

## **SUPPLEMENTAL MATERIALS**

**Title: Proteomic predictors of incident diabetes: Results from the Atherosclerosis Risk in Communities (ARIC) Study**

**Authors: Mary R. Rooney, Jingsha Chen, Justin B. Echouffo-Tcheugui, Keenan A. Walker, Pascal Schlosser, Aditya Surapaneni, Olive Tang, Jinyu Chen, Christie M. Ballantyne, Eric Boerwinkle, Chiadi E. Ndumele, Ryan T. Demmer, James S. Pankow, Pamela L. Lutsey, Lynne E. Wagenknecht, Yujian Liang, Xueling Sim, Rob van Dam, E Shyong Tai, Morgan E. Grams, Elizabeth Selvin, Josef Coresh**

**Supplemental Tables: 10**

**Supplemental Figures: 4**

## SUPPLEMENTAL METHODS

### Measurement of Covariates in ARIC

At the baseline (visit 2, 1990-1992) clinic visit, participants reported their smoking status and family history of diabetes (self-reported only at visit 1, maternal or paternal). Body mass index (BMI) was calculated based on measured weight and height (height was measured at visit 1 which occurred in 1987-1989, ~3 years prior to baseline). Estimated glomerular filtration rate (eGFR) was calculated based on cystatin-C and creatinine using the 2021 Chronic Kidney Disease Epidemiology equation.(1) Total and HDL cholesterol were measured using standard methods. Systolic blood pressure was based on the average of the 2<sup>nd</sup> and 3<sup>rd</sup> blood pressure reading after 5 minutes rest. Participants were asked to bring bottles of any medications used in the prior 2 weeks; medication bottles were transcribed and coded. Physical activity was measured at visit 1 (1987-1989) based on the validated Baecke sport index questionnaire which yields a score ranging from 1 to 5.(2) Hemoglobin A1c (HbA1c) was measured in whole blood using high-performance liquid chromatography instruments (Tosoh A1c 2.2 Plus Glycohemoglobin Analyzer method in 2003–2004 and the Tosoh G7 method in 2007–2008; Tosoh Corporation) that were standardized to the Diabetes Control and Complications Trial assay.(3) Fasting glucose was measured in serum using a hexokinase method.

### QC for Protein Measurements in ARIC

We ran blind duplicates for 514 (~5%) participants with available SOMAmer data at visit 2. The median coefficient of variation based on the Bland-Altman ( $CV_{BA}$ ) method(4) was 6.3%. The median split sample reliability coefficient (intraclass correlation coefficient) for visit 2 QC was 0.93. Of the 5,284 SOMAmers quantified using visit 2 samples, we excluded 329 SOMAmers with a  $CV_{BA} > 50\%$  or variance  $< 0.01$  at visits 2, 3 (1993-1995) and 5 (2011-2013), SOMAmers that bound to mouse Fc-fusion or a contaminant, or non-proteins.

### External Replication in Singapore MEC

#### *Study Cohort*

The Multi-Ethnic Cohort (MEC) is a cohort study under the Singapore Population Health Studies (SPHS) that aims to discover how lifestyle factors, physiological factors, genetic factors and their interactions impact the development of common health conditions in Singapore (<https://blog.nus.edu.sg/sphs/>). The baseline recruitment was completed between 2004 and 2010 and details of the study are described here [1]. Between 2011 and 2016, the participants were invited for a follow-up. At both baseline and revisit, participants completed an interview-administered questionnaire and clinical examination. The questionnaires collected demographic information, detailed lifestyle behaviors, personal and family medical histories, and medication usage. At the health examination, anthropometric measures, blood pressures, and blood was taken for lipid and glycemic biomarker measurements. Informed consent was obtained from all participants. All human biological samples were collected in accordance with ethical regulations and protocols.

