



Genetic Subtypes of Prediabetes, Healthy Lifestyle, and Risk of Type 2 Diabetes

Yang Li,¹ Guo-Chong Chen,² Jee-Young Moon,¹ Rhonda Arthur,¹ Daniela Sotres-Alvarez,³ Martha L. Daviglus,⁴ Amber Pirzada,⁴ Josiemer Mattei,⁵ Krista M. Perreira,⁶ Jerome I. Rotter,⁷ Kent D. Taylor,⁷ Yii-Der Ida Chen,⁷ Sylvia Wassertheil-Smoller,¹ Tao Wang,¹ Thomas E. Rohan,¹ Joel D. Kaufman,⁸ Robert Kaplan,¹ and Qibin Qi¹

Diabetes 2024;73:1178–1187 | <https://doi.org/10.2337/db23-0699>

Prediabetes is a heterogenous metabolic state with various risks for development of type 2 diabetes (T2D). In this study, we used genetic data on 7,227 US Hispanic/Latino participants without diabetes from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) and 400,149 non-Hispanic White participants without diabetes from the UK Biobank (UKBB) to calculate five partitioned polygenetic risk scores (pPRSs) representing various pathways related to T2D. Consensus clustering was performed in participants with prediabetes in HCHS/SOL ($n = 3,677$) and UKBB ($n = 16,284$) separately based on these pPRSs. Six clusters of individuals with prediabetes with distinctive patterns of pPRSs and corresponding metabolic traits were identified in the HCHS/SOL, five of which were confirmed in the UKBB. Although baseline glycemic traits were similar across clusters, individuals in cluster 5 and cluster 6 showed an elevated risk of T2D during follow-up compared with cluster 1 (risk ratios [RRs] 1.29 [95% CI 1.08, 1.53] and 1.34 [1.13, 1.60], respectively). Inverse associations between a healthy lifestyle score and risk of T2D were observed across different clusters, with a suggestively stronger association observed in cluster 5 compared with cluster 1. Among individuals with a healthy lifestyle, those in cluster 5 had a similar risk of T2D compared with those in cluster 1 (RR

ARTICLE HIGHLIGHTS

- Individuals with prediabetes differ in risk of progression to type 2 diabetes (T2D).
- This study aims to cluster individuals with prediabetes based on five partitioned polygenetic risk scores and to examine risk of T2D across clusters.
- Six subtypes of prediabetes were identified with distinctive patterns of genetic scores and corresponding metabolic traits, and two subtypes showed relatively high risk of T2D.
- This study provides insights into the use of genetic information to further stratify disease risk for early prevention and intervention among people with high T2D risk.

1.03 [0.91, 1.18]]. This study identified genetic subtypes of prediabetes that differed in risk of progression to T2D and in benefits from a healthy lifestyle.

Prediabetes is a condition of impaired glucose metabolism often preceding the ascertainment of overt diabetes (1). The term prediabetes can refer to a clinical state defined alternatively upon an observed fasting glucose between

¹Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY

²Department of Nutrition and Food Hygiene, School of Public Health, Soochow University, Suzhou, China

³Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC

⁴Institute for Minority Health Research, University of Illinois at Chicago, Chicago, IL

⁵Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA

⁶Department of Social Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC

⁷The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA

⁸Environmental and Occupational Health Sciences, Medicine, and Epidemiology, University of Washington, Seattle, WA

Corresponding author: Qibin Qi, qibin.qi@einsteinmed.edu

Received 30 August 2023 and accepted 1 April 2024

This article contains supplementary material online at <https://doi.org/10.2337/figshare.25571796>.

© 2024 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered. More information is available at <https://www.diabetesjournals.org/journals/pages/license>.

5.6 and 7.0 mmol/L (impaired fasting glucose [IFG]), a glucose level measured 2 h after oral glucose tolerance test (OGTT) between 7.8 and 11.1 mmol/L (impaired glucose tolerance [IGT]), or an HbA_{1c} level between 5.6 and 6.5% (2). Although all these subtypes carry the same designation of prediabetes, prediabetes is a heterogeneous metabolic state that varies in many aspects of its pathogenesis (3). In turn, subtypes of prediabetes may differ in risks of future development of type 2 diabetes (T2D) and its complications (3). In addition to the standard approach to prediabetes classification based on glucose and HbA_{1c} categories, more metabolic features have been proposed to improve the classification of prediabetes. Wagner et al. (4) used phenotypic variables related to glucose tolerance, insulin sensitivity, insulin secretion, lipids, adiposity, and a T2D polygenic risk score to identify six clusters of subphenotypes in individuals who were at increased risk for T2D. This demonstrated the pathophysiological heterogeneity among individuals before diagnosis of T2D, suggesting that subtyping of prediabetes based on multiple features could help to improve stratification of disease risk (3).

Previous efforts on subtyping individuals with T2D or at high risk of T2D have mainly focused on phenotypic variables (3–6), while many T2D-associated genetic variants have been identified through genome-wide association studies (GWAS), with various pathophysiological pathways (7). Compared with phenotypic variables, genetic variants are more likely to point to disease causation and do not change with disease progression or treatment (7). Thus, subtyping individuals with a high risk of T2D (e.g., prediabetes) based on genetic variants will not only help to better understand pathophysiological heterogeneity but also provide useful information in early prediction and prevention before onset of clinically detectable metabolic changes and diagnosis of T2D. Based on associations between 94 T2D-associated variants and 47 T2D-related traits, Udler et al. (7) created five pathway-based partitioned polygenic risk scores (pPRSs) representing two mechanisms of β -cell dysfunction (i.e., β -cell dysfunction score and low proinsulin score) and three mechanisms of insulin resistance (i.e., obesity score, lipodystrophy-like score, and disrupted liver lipid metabolism score) in relation to T2D. However, these pPRSs have not been examined among individuals with prediabetes, and it is unknown whether these various pPRSs can be used to cluster individuals with prediabetes.

In this study, we aimed to cluster individuals with prediabetes into subtypes based on previously designed pPRSs (7) reflecting different pathophysiological pathways of T2D in a prospective cohort of diverse Hispanic/Latino individuals with a high burden of diabetes. We then compared risk of incident T2D among individuals in different clusters of prediabetes during an average 6-year follow-up. In addition, we examined associations of healthy lifestyle and risk of T2D across different clusters of prediabetes. We replicated the analysis in a large, independent, prospective cohort of

European non-Hispanic White adults from the UK Biobank (UKBB).

