support vector machine algorithm was applied for feature selection and prediction. Model performance was assessed using area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve, and accuracy. The model was validated in the external Zhongyuan cohort. We compared our model with two established omics-based diabetes prediction models: the Episcore model (31), which predicts the 10-year risk of type 2 diabetes (T2D) using blood DNA methylation data, and the Kim et al. (32) model, which estimates T2D risk based on eight differentially methylated sites.

Data and Resource Availability

The proteomics and metabolomics data have been deposited in https://www.biosino.org/node/run/detail/OER475675. The methylation data were previously reported by Peng et al. (13). Summary statistics of global association analysis were uploaded in https://www.biosino.org/node/analysis/

detail/OEZ00021308. Other data generated or analyzed during this study are included in the supplementary files.

RESULTS

Discovery and Replication of Multiomics Biomarkers for Diabetes

EWAS analysis identified 175 CpGs significantly associated with diabetes in the NSPT cohort ($P < 1 \times 10^{-6}$), with 43 subsequently replicated in the CF cohort ($P < 0.05$, consistent direction of effects) (Fig. 2*A* and Supplementary Table 2). Among these, 8 CpGs had been previously reported in the EWAS studies of T2D (33–36), and the other 35 were newly identified as biomarkers of diabetes. These CpGs were enriched in positive regulation of steroid biosynthetic process and cholesterol metabolism (Fig. 2*D*, Supplementary Fig. 3, and Supplementary Table 3). In addition, they showed enrichment in DNase I hypersensitive sites specific to monocytes (Supplementary Texts), suggesting that these CpGs predominantly reflect epigenetic variation in innate immune cells. PWAS revealed 29 proteins significantly associated with diabetes (FDR < 0.05) (Fig. 2*B* and Supplementary Table 4), which were mainly enriched in cholesterol transport, lipid localization, and cholesterol metabolism (Fig. 2*E*, Supplementary Fig. 4, Supplementary Table 3). Among the 93 diabetes-associated metabolites identified in the NSPT cohort, 25 were replicated in the CF cohort (Fig. 2*C* and Supplementary Table 5). Of the 25 diabetes-associated metabolites, 68% were lipoproteins and subfractions linked to insulin secretion; the rest were low-molecular-weight metabolites, including amino acids and glycolysis intermediates involved in glucose metabolism (Fig. 2*F*). Consistent effect sizes were observed between the NSPT and CF cohorts for both CpGs and metabolites (Supplementary Fig. 5).

MR and Mediation Analysis

MR and mediation analysis were conducted to explore causal relationships between different omics layers and diabetes (Figs. 3 and 4). MR analysis revealed 11 CpGs, 5 proteins, and 4 metabolites causally linked to diabetes (Fig. 3*B–E* and Supplementary Table 6). Furthermore, a cross-lagged causal model using longitudinal data from the CF cohort validated the causal effect of two CpGs and one metabolite on FBG changes (Fig. 3*F–H* and Supplementary Texts). Mediation analysis showed that the association between CpGs and diabetes was mediated by 13 proteins and 19 metabolites, and the associations between proteins and diabetes were mediated by 5 metabolites (Fig. 4*B–D* and Supplementary Tables 7–9). Serial mediation analysis further identified 190 molecular signaling pathways, reflecting the regulatory processes from CpGs, protein, and metabolites to diabetes (Fig. 4*E* and *F* and Supplementary Table 10). For example, the CpG cg11024682 in SREBF1, associated with diabetes, was shown to mediate its effect via specific lipoproteins and glucose, illustrating a pathway from epigenetic regulation to diabetes via lipid metabolism (Supplementary Fig. 6).

Prioritization of Diabetes Biomarkers

To provide a practical resource for the community, we integrated multiple lines of omics evidence to prioritize the identified biomarkers (Supplementary Fig. 7). Biomarkers were scored based on their significance in the association analysis, causality in MR and mediation analysis, functional consistency across omics (Fig. 5*A*), expression specificity in diabetes-related tissues (Fig. 5*B–D*, Supplementary Texts, and Supplementary Figs. 8 and 9), and related drug targets (Fig. 5*E* and Supplementary Texts). The integrated omics score, reflecting the relevance of biomarkers to diabetes, was used for prioritization. Totally, 64 biomarkers were assigned with omics scores >5 (Supplementary Table 11) and marked as key biomarkers of diabetes across different omics. Among them, 40 had been previously reported in association with diabetes, obesity, or metabolic syndrome; the other 24 are novel, to our knowledge (Supplementary Texts and Supplementary Table 12).

COLEC11, a Potential Therapeutic Target for Diabetes

COLEC11, a novel biomarker identified in our study, was ranked among the top candidates and found to be a downstream protein of previously reported diabetes biomarkers (namely *TXNIP*, *SREBF1*, and *PTPRT*) within the signaling pathways (Fig. 4*F*). Its strong specificity in liver tissue highlights its potential role in diabetes regulation, leading us to propose COLEC11 as a promising therapeutic target. To further validate its biological function, a series of cell experiments were conducted.

First, in the mouse model, *Colec11* expression was upregulated in the livers of hyperglycemic *ob/ob* mice and high-fat, high-sucrose diet–induced mice (Fig. 6*A*). Then, the knockdown of COLEC11 in Huh7 hepatocytes inhibited the expression of lipogenesis-related genes (*FAS* and *SCD1*), which were stimulated by high glucose and insulin concentrations, indicating COLEC11's role in glucose and lipid metabolism (Fig. 6*B–D*). Furthermore, knockdown of COLEC11 improved the insulin-induced activation of the insulin-signaling pathway, as indicated by the increased phosphorylation of insulin receptor and Akt (Fig. 6*E–F*). Consistently, *Colec11* deficiency remarkably prevented the high-concentration glucose- and insulin-stimulated lipid accumulation in mouse primary hepatocytes (Fig. 6*G*). Taken together, these findings demonstrate that COLEC11 deficiency promotes insulin sensitivity and prevents lipid accumulation in hepatocytes.