For the incident type 2 diabetes (T2D) cases, they were identified by the date of first identification of diabetes (diagnosis date through national record linkage with health records or date of revisit in which diabetes was identified based on personal history of T2D, fasting glucose  $\geq 7$  mmol/l or random glucose

$\geq 11 \text{ mmol/l}$  or  $\text{HbA1c} \geq 6.5\%$ ). Controls were selected by incident density sampling matched (1 case to 2 controls) by age ( $\pm 5$  years), gender, ethnicity and date of blood collection ( $\pm 2$  year).

### *Statistical Analysis*

In MEC, proteomic biomarkers were log<sub>2</sub>-transformed and then were standardized for analyses with T2D. The logistic model adjusted for baseline age (continuous), sex (male/female), ethnicity (Chinese/Malay/Indian), cystatin-C measured on SomaScan (log<sub>2</sub>transformed, continuous), current smoking status (yes/no), total cholesterol (continuous), HDL cholesterol (continuous), systolic blood pressure (continuous), BMI (continuous). As MEC did not contain a comparable physical activity variable, the model utilized the weekly activity level (met-hr/wk) for all leisure-time physical activities variable instead. However, with or without the leisure-time physical activity variable, similar associations results were observed.

**LIST OF SUPPLEMENTAL FIGURES**

**Supplemental Figure 1.** Scatterplot of effect estimates for the protein-diabetes associations in ARIC and MEC.

**Supplemental Figure 2.** Calibration curves of the risk model (Model 4 covariates and proteins identified using elastic net with Model 4 adjustment in the 2/3 discovery) within discovery (Panel A) and within the 1/3 internal validation set (Panel B).

**Supplemental Figure 3.** (A) Biological pathways and (B) upstream regulators associated with ~20 year diabetes risk in discovery.

**Supplemental Figure 4.** Regional association plots of SHBG (A), ATP1B2 (B), and type 2 diabetes (C) GWAS.

**LIST OF SUPPLEMENTAL TABLES**

**Supplemental Table 1.** Protein biomarkers associated with ~20 year risk of incident diabetes in discovery: The ARIC Study

**Supplemental Table 2.** Results of the proteins associated with ~20 year diabetes risk, adjusted for demographics, cardiometabolic risk factors, fasting glucose, and HbA1c: The ARIC Study

**Supplemental Table 3.** Results of the top 50 proteins associated with ~20 year risk of diabetes, adjusted for demographics, cardiometabolic risk factors, and 10 principal components of proteins: The ARIC Study

**Supplemental Table 4.** List of proteins identified in discovery that replicated in the internal validation

**Supplemental Table 5.** Plasma protein associations with type 2 diabetes in the MEC cohort (n = 1,838): with 'Leisure-time physical activities' variable

**Supplemental Table 6.** Calibration of predictive model derived within the overall ARIC discovery: mean observed and predicted risk of deciles within the internal validation set