RESEARCH DESIGN AND METHODS

Study Population

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a population-based cohort study that recruited 16,415 adults who self-identified as Hispanic or Latino and were aged 18–74 years at baseline (8,9). A comprehensive battery of interviews and a clinical assessment that included fasting and post-OGTT blood draw with central analysis of laboratory analytes were conducted during 2008–2011 (baseline) and 2014–2017 (visit 2). The current analysis included 7,227 participants who were free of diabetes at baseline, had genetic data, and completed the 6-year follow-up (visit 2) examination. The study was approved by the institutional review boards at all participating institutions. All individuals gave written informed consent. UKBB is a large, prospective, observational study that recruited ~500,000 men aged 37–73 years between 2006 and 2010 (10). In the current analysis, 400,149 non-Hispanic White participants without diabetes at baseline and with genetic data were included, with a median 10.1 years of follow-up. UKBB received research ethics committee approval (reference no. UKBB 11/NW/0382), and participants provided written informed consent.

Ascertainment of Prediabetes and T2D

In HCHS/SOL, prediabetes was defined as fasting glucose between 100 and 125 mg/dL, 2-h glucose after OGTT between 140 and 199 mg/dL, or HbA_{1c} between 5.7 and 6.5% and no self-reported physician-diagnosed diabetes or medically treated diabetes (2). Diabetes was defined as fasting glucose levels ≥ 126 mg/dL, 2-h glucose after OGTT ≥ 200 mg/dL, HbA_{1c} $\geq 6.5\%$, current use of glucose-lowering medication, or self-reported physician-diagnosed diabetes (2). Participants free of diabetes at baseline (visit 1) who were identified as having diabetes at visit 2 were deemed to have incident T2D. In UKBB, diabetes was defined through multiple procedures and sources of the diagnosis, including primary care, hospital admissions, and self-report, or by participants having an HbA_{1c} $\geq 6.5\%$ (2,10,11). Participants with prediabetes were defined as those with HbA_{1c} between 5.7 and 6.5% (2). Participants free of diabetes at baseline who were identified as having diabetes during follow-up were deemed to have incident T2D.

Measurements of Metabolic Traits

In HCHS/SOL, BMI; waist-to-hip ratio; plasma fasting glucose; 2-h glucose after OGTT; fasting insulin; HbA_{1c}; serum lipids, including triglycerides and total, LDL, and HDL cholesterol; and HOMA of insulin resistance (HOMA-IR) and HOMA of β -cell function (HOMA-B) were collected. In UKBB, metabolic traits included BMI, waist-to-hip ratio, random glucose, HbA_{1c}, triglycerides, and total, LDL, and

HDL cholesterol. HOMA-IR and HOMA-B in HCHS/SOL were estimated using the following equations:

$$\text{HOMA-IR} = \text{fasting glucose (mg/dL)} \\ * \text{fasting insulin (mU/L)} / 405$$

$$\text{HOMA-B} = 360 * \text{fasting insulin (mU/L)} / \\ [\text{fasting glucose (mg/dL)} - 63]$$

Genotyping, pPRS Calculation, and Clustering on Individuals

In HCHS/SOL, genome-wide genotyping was performed in 12,633 participants using a customized Illumina array (15041502 B3; Illumina Omni 2.5M array) plus ~150,000 custom single-nucleotide polymorphisms, and imputation was performed based on the 1000 Genomes Project phase 3 reference panel (12). In UKBB, genome-wide genotyping was performed in ~500,000 participants using Affymetrix UK BiLEVE Axiom Array or Affymetrix UK Biobank Axiom Array (13).

Five pPRSs, including β -cell score, proinsulin score, obesity score, lipodystrophy-like score, and liver-lipid score, were calculated based on 94 T2D-associated genetic variants using a previously reported method (7):

$$\text{pPRS}_i = \sum_{j=1}^{N_j} x_{ij} \beta_{ij}$$

where x_{ij} is the genotype for the j th single nucleotide polymorphism in i th pPRS (encodes as 0, 1, or 2 for the effect of allele dosage), and β_{ij} is the estimated effect size for the corresponding genotype x_{ij} (obtained from GWAS summary statistics) (Supplementary Table 1).

To cluster participants with prediabetes, k -means-based consensus clustering was performed for 3,677 participants with prediabetes in HCHS/SOL and for 16,284 participants with prediabetes in UKBB, using the ConsensusClusterPlus R package (14). The consensus clustering included three steps:

1. k -Means clustering based on pPRSs: The process started by randomly assigning each sample to one of the k clusters, where k was predefined. The algorithm was then iterated to optimize the cluster centroids until convergence, with the goal of minimizing the within-cluster variance.
2. Consensus matrix: The k -means clustering at the first step was repeated 1,000 times by resampling the data set. Each run produced a k -mean clustering result, and the cluster assignments for each data point were recorded. A consensus matrix was constructed to measure the similarity between any pair of data points across all clustering solutions. Each cell (i, j) in the consensus matrix represents how often data point i and data point j were clustered together.
3. Hierarchical clustering: The consensus matrix was used as input for hierarchical clustering. This process generates the final clustering solution. The optimal

number of clusters was determined by the largest average silhouette coefficient.

The complete process of calculating the pPRS for each participant, clustering participants with prediabetes into subgroups, and comparing metabolic traits and risk of incident T2D across clusters are depicted in Supplementary Fig. 1.

Lifestyle Score Calculation

Adherence to a healthy lifestyle was measured by a lifestyle score based on five well-established modifiable factors, including BMI, smoking, alcohol drinking, physical activity, and diet components, relevant to risk of T2D (15). Each factor was scored individually, and then the overall lifestyle score was calculated by summing the five scores (Supplementary Tables 2 and 3). For both studies, participants within the 2nd and 3rd tertiles of the lifestyle score were defined as being adherent to a healthy lifestyle.