Prediction of the Diabetes Onset

An early-prediction model for diabetes was developed using biomarkers from clinical, epigenetic, and metabolic omics layers in the CF longitudinal cohort; proteomic data were not included, because the CF cohort data do not contain

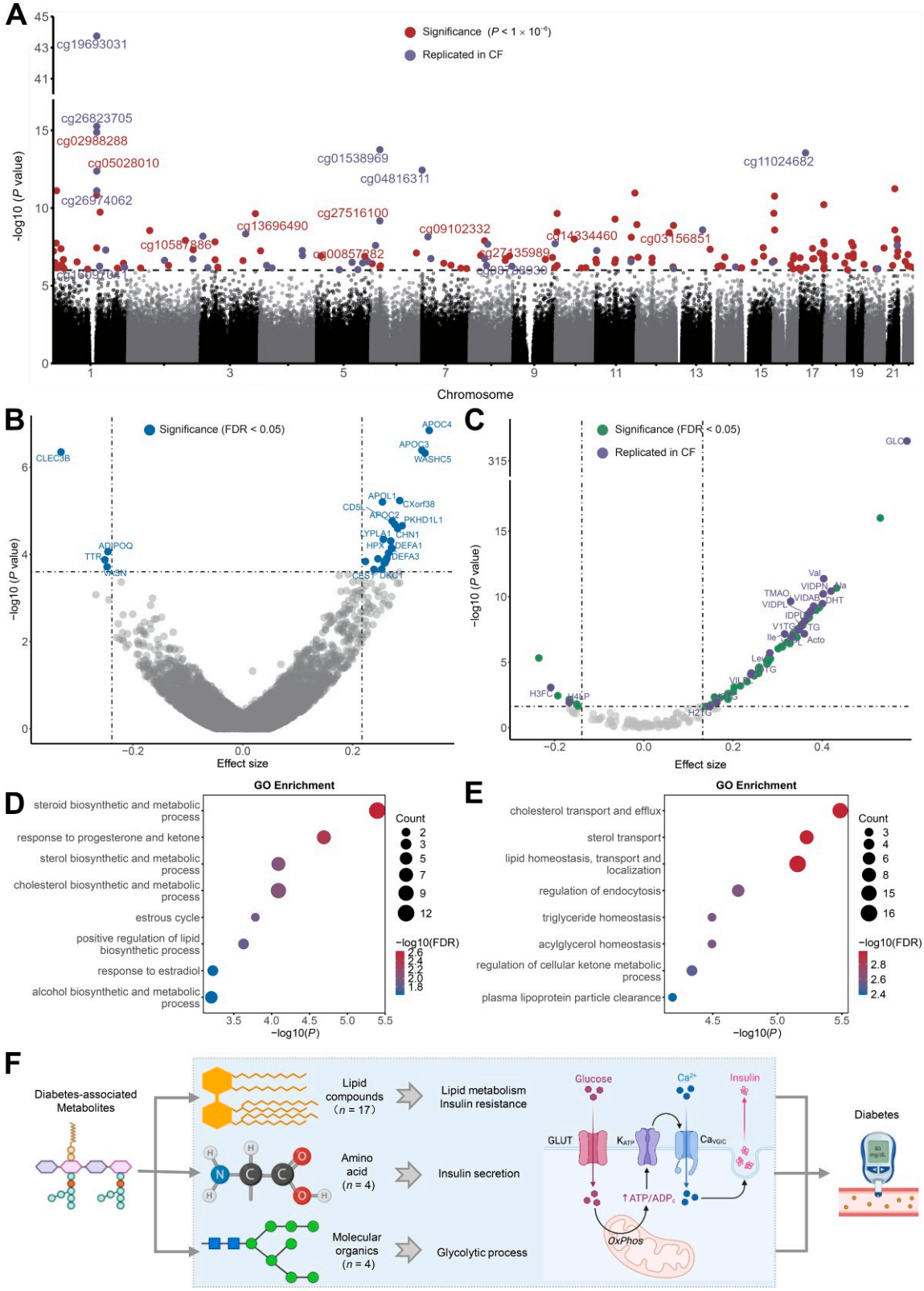


Figure 2—Diabetes biomarkers identified by EWAS, PWAS, and MWAS. *A*: Manhattan plot of the EWAS results in the NSPT cohort, showing the significance of CpG sites associated with diabetes. The horizontal line indicates the threshold of $P < 1 \times 10^{-6}$. *B* and *C*: Volcano plot illustrating the PWAS results (*B*) and MWAS results (*C*), with top significant proteins labeled in blue and green (FDR < 0.05). CpGs and metabolites replicated by the CF cohort are labeled purple. Functional enrichment of genes where the CpGs locate (*D*) and proteins (*E*) in the GO database, with functionally related terms aggregated together. *F*: Functional interpretation for the identified metabolites.

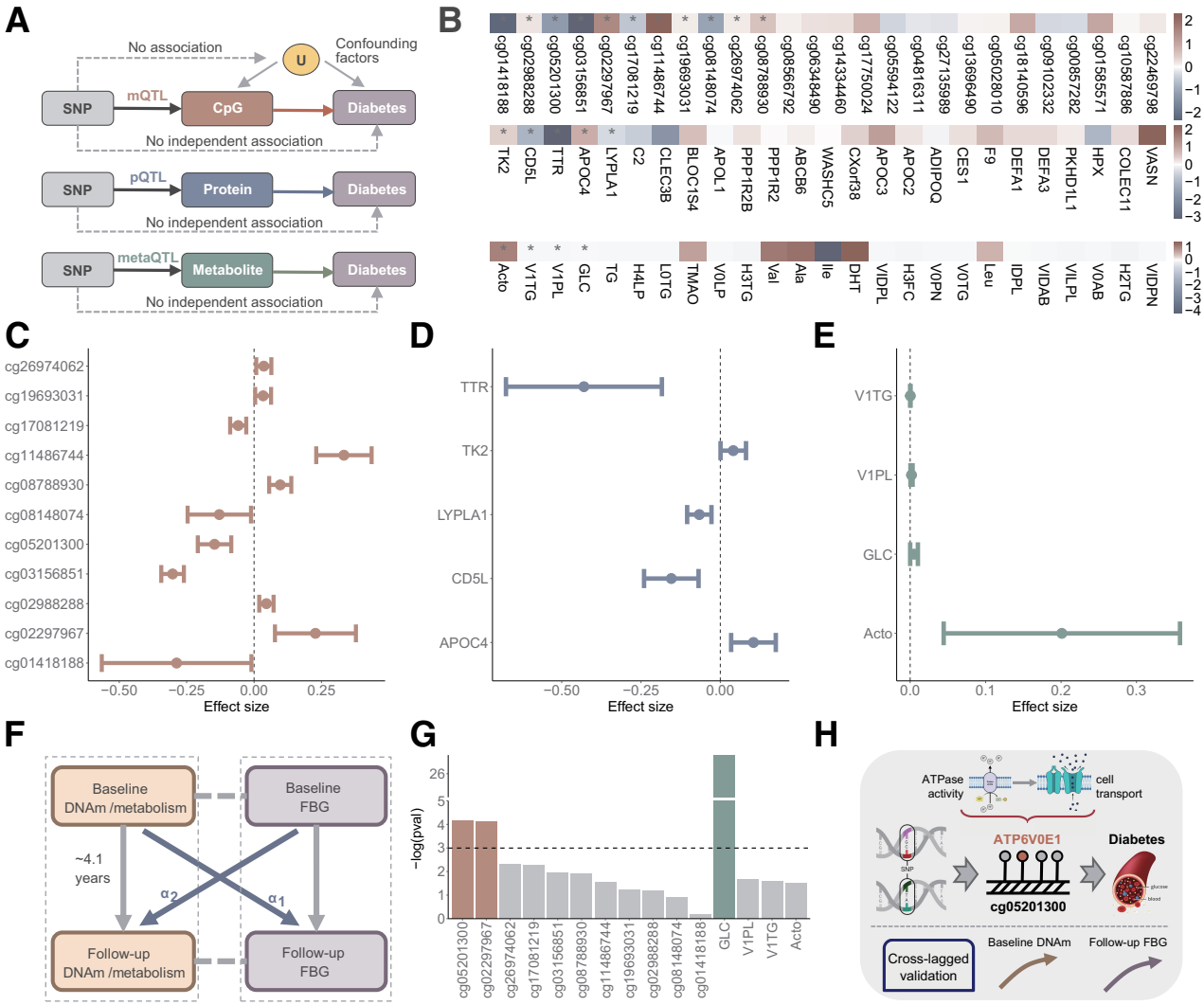


Figure 3—Causal links detected by the MR analysis. *A*: Schematic representation of MR analysis for CpGs (red), proteins (blue), and metabolites (green), with SNPs serving as instruments. *B*: Results of MR model for the identified biomarkers. Biomarkers causally linked to diabetes are marked with asterisks (*). The boxes are colored according to the effect sizes calculated by EWAS, PWAS, and MWAS. *C–E*: Forest plots of MR results for CpGs (*C*), proteins (*D*), and metabolites (*E*). *F*: Illustration of a cross-lagged path analysis model of FBG and DNA methylation/metabolites. *G*: Validation by cross-lagged path analysis. CpGs validated by the longitudinal model are marked in red; the validated metabolites are marked in green. *H*: Example of the causal link between CpG (cg05201300) and diabetes identified by MR and validated by cross-lagged path analysis. In cross-lagged validation, the follow-up FBG level was significantly associated with the baseline DNA methylation (DNAm) of cg05201300.

proteomic measurements (Fig. 7*A* and Supplementary Texts). The clinical baseline model included sex, age, waist circumference, BMI, and FBG, HDL cholesterol, and LDL cholesterol values. We compared the performance of the baseline model with that of other single-omics models, as well as the combined model, which integrated omics biomarkers as additional features. Results showed that the FBG levels of the positive predictions were significantly higher than those of negative predictions in both the single-omics models and the combined model (Supplementary Fig. 10). The combined model also demonstrated significantly better performance (AUROC = 0.788) when combining omics biomarkers with clinical indicators, surpassing the baseline clinical model by 27.5% (Fig. 7*B*).

Then, the early-prediction model was validated in the Zhongyuan cohort, using baseline clinical indicators, DNA methylation, and metabolism level as input to predict diabetes onset. The model effectively distinguished positive from negative predictions based on FBG levels and predicted diabetes onset with an AUROC of 0.751 (Fig. 7*C*). Our model also outperformed the state-of-the-art omics models by Kim et al. (32) (AUROC = 0.594) and Episcore (31) (AUROC = 0.666) (Fig. 7*D*). We observed that the majority of the predictive gain was driven by CpGs (Supplementary Fig. 11), possibly because DNA methylation changes occur earlier and more stably reflect long-term disease risk compared with metabolites. To further explain the effectiveness of the predictors in the combined model, 79 individuals without