Statistical Analysis

In both studies, linear regression was used to examine associations of the five pPRSs with baseline metabolic traits and to test differences in pPRSs and baseline metabolic traits between any two clusters of prediabetes after adjustment for covariates. Robust Poisson regression and Cox proportional hazards regression adjusting for covariates were performed in HCHS/SOL and UKBB, respectively, to examine the following:

1. The associations of the five pPRSs with risk of T2D (per SD increment) among participants without diabetes at baseline
2. The relative risks of T2D across clusters of prediabetes (cluster 1 defined as reference)
3. The associations of the other four pPRSs and a summed score (sum of these four pPRSs) with risk of T2D (per SD increment) stratified by the median of the proinsulin score among participants without diabetes at baseline (interactions were tested by including the respective interaction terms in the models: T2D outcome = $\beta_1 * \text{pPRS}_1 + \beta_2 * \text{pPRS}_2 + \beta_3 * \text{pPRS}_1 * \text{pPRS}_2 + \beta_4 * \text{covariate}$, where β_3 is the effect size of interaction term)
4. The associations between adherence to a healthy lifestyle and risk of T2D in all participants with prediabetes, as well as across clusters of prediabetes
5. The relative risks of T2D comparing cluster 5 with cluster 1 while stratifying according to the lifestyle score (cluster 1 with unhealthy lifestyle defined as reference)

ANOVA and χ^2 tests were applied to test differences in participant characteristics, including continuous variables (e.g., baseline glycemic traits, changes in glycemic traits) and categorical variables (e.g., glycemic status [IFG, IGT, or both], Hispanic/Latino background) across clusters of prediabetes, respectively. Results from HCHS/SOL and

UKBB were pooled by inverse variance-weighted, fixed-effects meta-analyses. In the combined analysis, the difference in the effect sizes of healthy lifestyle on risk of T2D between cluster 5 and cluster 1 was tested using Cochran Q test for heterogeneity. We considered a two-sided $P < 0.05$ as statistically significant and used false discovery rate-adjusted P values for multiple tests. All analyses were performed using R 4.1.2 statistical software. An expanded description of the methods is available in the Supplementary Material.

Data and Resource Availability

The data sets analyzed during the current study can be requested from HCHS/SOL (<https://sites.csc.unc.edu/hchs>) and UKBB (<https://www.ukbiobank.ac.uk>). The data sets generated during the current study are available from the corresponding author upon reasonable request.

RESULTS

Baseline Characteristics of Participants

Baseline characteristics of participants from HCHS/SOL and UKBB are shown in Supplementary Tables 4 and 5. Differences in characteristics between participants with normal glucose and those with prediabetes were generally similar for the two studies. Compared with those with normal glucose, participants with prediabetes were older, were more likely to be males and current smokers, and had lower education levels. Participants with prediabetes had higher levels of obesity measures and glycemic traits and less favorable lipid profiles (i.e., higher triglycerides, lower HDL cholesterol) compared with those with normal glucose.

Genetic Risk Scores, Metabolic Traits, and Risk of T2D

We first calculated five pPRSs among 7,227 participants without diabetes at baseline in HCHS/SOL and 400,149 participants without diabetes at baseline in UKBB (Supplementary Fig. 2). Relatively weak correlations were observed among these five pPRSs both in HCHS/SOL ($r = 0.07$ – 0.26) and in UKBB ($r = 0.10$ – 0.28) (Supplementary Fig. 3).

We then examined correlations of these five scores with multiple diabetes-related metabolic traits. Correlations among these traits are shown in Supplementary Fig. 4. In HCHS/SOL, β -cell score and low proinsulin score were inversely associated with β -cell function measured by HOMA-B, while obesity score, lipodystrophy-like score, and liver-lipid score were positively associated with insulin resistance measures (Supplementary Fig. 5). All five scores were positively correlated with fasting glucose, 2-h glucose after OGTT, and HbA_{1c} in HCHS/SOL. Previously reported associations of these pPRSs with other corresponding phenotypic traits (11) were also confirmed in HCHS/SOL. Similar correlations of these pPRSs with glycemic traits (e.g., HbA_{1c}) and other metabolic traits were

observed in UKBB (Supplementary Fig. 5), although fasting glucose, 2-h glucose after OGTT, HOMA-B, or HOMA-IR were not measured.

We also examined associations of these pPRSs with incident T2D. Among 7,227 participants without diabetes at baseline in HCHS/SOL, 888 incident T2D cases were identified after an average 6-year follow-up. Among 400,149 participants without diabetes at baseline in UKBB, 10,806 incident T2D cases were identified during a median 10.1 years of follow-up. All five scores were positively associated with risk of T2D in both studies (all $P < 0.05$) (Supplementary Fig. 5).

Genetic Subtypes of Prediabetes and Baseline Metabolic Profiles

Six clusters of prediabetes were identified among 3,677 participants with prediabetes at baseline in HCHS/SOL according to the highest average silhouette coefficients (Supplementary Fig. 6). The numbers of participants within each cluster from cluster 1 to cluster 6 were 488 (13.2%), 569 (15.5%), 547 (14.8%), 742 (20.0%), 719 (19.5%), and 612 (16.6%), respectively, and each cluster displayed distinctive pPRS patterns (Fig. 1A). Metabolic trait patterns were generally consistent with the pPRS patterns in each cluster (Fig. 1B). Differences in the pPRSs and metabolic traits between any two clusters of prediabetes are shown in Supplementary Fig. 6. Main features of the pPRSs and metabolic traits for these six clusters are summarized in Supplementary Table 6. We then used the same consensus clustering approach and identified five genetic subtypes among 16,284 participants with prediabetes at baseline in UKBB (Supplementary Fig. 6). These clusters were highly consistent with those identified in HCHS/SOL, except for cluster 3, which was not detected in UKBB (Fig. 1C). Additionally, we used t -distributed stochastic neighbor embedding (16) to reduce five pPRSs into two components and thus visualize the six clusters in two dimensions (Supplementary Fig. 7).

Although these clusters had a distinctive metabolic trait pattern corresponding to their various pPRS patterns, none of the clusters showed an overall better or worse metabolic pattern at baseline than others (Supplementary Tables 7 and 8). In particular, proportions of participants with IFG, IGT, or both were similar ($P = 0.13$), and levels of fasting glucose ($P = 0.62$), 2-h glucose after OGTT ($P = 0.35$), and HbA_{1c} ($P = 0.98$) were similar across the six clusters in HCHS/SOL (Fig. 2A–D). In UKBB, levels of random glucose ($P = 0.43$) and HbA_{1c} ($P = 0.71$) were also similar across five clusters (Fig. 2E and F). There were no differences in demographic, socioeconomic, and behavioral factors across clusters in either study (Supplementary Tables 7 and 8), except for Hispanic/Latino background in HCHS/SOL. However, all six clusters were identified in each of the six Hispanic/Latino groups, and no group was dominated by one or two clusters (Supplementary Fig. 9).