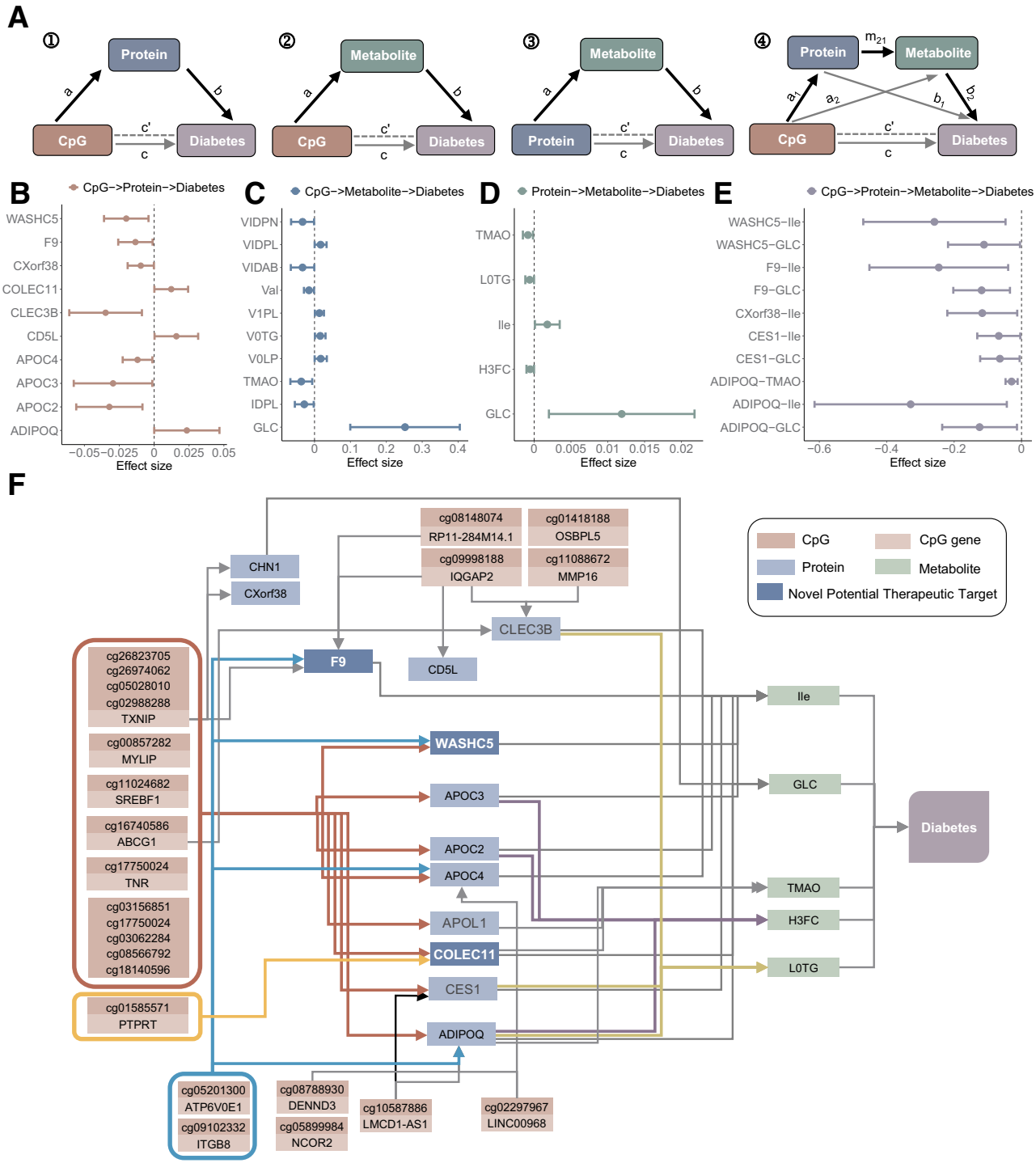


Figure 4—Signaling pathways identified by the mediation analysis. *A*: The schematic of mediation analysis among CpGs, proteins, metabolites, and diabetes. & Mediation for CpGs, proteins, and diabetes, with proteins as mediators. & Mediation for CpGs, metabolites, and diabetes, with metabolites as mediators. & Mediation for proteins, metabolites, and diabetes, with metabolites as mediators. & Serial mediation for CpGs, proteins, metabolites, and diabetes. *B–E*: Forest plot for mediation results of CpG→protein→diabetes (*B*), CpG→metabolite→diabetes (*C*), protein→metabolite→diabetes (*D*), and CpG→protein→metabolite→diabetes (*E*). *F*: Illustration of the results of serial mediation analysis.

diabetes at baseline but who were diagnosed with diabetes at follow-up in the CF cohort were selected. Significant changes in 12 CpGs and 4 metabolites were observed in their DNA methylation and metabolism levels before and after diabetes onset (Supplementary Fig. 12).

DISCUSSION

In this study, we conducted multilayered omics profiling in a large Chinese cohort (NSPT, n = 3,451) and identified 97 biomarkers (n = 43 CpGs, 29 proteins, and 25 metabolites) significantly associated with diabetes, with a subset

A

GO:0030301 (cholesterol transport)	GO:0045834 (positive regulation of lipid metabolic process)	GO:0009749 (response to glucose)	GO:0034381 (plasma lipoprotein particle clearance)
cg01418188(*OSBPL5*);cg16740586(*ABCG1*); *CES1*;*APOC2*; *APOC3*;*ADIPOQ*	cg22469798(*IGF1R*); cg11024682(*SREBF1*); cg16740586(*ABCG1*); *CES1*; *APOC2*; *ADIPOQ*	cg19693031, cg26823705, cg02988288(*TXNIP*); cg11024682(*SREBF1*); cg22469798(*IGF1R*); *ADIPOQ*	cg00857282(*MYLIP*); *APOC2*; *APOC3*;*ADIPOQ*

hsa04979 (Cholesterol metabolism)	hsa04152 (AMPK signaling pathway)	hsa04936 (Alcoholic liver disease)	hsa04144 (Endocytosis)
cg01418188 (*OSBPL5*); cg00857282(*MYLIP*); *APOC2*; *APOC3*	cg22469798(*IGF1R*); cg11024682(*SREBF1*); *ADIPOQ*	cg11024682(*SREBF1*); *ADIPOQ*	cg22469798(*IGF1R*); *WASHC5*

R-HSA-9024446 (NR1H2 and NR1H3-mediated signaling)	R-HSA-174824 (Plasma lipoprotein assembly, remodeling, and clearance)	R-HSA-8964058 (HDL remodeling)	R-HSA-381340 (Transcriptional regulation of white adipocyte differentiation)
cg00857282(*MYLIP*); cg16740586(*ABCG1*); cg11024682(*SREBF1*); cg05899984(NCOR2); *APOC2*;*APOC4*;	cg00857282(*MYLIP*); cg16740586(*ABCG1*); *APOC2*; *APOC3; APOC4*;	cg16740586(*ABCG1*); *APOC2*; *APOC3;*	cg11024682(*SREBF1*); cg05899984(NCOR2); *ADIPOQ*

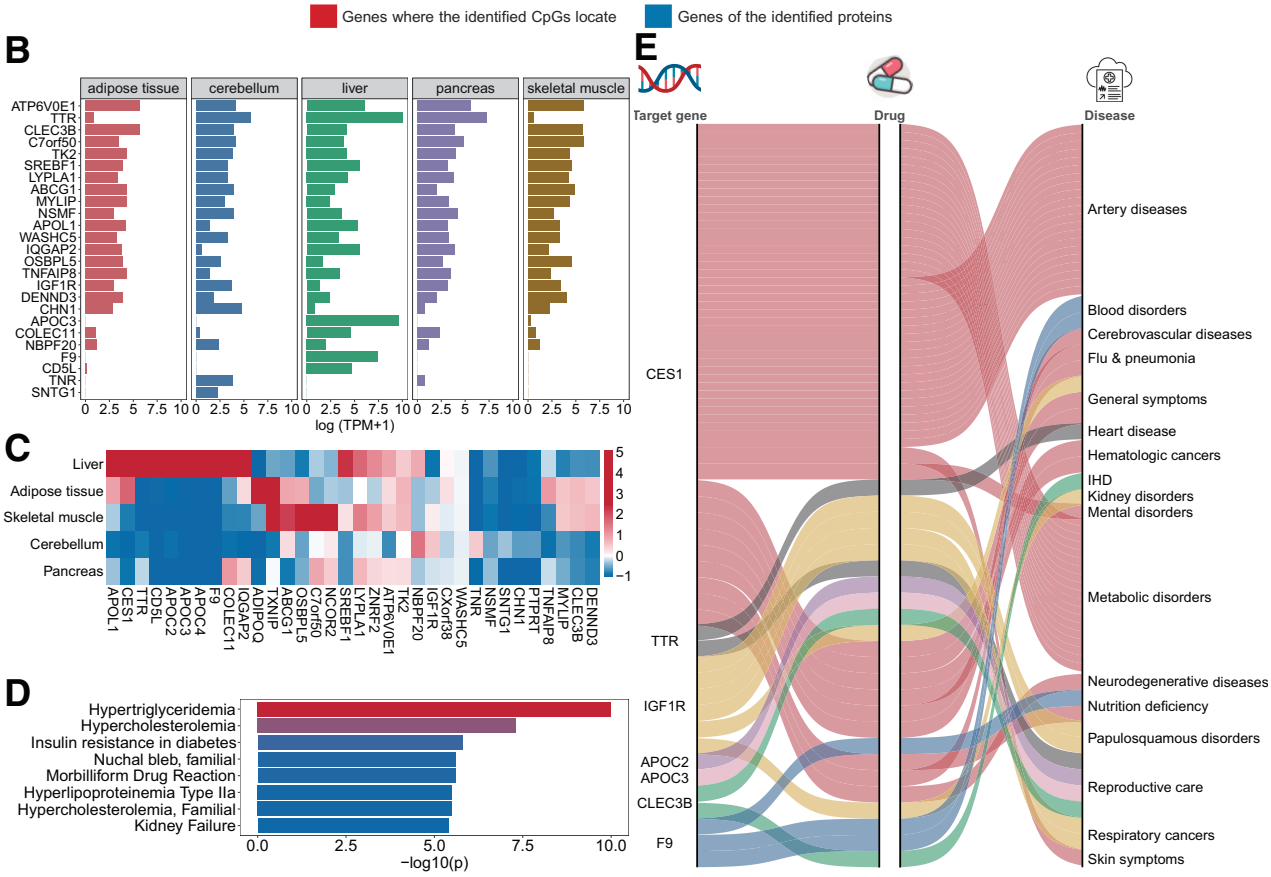


Figure 5—Prioritization for the identified biomarkers. *A*: Cofunction of diabetes-related CpGs and proteins. Biomarkers in red represent CpGs and genes where those CpGs locate; biomarkers in blue are the identified proteins. *B*: The expression levels of the identified genes of CpGs and proteins in diabetes-associated tissues, including adipose, cerebellum, liver, pancreas, and skeletal muscle tissues. TPM, transcripts per million. *C*: Heat map of the tissue-specificity scores of the identified biomarkers. *D*: Enrichment of the biomarkers in DisGeNet. *E*: Gene-drug-disease network of the biomarkers overlapped with the GREP database. Only biomarkers reported by MR and mediation analyses are demonstrated. IHD, ischemic heart disease.

of 43 CpGs and 25 metabolites replicated in an independent cohort (CF, n = 532). Notably, 20 biomarkers (n = 11 CpGs, 5 proteins, 4 metabolites) demonstrated consistent causal associations with diabetes in MR analyses, and >190 mediation pathways were identified across omics layers. These findings provide a multilayered molecular landscape of diabetes and lay the foundation for integrated analyses that can go beyond correlation.

Although previous studies have identified diabetes-associated biomarkers through various omics platforms, most