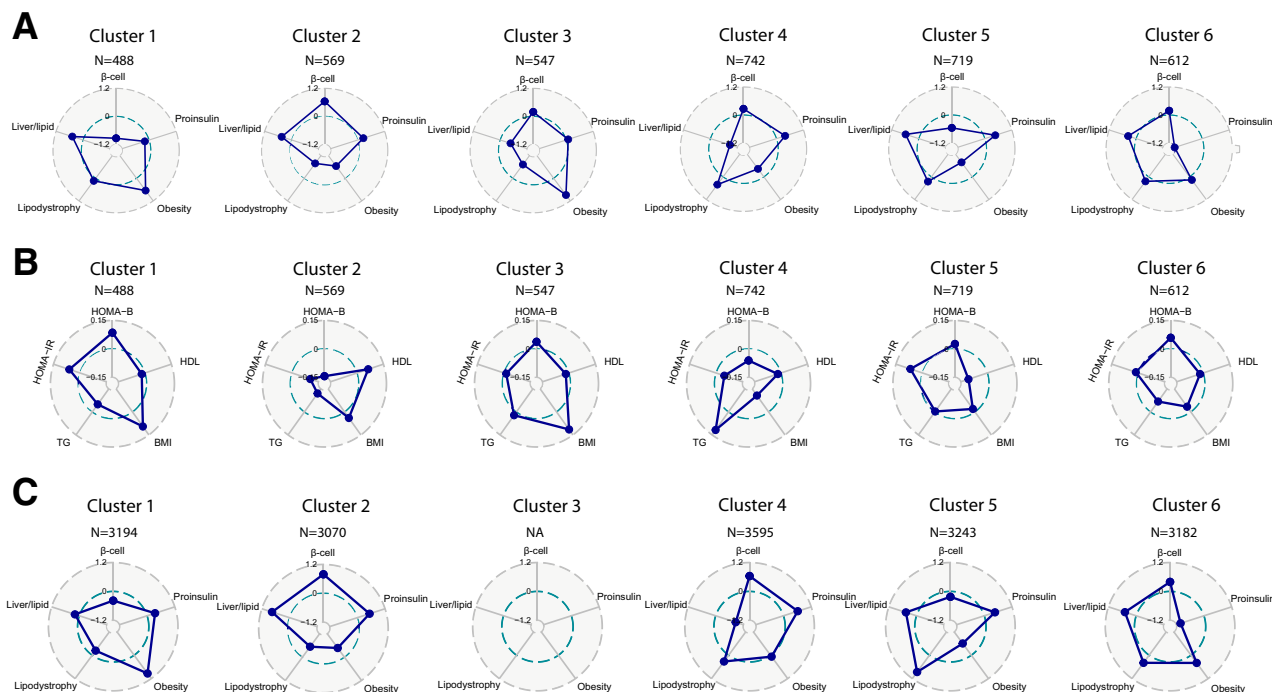


Figure 1—Patterns of genetic scores and metabolic traits pattern across clusters of prediabetes. **A:** Patterns of five pPRSs across six clusters of individuals with prediabetes in HCHS/SOL. **B:** Patterns of five T2D-related metabolic traits across six clusters of individuals with prediabetes in HCHS/SOL. **C:** Patterns of five pPRSs across five clusters of individuals with prediabetes in UKBB (cluster 3 was not detected in UKBB). Blue dots on five axes in radar plot show the median values of standardized pPRSs or standardized metabolic traits. TG, triglyceride.

Genetic Subtypes of Prediabetes and Incident T2D

We then compared risk of incident T2D among these clusters of participants with prediabetes at baseline. In HCHS/SOL, participants in cluster 6, characterized by a very low proinsulin score, had the highest risk of T2D (risk ratio [RR] 1.39 [95% CI 1.10, 1.76] compared with cluster 1; $P = 0.006$), followed by those in cluster 5 characterized by high liver-lipid score, lipodystrophy-like score, and proinsulin score and low β -cell score and obesity score (RR 1.23 [0.96, 1.57] compared with cluster 1; $P = 0.096$) (Fig. 3A). In UKBB, participants in cluster 5 and cluster 6 also showed the highest risk of T2D compared with cluster 1 (hazard ratios 1.35 [95% CI 1.05, 1.75] and 1.29 [1.00, 1.67]; $P = 0.021$ and 0.049, respectively). In the combined analysis of two cohorts, participants in cluster 5 and cluster 6 had a 29% [95% CI 8, 53%] and 34% [95% CI 13, 60%] increased risk of T2D compared with those in cluster 1 ($P = 0.005$ and <0.001 , respectively).

We also examined changes in glycemic traits over 6 years across these six clusters in HCHS/SOL and found that changes in fasting glucose (overall $P = 0.02$) (Fig. 3B), 2-h glucose after OGTT (overall $P = 0.003$) (Fig. 3C), and HOMA-IR (overall $P = 0.03$) (Fig. 3D) differed across clusters, despite baseline levels of these traits being similar across clusters. Compared with those in cluster 1, participants in cluster 5 and cluster 6 had a greater increase in fasting glucose and 2-h glucose after OGTT over 6 years (all $P < 0.05$) (Fig. 3B and C).

It was unexpected that participants with prediabetes in cluster 6, which had very low proinsulin scores and only slightly above-average levels of the other four scores, showed the highest risk of T2D (Figs. 1A and 3A), since lower proinsulin score was associated with lower risk of T2D and favorable glycemic traits (Supplementary Fig. 5). Revisiting the associations of the proinsulin score and other genetic scores with risk of T2D, we found evidence of an interaction between the proinsulin score and other genetic scores on risk of T2D (Supplementary Fig. 10). In HCHS/SOL, the associations of β -cell score and liver-lipid score with risk of T2D were stronger among participants with a low proinsulin score than among those with a high proinsulin score (P for interaction = 0.04 and 0.06 respectively). The combined genetic effect of the other four scores (summed score) was stronger among participants with a low proinsulin score (RR 1.31 [95% CI 1.20, 1.42]; $P < 0.001$) compared with those with a high proinsulin score (RR 1.09 [0.99, 1.21]; $P = 0.09$) (P for interaction = 0.01). These potential interactions were replicated in UKBB (all P for interaction < 0.05), with an additional interaction between lipodystrophy-like score and proinsulin score identified (P for interaction = 0.01).

Healthy Lifestyle, Genetic Subtypes of Prediabetes, and Risk of T2D

We then examined the association between adherence to a healthy lifestyle in HCHS/SOL (Supplementary Table 2)

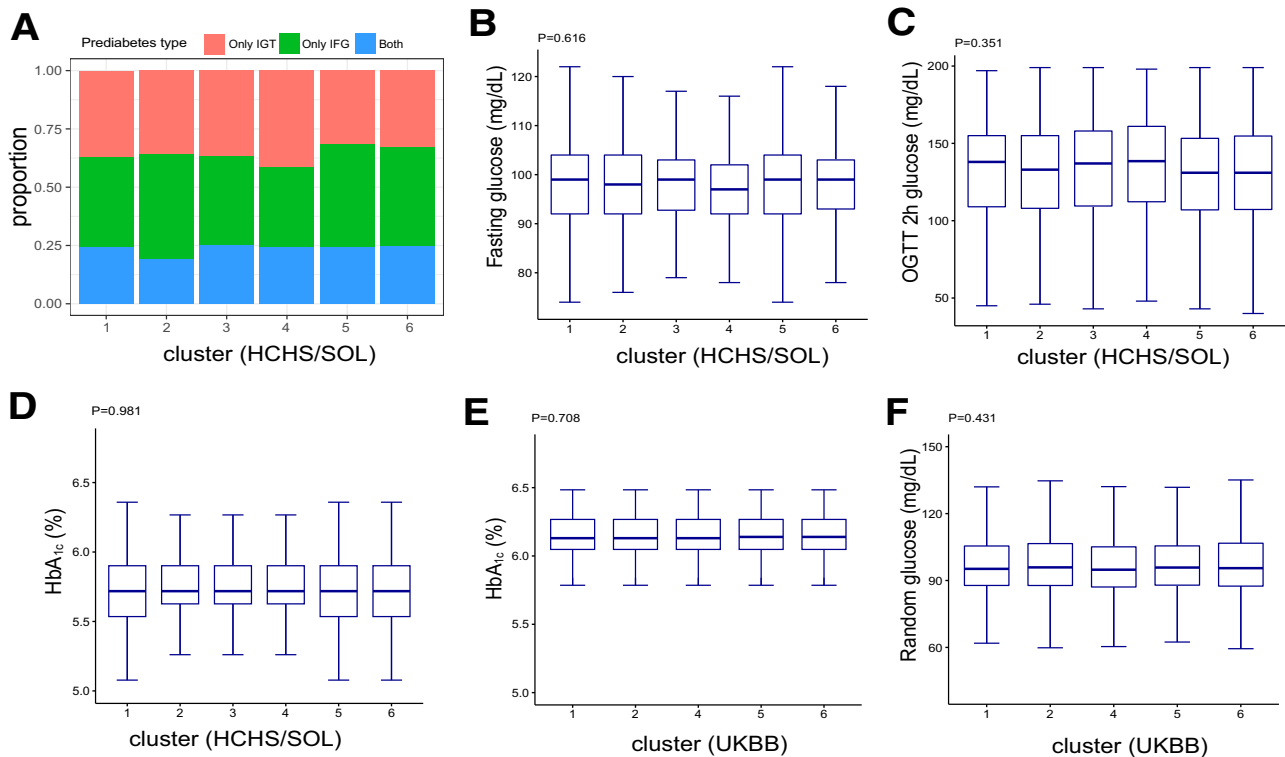


Figure 2—Comparison of baseline metabolic traits and glycemic status across clusters of prediabetes. **A**: Proportion of individuals with only IGT, only IFG, or both across six clusters in HCHS/SOL. **B–D**: Box plots of fasting glucose, 2-h glucose after OGTT, and HbA_{1c} at baseline across six clusters in HCHS/SOL. **E** and **F**: Box plots of HbA_{1c} and random glucose at baseline across five clusters in UKBB.

and in UKBB (Supplementary Table 3) and risk of T2D across different clusters of participants with prediabetes (Fig. 4A). In HCHS/SOL, healthy lifestyle (i.e., 2nd and 3rd tertiles of the lifestyle score) was associated with a lower risk of T2D among all participants with prediabetes (RR 0.73 [95% CI 0.63, 0.84]; $P < 0.001$), those in cluster 2 (RR 0.67 [0.47, 0.95]; $P = 0.02$), and those in cluster 5 (RR 0.65 [0.47, 0.89]; $P = 0.008$), but not among those in other clusters. In UKBB, healthy lifestyle was significantly associated with a lower risk of T2D in all participants with prediabetes (RR 0.80 [0.75, 0.85]; $P < 0.001$), as well as in each cluster separately, with the highest effect size observed in cluster 5 (RR 0.71 [0.62, 0.81]; $P < 0.001$) and the lowest effect size observed in cluster 1 (RR 0.86 [0.75, 1.00]; $P = 0.048$). In the combined analysis of two cohorts, the inverse association between healthy lifestyle and risk of T2D was observed across all five clusters, with a suggestive stronger association observed in cluster 5 (RR 0.70 [0.62, 0.79]; $P < 0.001$) compared with cluster 1 (RR 0.85 [0.74, 0.97]; $P = 0.02$). We did not find an overall difference in the association between adherence to a healthy lifestyle and risk of T2D across all clusters using Cochran Q test for heterogeneity. However, suggestive different effect sizes were detected between cluster 1 and cluster 5 ($P = 0.04$ for difference in effect sizes).

To further illustrate potentially extra beneficial effects of healthy lifestyle on risk of T2D in cluster 5 compared

with cluster 1, we compared risk of T2D between cluster 1 and cluster 5 according to the lifestyle score (Fig. 4B). In HCHS/SOL, compared with those in cluster 1 with a healthy lifestyle, participants in cluster 5 with an unhealthy lifestyle had a much higher risk of T2D (RR 1.77 [95% CI 1.19, 2.64]; $P = 0.005$), while those in cluster 5 with a healthy lifestyle did not show a significant higher risk of T2D (RR 1.14 [0.76, 1.70]; $P = 0.54$). Similar results were observed in UKBB. In the combined analysis of two cohorts, given a healthy lifestyle, participants in cluster 5 had a similar risk of T2D compared with those in cluster 1 (RR 1.03 [0.91, 1.18]; $P = 0.60$).

DISCUSSION

Based on five genetic risk scores representing different pathophysiological pathways related to T2D, we identified six clusters of individuals with prediabetes, with distinctive patterns of genetic risk scores and corresponding metabolic traits in U.S. Hispanic/Latino participants from HCHS/SOL, and confirmed five clusters in non-Hispanic White participants from UKBB. Different clusters of individuals with diabetes had similar levels of blood glucose and HbA_{1c} at baseline but differed in risk of progression to T2D during follow-up.