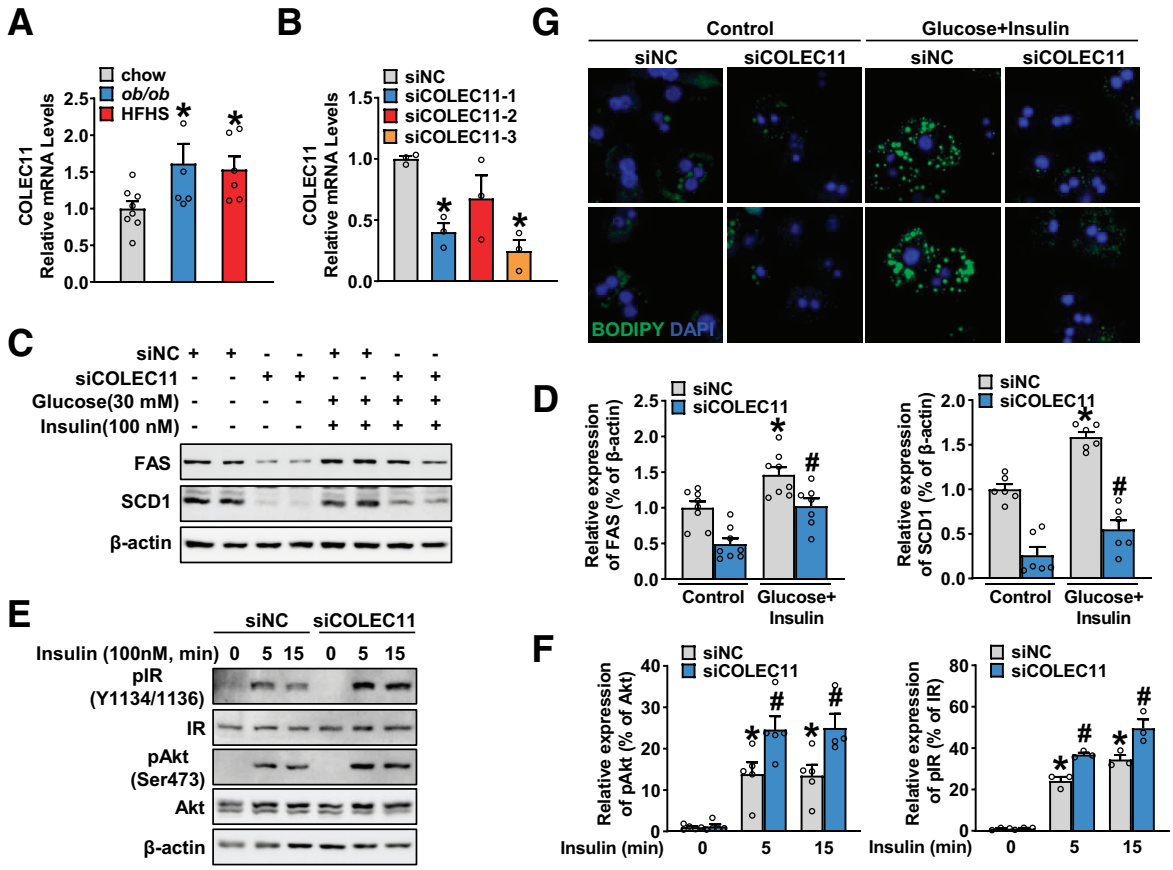


Figure 6—Validation of the function of COLEC11. *A*: Expression levels of *Colec11* are increased in livers of hyperglycemic *ob/ob* mice or high-fat high-sucrose diet–induced hyperglycemic mice. $n = 5–8$. *$P < 0.05$, versus chow. *B*: Efficiency of siRNA-mediated COLEC11 knockdown was verified in Huh7 cells. $n = 3$. *$P < 0.05$, versus sinegative control (siNC). *C*: COLEC11 deficiency prevents activation of lipogenesis induced by a high concentration of glucose plus insulin concentrations in hepatocytes. After a 24-h period of transfection with siNC or siCOLEC11, Huh7 cells were incubated in serum-free DMEM containing 5.5 mmol/L glucose for 24 h and then treated with a high-concentration of glucose plus insulin for 24 h. Lipogenesis-related genes were analyzed via immunoblots. *D*: The band intensity was quantified by densitometry. $n = 6–8$. *$P < 0.05$, versus siNC and control; #$P < 0.05$, versus siNC and glucose and insulin. *E*: COLEC11 deficiency increases insulin sensitivity in hepatocytes. Huh7 cells were transfected with siNC or siCOLEC11 for 48 h, followed by the treatment with insulin for the indicated time course, phosphorylation of IR, IRS, and Akt were analyzed through immunoblots. *F*: The band intensity was quantified by densitometry. $n = 3–5$. *$P < 0.05$, versus siNC and control; #$P < 0.05$, versus siNC and insulin. *G*: Knockdown of *Colec11* alleviates lipid accumulation in primary hepatocytes stimulated by a high concentration of glucose plus insulin. Mouse primary hepatocytes were isolated and transfected with siNC or siCOLEC11 for 24 h and then incubated in serum-free medium for 24 h, followed by a 24-h high concentration of glucose and insulin treatment. The neutral lipid droplets were analyzed by BODIPY staining.

of them focused on single-omics layers or limited integration (37–40). For instance, Yao et al. identified novel protein associations with T2D, supporting their causal relevance as drug targets (38). Wigger et al. demonstrated progressive β-cell remodeling in pancreatic islets from individuals with diabetes (39). Zaghlool et al. identified cluster-specific signatures of metabolites and proteins in patients with T2D (40). However, few used a systematic multilayered causal prioritization framework, particularly in East Asian populations.

An important strength of our study is the harmonized profiling of multiple circulating omics layers in an East Asian population, which remains underrepresented in diabetes research. To explore population relevance, we conducted cross-population comparisons with large-scale European cohorts and identified both shared and ancestry-specific molecular signatures of T2D (Supplementary Texts and Supplementary Fig. 13). These findings emphasize the importance of including diverse ancestries in omics studies to capture the full spectrum of disease biology and to support more equitable precision medicine strategies.

To dissect the molecular mechanisms underlying diabetes, we applied both MR and mediation analyses. MR identifies direct causal biomarkers, and mediation analysis reveals indirect regulators in molecular pathways, thereby offering a more nuanced and comprehensive view of molecular regulation. In the mediation analysis, we focused primarily on the biologically plausible directional cascade from DNA methylation, proteins, metabolites, to diabetes. This direction was guided by prior knowledge of molecular biology and the assumption that epigenetic modifications often precede transcriptomic and proteomic changes. For