Clusters of prediabetes identified by our approach exhibited distinctive patterns of genetic scores and corresponding patterns of diabetes-related metabolic traits, which reflect

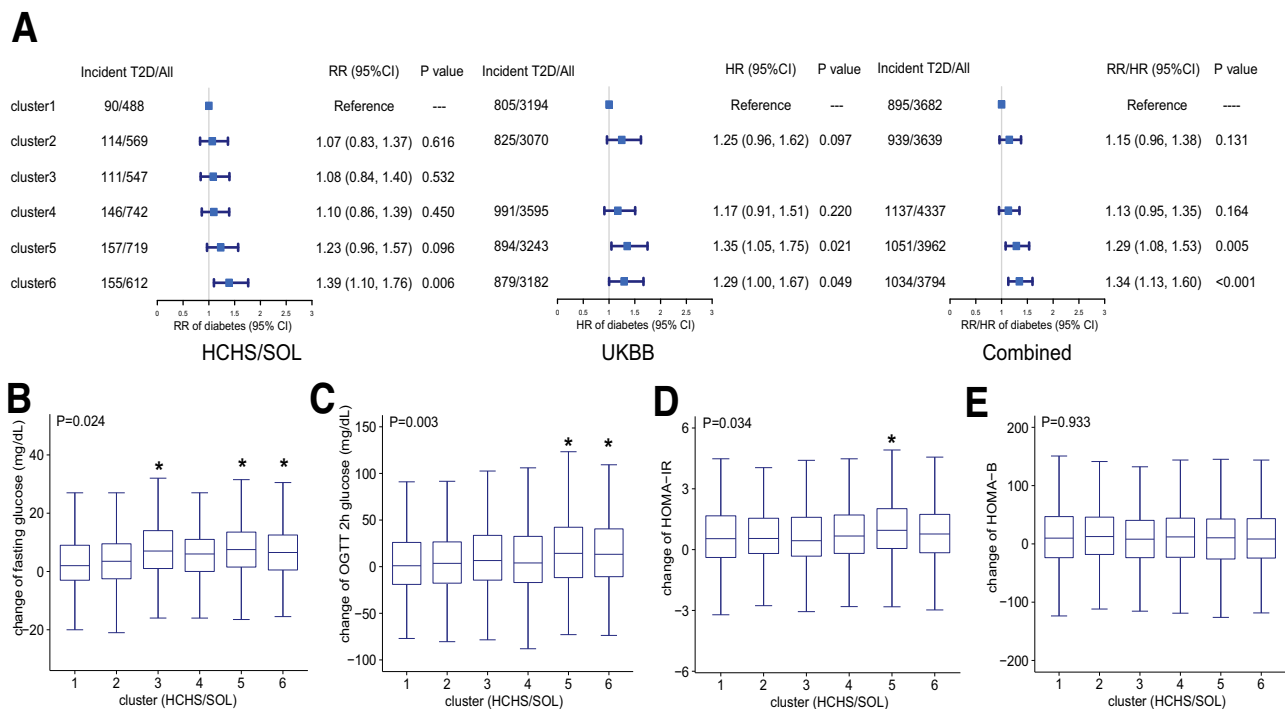


Figure 3—Clusters of prediabetes, incident T2D, and changes in glycemic traits during follow-up. **A**: Clusters of prediabetes and risk of T2D in HCHS/SOL, UKBB, and the combined studies. In HCHS/SOL, data are RRs and 95% CIs estimated by Poisson regression after adjustment for age, sex, U.S.-born status, Hispanic/Latino background, education, annual income, Alternate Healthy Eating Index 2020, smoking, drinking, physical activity, and eigenvectors derived from GWAS. In UKBB, data are hazard ratios (HRs) and 95% CIs estimated by Cox proportional hazards regression after adjustment for age, sex, education, Townsend deprivation score, diet score, smoking, drinking, physical activity, and eigenvectors derived from GWAS. In the combined analysis, results from HCHS/SOL and UKBB were combined using fixed-effects meta-analysis. **B–E**: Box plots of changes in fasting glucose, 2-h glucose after OGTT, HOMA-IR, and HOMA-B over 6 years across six clusters of individuals with prediabetes in HCHS/SOL. **P* < 0.05.

β -cell function, insulin resistance, obesity, and lipid metabolism. However, interestingly, baseline glycemic traits (e.g., fasting and 2-h post-oral load glucose levels, HbA_{1c}) and proportions of individuals with IGT and/or IFG were similar across different clusters, even for those with a higher risk of T2D during follow-up (i.e., cluster 5 and cluster 6), which did not show worse glycemic status at baseline compared with the other clusters. These findings imply that our genetic subtyping approach might help to identify individuals with a high risk of future T2D development before they manifest obvious glycemic change during T2D development and reflect the advantage of using the constant genetic information compared with T2D-related metabolic phenotypes, which are subject to change with disease progression or treatment (7). Genetic-based subtyping of prediabetes may have important implications in the early prevention of T2D, since previous phenotype-based studies identified high-T2D-risk subgroups that already showed worse glycemic status at baseline compared with low-T2D-risk subgroups (4). It is worth mentioning that clusters of prediabetes identified in this study showed no significant differences in demographic, socioeconomic, and behavioral factors in both cohorts, further supporting the robustness of results using genetic information with minimum confounding issues.

It is unexpected and intriguing that individuals with prediabetes in cluster 6, who had a very low proinsulin genetic score, showed the highest risk of T2D, since a lower proinsulin score was associated with a favorable metabolic profile (e.g., high HOMA-B, low levels of glycemic traits) and reduced risk of T2D in the current and previous studies (7). However, low proinsulin genetic score was associated with high fasting proinsulin levels adjusted for fasting insulin (7). The proinsulin-to-insulin ratio has been used to differentiate disproportionately elevated proinsulin from compensatory hyperinsulinemia (17), and a high proinsulin-to-insulin ratio might indicate disturbed insulin secretion (18) and has been associated with an increased risk of T2D (17). Thus, the high risk of T2D might be partially explained by disturbed insulin secretion among individuals with prediabetes in this cluster. Our further analysis showed a potential interaction between the proinsulin score and other genetic scores, suggesting that the genetic effects of other biological pathways on risk of T2D might be strengthened among individuals with a low proinsulin score, indicating disturbed insulin secretion. This might help to explain the high risk of T2D for individuals with prediabetes in cluster 6, since levels of the other four genetic scores were slightly above average in this cluster. However, the underlying biological mechanisms are unclear and need to be clarified.