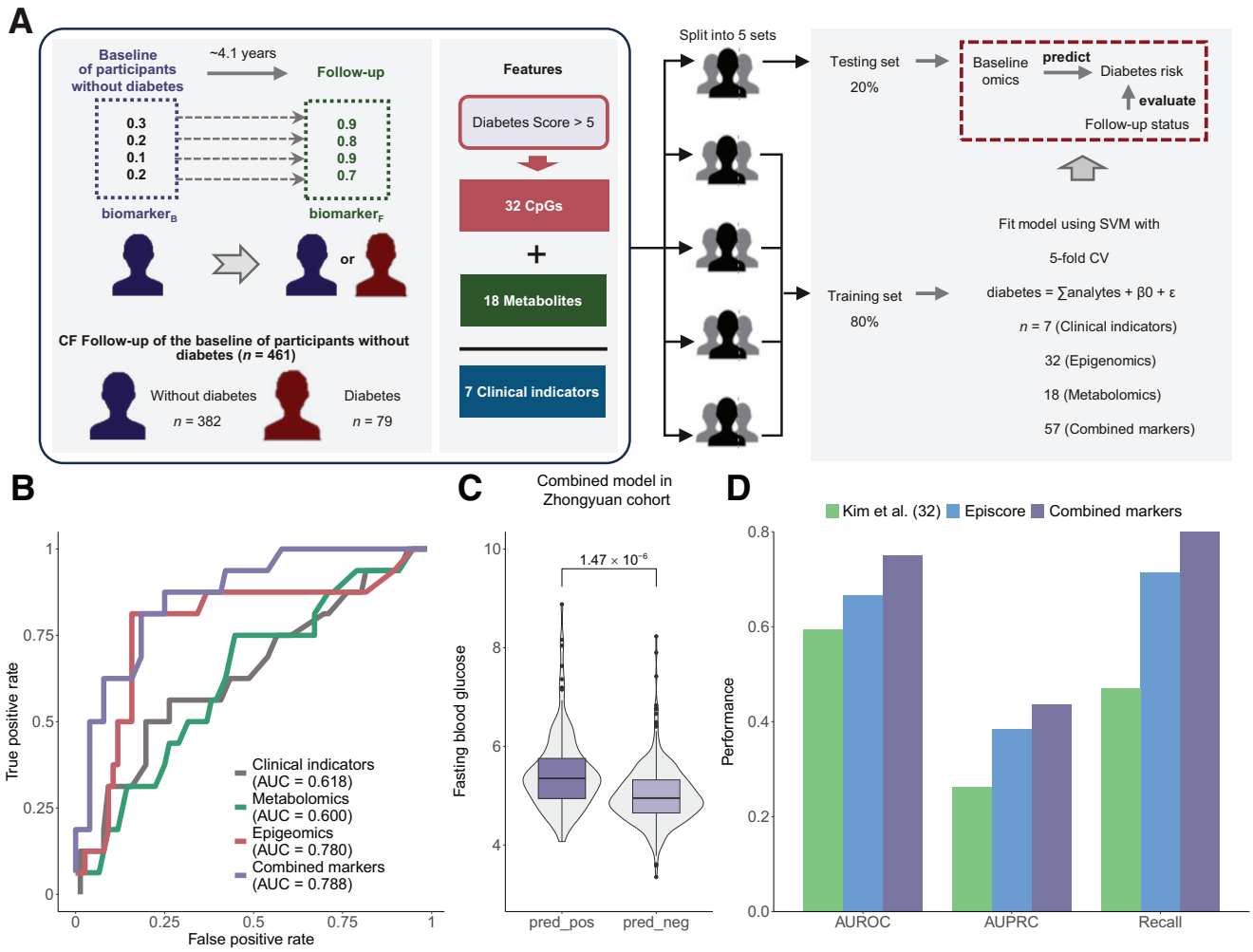


Figure 7—Model performance in predicting onset of diabetes. *A*: Schematic of the diabetes prediction model design. *B*: AUROC comparison of models with different input types. *P* values indicate the significance of difference between each model with the baseline clinical model, calculated using DeLong tests. *C*: Comparison of FBG levels between positive predictions (pred-pos) and negative predictions (pred_neg) in the Zhongyuan cohort. *D*: Overall performance comparison of our models and other state-of-the-art models in the external cohort. AUPRC, area under the precision-recall curve; CV, coefficient of variation; SVM, support vector machine.

clarity and interpretability, we only focused on the canonical cascade in the main text. However, we acknowledge that biological systems are complex and may exhibit feedback regulation or reverse causality. To explore this, we conducted additional mediation analyses covering other possible directional combinations among biomarkers and diabetes. These are detailed in the Supplementary Texts and Supplementary Fig. 14.

Existing biomarker prioritization methods often rely on limited types of data or a single analytical approach (41,42). These methods typically focus on one dimension of evidence, without considering the broader biological context. To address this limitation, we constructed a multilayered integrative framework that combines differential association, replication, functional annotation, MR, mediation analysis, longitudinal inference, and drug conversion assessment to comprehensively evaluate each biomarker. This framework reduces the likelihood of false-positive results and highlights both direct causal drivers and biologically meaningful intermediates. As a result, we provide a refined and biologically informed list of candidate biomarkers that offers both scientific insight into diabetes pathophysiology and practical utility for clinical translation.

We used COLEC11 as an illustrative example of how multiomics signaling pathways can guide therapeutic target discovery. Identified through our integrative analysis as liver-enriched, diabetes-associated protein with regulatory links to upstream and downstream biomarkers, COLEC11 emerged as a promising candidate. Population-level data showed that higher circulating COLEC11 levels were associated with increased diabetes risk. Consistently, COLEC11 knockdown in hepatocytes reduced lipid accumulation, suppressed lipogenic gene expression, and improved insulin sensitivity, supporting its functional role in hepatic metabolic dysregulation. Given its involvement in insulin resistance and lipid metabolism, COLEC11 may contribute to diabetes

pathogenesis via hepatic mechanisms. These findings demonstrate the biological relevance of our integrative framework for identifying putative therapeutic targets. Furthermore, by combining biomarkers with causal and mediation evidence, we developed an early prediction model that outperformed single-omics and existing tools, underscoring the clinical utility of multiomics integration.

Despite the promising findings of our study, several limitations should be acknowledged. First, our model reported biomarkers mainly identified in the NSPT cohort, including epigenome, proteome, and metabolites, which may not capture all relevant biological processes, such as transcriptome and microbiome. Incorporating more omics layers and leveraging advances in single-cell technologies could uncover additional biomarkers and pathways involved in diabetes pathogenesis. Second, MR analysis and mediation analysis in our study revealed potential pathways from biomarkers to diabetes. However, these analyses are limited by the complexity of biological interactions and the potential for unidentified confounding factors. Reverse analyses and other hypotheses of regulatory processes need to be explored in future studies. Third, although our early-prediction model demonstrates improved accuracy over existing models, the study population was predominantly Chinese. Future studies should aim to validate and refine the model in more diverse cohorts to ensure broader applicability.

In conclusion, our study presents an important advancement in the field of diabetes research by integrating multiomics data to identify and prioritize biomarkers, elucidate signaling pathways, and improve predictive performance for diabetes. Collectively, our findings offer a comprehensive and prioritized list of multiomics biomarkers and elucidate specific signaling pathways involved in diabetes, contributing significantly to the understanding of diabetes pathophysiology and the selection of therapeutic targets.

Funding. This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (grant XDB38020400 to S.W.), the National Natural Science Foundation of China (grants 92249302 and 32325013 to S.W. and 32200472 to W.L.), Chinese Academy of Sciences Young Team Program for Stable Support of Basic Research (grant YSBR-077 to S.W.), Shanghai Science and Technology Commission Excellent Academic Leaders Program (grant 22XD1424700 to S.W.), National Key R&D Program of China (grants 2022YFC3400700 and 2022YFA0806400 to H.T., 2023YFA1801100 and 2019YFA0802502 to Y.L., 2022YFA1004800 to Y.W., and 2022YFA1004804 to H.L.), Shanghai Municipal Science and Technology Major Project (grants 2017SHZDZX01 to S.W. and 2022MVDKL-K2 to Y.L.), and CAMS Innovation Fund for Medical Science (grant 2019-I2M-5-066 to J.W.).

Duality of Interest. No potential conflicts of interest relevant to this article were reported.