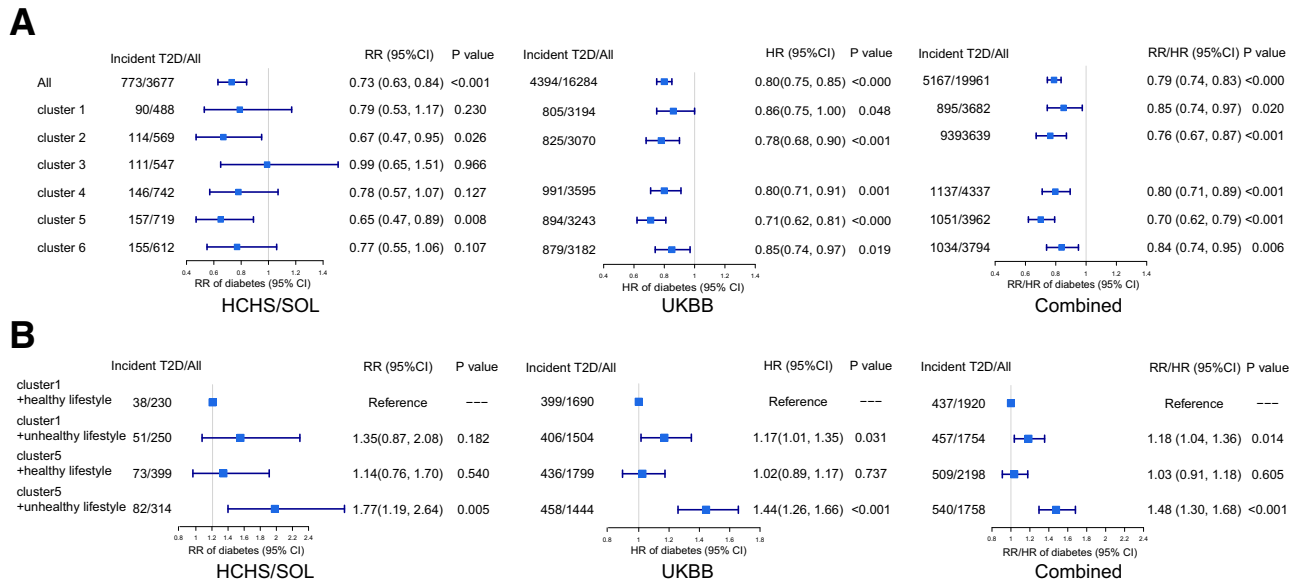


Figure 4—Association between healthy lifestyle and incident T2D across clusters of prediabetes. *A*: Association between healthy lifestyle and incident T2D among all individuals with prediabetes and across clusters of prediabetes. In HCHS/SOL, data are RRs and 95% CIs for incident T2D comparing healthy lifestyle (2nd and 3rd tertiles of the lifestyle score) with unhealthy lifestyle (1st tertile of the lifestyle score), estimated by Poisson regression after adjustment for age, sex, U.S.-born status, Hispanic/Latino background, education, annual income, and eigenvectors derived from GWAS. In UKBB, data are hazard ratios (HRs) and 95% CIs for incident T2D comparing healthy lifestyle (2nd and 3rd tertiles of the lifestyle score) with unhealthy lifestyle (1st tertile of the lifestyle score), estimated by Cox proportional hazards regression after adjustment for age, sex, education, Townsend deprivation score, and eigenvectors derived from GWAS. In the combined analysis, results from HCHS/SOL and UKBB were combined using fixed-effects meta-analysis. *B*: Risk of T2D in cluster 5 compared with the cluster 1 according to lifestyle (healthy lifestyle: 2nd and 3rd tertiles of the lifestyle score; unhealthy lifestyle: 1st tertile of the lifestyle score). In HCHS/SOL, data are RRs and 95% CIs for incident T2D estimated by Poisson regression after adjustment for the covariates mentioned above. In UKBB, data are HRs and 95% CIs for incident T2D estimated by Cox proportional regression after adjustment for the covariates mentioned above. In the combined analysis, results from HCHS/SOL and UKBB were combined using fixed-effects meta-analysis.

This study identified another potential genetic subtype of prediabetes with high T2D risk, cluster 5, which was characterized by high levels of lipodystrophy-like, liver-lipid, and proinsulin scores but low β -cell and obesity scores. The metabolic trait profile of this cluster was also mixed. It had not only the lowest levels of HDL cholesterol and highest levels of HOMA-IR among all clusters but also some favorable levels of metabolic traits (e.g., high HOMA-B, low BMI). Individuals with prediabetes in this cluster might share some features with the previously suggested normal-weight metabolically obese phenotype (19), and their elevated risk of T2D might be due to lipodystrophy or fat distribution-related (e.g., reduced subcutaneous adiposity) insulin resistance (7,19). However, lipodystrophy or fat distribution was not measured in this study, and future studies with these measures might help to explain the elevated risk of T2D in this cluster of individuals with prediabetes.

Another major finding of this study is that adherence to a healthy lifestyle was associated with a decreased risk of T2D across different clusters of individuals with prediabetes, except for cluster 3. This cluster was only identified in HCHS/SOL (U.S. Hispanic/Latino individuals) but not in UKBB (non-Hispanic White individuals), which might be due to genetic diversity of Hispanic/Latino admixed

populations with Amerindian, European, and African ancestries (12). The nonsignificant association between the healthy lifestyle score and risk of T2D in this cluster might be due to a relatively small sample size, but more studies in admixed populations are needed to validate our findings. Interestingly, we identified a cluster of individuals with prediabetes, cluster 5, which might have greater benefits from adherence to a healthy lifestyle in the prevention of T2D. Individuals with prediabetes in cluster 5 had an ~30% greater risk of T2D compared with those in cluster 1, but adherence to a healthy lifestyle could lower the risk of T2D to a similar level for both clusters. Our study further emphasizes the importance of healthy lifestyle for all individuals with prediabetes in reducing their risk of progression to diabetes (20,21) and suggests a potential genetic subtype of prediabetes that could have extra benefits of a healthy lifestyle. This may have important public health implications.

Our study has several strengths. We used data from two prospective cohorts with large sample sizes of individuals with prediabetes, especially in UKBB, which help to reduce inconsistency of clustering results among multiple studies caused by small sample sizes. Our subtyping approach only used genetic information, and thus, unlike phenotypic subtyping methods, the genetic subtyping results would not

change over time. Longitudinal data helped with examining the risk of incident T2D across clusters to point out the advantage of this approach. With well-collected lifestyle factors, we were able to examine the relationship between adherence to a healthy lifestyle and risk of T2D across different clusters.