Author Contributions. S.W., Y.L., L.J., B.-H.J., H.T., and W.L. designed the work. W.L. and Y.C. performed the statistical analyses. A.C. and M.H. performed the biological experiments. W.L., Y.C., and A.C. wrote the manuscript. B.-H.J., H.T., X.G., C.D., J.W., M.X., J.Q., Q.H., Q.W., Y.Z., Q.P., and J.L. contributed to data preparation. S.W., Y.L., L.J., B.-H.J., and H.T. supervised the project and revised the manuscript. G.Z., Y.W., and H.L. revised the manuscript. All the authors approved the submitted version. S.W. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Tomic D, Shaw JE, Magliano DJ. The burden and risks of emerging complications of diabetes mellitus. Nat Rev Endocrinol 2022;18:525–539
2. Yong J, Johnson JD, Arvan P, et al. Therapeutic opportunities for pancreatic β-cell ER stress in diabetes mellitus. Nat Rev Endocrinol 2021; 17:455–467
3. Florath I, Butterbach K, Heiss J, et al. Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults. Diabetologia 2016;59:130–138
4. Chambers JC, Loh M, Lehne B, et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. Lancet Diabetes Endocrinol 2015;3:526–534
5. Meeks KAC, Henneman P, Venema A, et al. Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: findings from the RODAM study. Int J Epidemiol 2019;48:58–70
6. Kim H, Bae JH, Park KS, et al. DNA methylation changes associated with type 2 diabetes and diabetic kidney disease in an East Asian population. J Clin Endocrinol Metab 2021;106:e3837–e3851
7. Beijer K, Nowak C, Sundström J, et al. In search of causal pathways in diabetes: a study using proteomics and genotyping data from a cross-sectional study. Diabetologia 2019;62:1998–2006
8. Kim SW, Choi J-W, Yun JW, et al. Proteomics approach to identify serum biomarkers associated with the progression of diabetes in Korean patients with abdominal obesity. PLoS One 2019;14:e0222032
9. Yao P, Iona A, Pozarickij A, et al.; China Kadoorie Biobank Collaborative Group. Proteomic analyses in diverse populations improved risk prediction and identified new drug targets for type 2 diabetes. Diabetes Care 2024;47:1012–1019
10. Rooney MR, Chen J, Echouffo-Tcheugui JB, et al. Proteomic predictors of incident diabetes: results from the Atherosclerosis Risk in Communities (ARIC) Study. Diabetes Care 2023;46:733–741
11. Wang TJ, Larson MG, Vasan RS, et al. Metabolite profiles and the risk of developing diabetes. Nat Med 2011;17:448–453
12. Bragg F, Trichia E, Aguilar-Ramirez D, et al. Predictive value of circulating NMR metabolic biomarkers for type 2 diabetes risk in the UK Biobank study. BMC Med 2022;20:159
13. Peng Q, Liu X, Li W, et al. Analysis of blood methylation quantitative trait loci in East Asians reveals ancestry-specific impacts on complex traits. Nat Genet 2024;56:846–860
14. Li W, Xia M, Zeng H, et al. Longitudinal analysis of epigenome-wide DNA methylation reveals novel loci associated with BMI change in East Asians. Clin Epigenetics 2024;16:70
15. Wu Q, Huang Q-X, Zeng H-L, et al. Prediction of metabolic disorders using NMR-based metabolomics: the Shanghai Changfeng Study. Phenomics 2021;1:186–198
16. Ritchie ME, Phipson B, Wu D, et al. *limma* Powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47
17. Teschendorff AE, Breeze CE, Zheng SC, et al. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. BMC Bioinformatics 2017;18:105–114
18. Yang J, Lee SH, Goddard ME, et al. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88:76–82
19. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575

20. Minelli C, Del Greco M F, van der Plaat DA, et al. The use of two-sample methods for Mendelian randomization analyses on single large datasets. Int J Epidemiol 2021;50:1651–1659
21. Cavalcante RG, Sartor MA. Annotatr: genomic regions in context. Bioinformatics 2017;33:2381–2383
22. Harris MA, Clark J, Ireland A, et al.; the Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32:D258–D261
23. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28:27–30
24. Fabregat A, Jupe S, Matthews L, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2018;46:D649–D655
25. Kutmon M, Riutta A, Nunes N, et al. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res 2016;44:D488–D494
26. Yu G, Wang L-G, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–287
27. Tingley D, et al. Mediation: R package for causal mediation analysis. J Stat Software 2014;59:1–38
28. Rosseel Y, et al. Package 'lavaan'. J Stat Software 2017;48:1–36
29. Roden M, Shulman GI. The integrative biology of type 2 diabetes. Nature 2019;576:51–60
30. Chen Y, Li Z, Chen Y, et al. Cerebellar gray matter and white matter damage among older adults with prediabetes. Diabetes Res Clin Pract 2024; 213:111731
31. Cheng Y, Gadd DA, Gieger C, et al. Development and validation of DNA methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes. Nat Aging 2023;3:450–458
32. Kim H, Bae JH, Park KS, et al. DNA methylation changes associated with type 2 diabetes and diabetic kidney disease in an East Asian population. J Clin Endocrinol Metab 2021;106:e3837–e3851
33. Chambers JC, Loh M, Lehne B, et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. Lancet Diabetes Endocrinol 2015;3:526–534
34. Al Muftah WA, Al-Shafai M, Zaghlool SB, et al. Epigenetic associations of type 2 diabetes and BMI in an Arab population. Clin Epigenetics 2016;8:13
35. Meeks KAC, Henneman P, Venema A, et al. Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: findings from the RODAM study. Int J Epidemiol 2019;48:58–70
36. Fraszczyk E, Spijkerman AMW, Zhang Y, et al. Epigenome-wide association study of incident type 2 diabetes: a meta-analysis of five prospective European cohorts. Diabetologia 2022;65:763–776
37. Gudmundsdottir V, Zaghlool SB, Emilsson V, et al. Circulating protein signatures and causal candidates for type 2 diabetes. Diabetes 2020;69: 1843–1853
38. Yao P, Iona A, Pozarickij A, et al.; China Kadoorie Biobank Collaborative Group. Proteomic analyses in diverse populations improved risk prediction and identified new drug targets for type 2 diabetes. Diabetes Care 2024;47:1012–1019
39. Wigger L, Barovic M, Brunner A-D, et al. Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories towards type 2 diabetes. Nat Metab 2021;3:1017–1031
40. Zaghlool SB, Halama A, Stephan N, et al. Metabolic and proteomic signatures of type 2 diabetes subtypes in an Arab population. Nat Commun 2022;13:7121
41. Lundgaard AT, Burdet F, Siggaard T, et al. BALDR: a web-based platform for informed comparison and prioritization of biomarker candidates for type 2 diabetes mellitus. PLoS Comput Biol 2023;19:e1011403
42. Chen J-X, Geng T, Zhang Y-B, et al. Associations of clinical risk factors and novel biomarkers with age at onset of type 2 diabetes. J Clin Endocrinol Metab 2023;109:e321–e329