This study also has several limitations. Both cohorts lacked data on some metabolic traits, such as blood proinsulin levels and fat distribution measures, which could help to better understand the underlying mechanisms for genetic subtypes of prediabetes with a high risk of T2D. Although our subtyping approach was based on multiple T2D-associated genetic variants and five different pathways related to T2D, these genetic variants might only explain a small portion of genetic risk for T2D. More studies are needed to reveal the genetic architecture of T2D and related quantitative traits, which will yield more biological pathways related to T2D and thus help with identifying more accurate subtypes of prediabetes based on genetic information. This study included Hispanic/Latino individuals and non-Hispanic White individuals, and although findings were consistent between these two study populations, further investigations in other racial/ethnic groups are needed given the heterogeneity in genetics of T2D across different populations (22). Another limitation is the absence of fasting glucose and 2-h OGTT measurements in UKBB, and thus, some undiagnosed incident T2D cases were not identified during follow-up, which may introduce misclassification of disease outcome, though diabetes diagnosis method based on primary care and hospital admission in the medical record has been widely used in UKBB (11).

In this analysis of U.S. Hispanic/Latino participants from HCHS/SOL and non-Hispanic White individuals from UKBB, we identified several clusters of individuals with prediabetes based on genetic information, and two potential genetic subtypes of prediabetes showed a relatively high risk of T2D over time. We also observed generally favorable relationships between healthy lifestyle and risk of T2D among individuals with prediabetes, regardless of their genetic subtypes, though individuals in one cluster with high T2D risk might have extra benefits in terms of risk reduction from a healthy lifestyle. This study extends the application of genetic information in the subtyping of prediabetes and provides useful information for prevention and intervention among people with a high T2D risk.

Acknowledgments. The authors thank Dr. Miriam S. Udler (Diabetes Unit, Endocrine Division, Department of Medicine, Massachusetts General Hospital, Boston, MA) for suggestions in constructing pPRSs. The authors also thank the staff and participants of HCHS/SOL for important contributions. A complete list of HCHS/SOL staff and investigators can be found in Lavange et al. (9) or at <https://sites.cscs.unc.edu/hchs>.

Funding. The HCHS/SOL is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001/N01-HC-65233), University of Miami (HHSN268201300004/N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002/N01-HC-65235), University of Illinois at Chicago

(HHSN268201300003/N01-HC-65236 Northwestern University), and San Diego State University (HHSN268201300005/N01-HC-65237). The following institutes, centers, and offices have contributed to 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 National Institutes of Health Office of Dietary Supplements. This work is supported by National Institute of Environmental Health Sciences grant R01ES030994 and National Institute of Diabetes and Digestive and Kidney Diseases grant R01DK119268. Other funding sources for this study include NHLBI grants R01HL060712, R01HL140976, R01HL105756, and R01HL136266; NIDDK grant R01DK120870; New York Regional Center for Diabetes Translation Research grant P30 DK111022; Diabetes Research Center grant DK063491 to the Southern California Diabetes Endocrinology Research Center; NIDDK grant UM1DK078616, and National Center for Advancing Translational Sciences Clinical and Translational Science Institute grant UL1TR00188. This research has been conducted using the UKBB Resource under application no. 56483.

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

Author Contributions. Y.L., G.-C.C., J.-Y.M., R.A., D.S.-A., M.L.D., A.P., J.M., K.M.P., J.I.R., K.D.T., Y.-D.I.C., S.W.-S., T.W., T.E.R., J.D.K., R.K., and Q.Q. interpreted the data and critically revised the manuscript. Y.L. and Q.Q. designed the study. Y.L. performed the statistical analyses and drafted the manuscript. Q.Q. supervised the study. All authors approved the final version of the manuscript and agreed to be accountable for the accuracy of the work. Q.Q. 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

- Hostalek U. Global epidemiology of prediabetes - present and future perspectives. *Clin Diabetes Endocrinol* 2019;5:5
- American Diabetes Association. 2. Classification and diagnosis of diabetes: *Standards of Medical Care in Diabetes—2021*. *Diabetes Care* 2021;44(Suppl. 1):S15–S33
- Stefan N, Fritsche A, Schick F, Häring HU. Phenotypes of prediabetes and stratification of cardiometabolic risk. *Lancet Diabetes Endocrinol* 2016;4:789–798
- Wagner R, Heni M, Tabák AG, et al. Pathophysiology-based subphenotyping of individuals at elevated risk for type 2 diabetes. *Nat Med* 2021;27:49–57
- Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol* 2018;6:361–369
- Zaharia OP, Strassburger K, Strom A, et al.; German Diabetes Study Group. Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: a 5-year follow-up study. *Lancet Diabetes Endocrinol* 2019;7:684–694
- Udler MS, Kim J, von Grothuss M, et al.; METASTROKE and the ISGC. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. *PLoS Med* 2018;15:e1002654
- Sorlie PD, Avilés-Santa LM, Wassertheil-Smoller S, et al. Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Ann Epidemiol* 2010;20:629–641
- Lavange LM, Kalsbeek WD, Sorlie PD, et al. Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Ann Epidemiol* 2010;20:642–649
- Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12:e1001779

11. Eastwood SV, Mathur R, Atkinson M, et al. Algorithms for the capture and adjudication of prevalent and incident diabetes in UK Biobank. *PLoS One* 2016;11:e0162388
12. 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;98:165–184
13. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;562:203–209
14. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 2010;26:1572–1573
15. Hu FB, Manson JE, Stampfer MJ, et al. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N Engl J Med* 2001;345:790–797
16. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;9:2579–2605
17. Then C, Gar C, Thorand B, et al. Proinsulin to insulin ratio is associated with incident type 2 diabetes but not with vascular complications in the KORA F4/FF4 study. *BMJ Open Diabetes Res Care* 2020;8:e001425
18. Lorenzo C, Hanley AJ, Rewers MJ, Haffner SM. Disproportionately elevated proinsulinemia is observed at modestly elevated glucose levels within the normoglycemic range. *Acta Diabetol* 2014;51:617–623
19. Yaghootkar H, Scott RA, White CC, et al. Genetic evidence for a normal-weight “metabolically obese” phenotype linking insulin resistance, hypertension, coronary artery disease, and type 2 diabetes. *Diabetes* 2014;63:4369–4377
20. Knowler WC, Barrett-Connor E, Fowler SE, et al.; Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002;346:393–403
21. Diabetes Prevention Program Research Group; Knowler WC, Fowler SE, Hamman RF, et al. 10-Year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet* 2009;374:1677–1686
22. Pofus LM, Darst BF, Highland H, et al.; 23andMe Research Team; DIAMANE Hispanic/Latino Consortium; Meta-Analysis of Type 2 Diabetes in African Americans Consortium. Genetic discovery and risk characterization in type 2 diabetes across diverse populations. *HGG Adv* 2021;2:100029