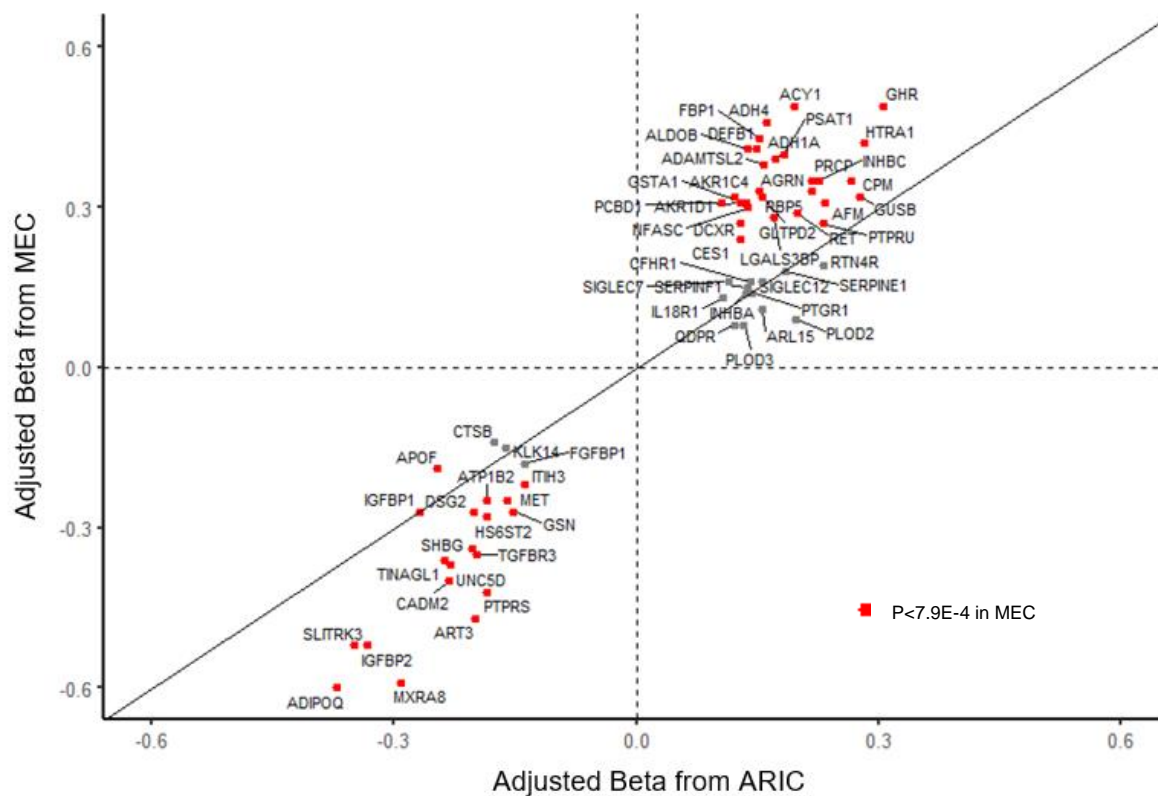
**Supplemental Table 7.** The top 25 canonical pathways determined by IPA from proteins associated with incident diabetes

**Supplemental Table 8.** The top 25 upstream regulators predicted to be activated and inhibited by diabetes-associated proteins

**Supplemental Table 9.** Genetic models for the proteins that validated internally in association with diabetes risk

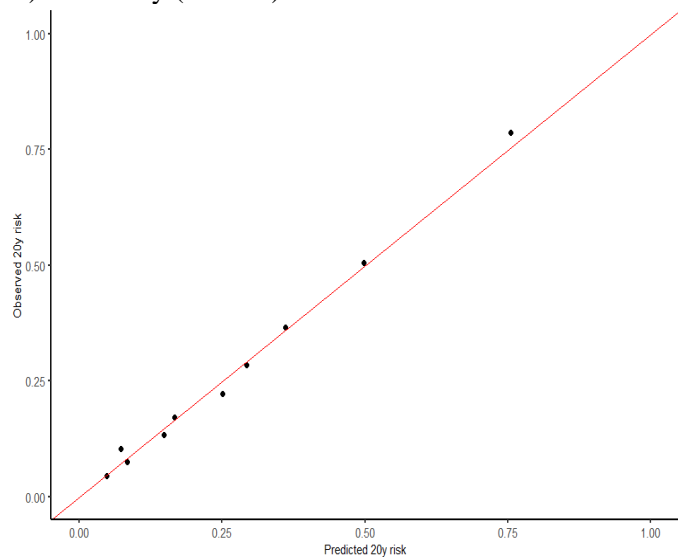
**Supplemental Table 10.** Summary of proteomics of incident diabetes publications

**Supplemental Figure 1.** Scatterplot of effect estimates for the 63 protein-diabetes associations in ARIC and MEC. Beta estimates in ARIC from Cox regression and in MEC from conditional logistic regression. The MEC case-control design used a risk-set sampling design so OR can approximate the HR. Betas adjusted for age, sex, race, eGFR (eGFRcrcls in ARIC,  $\log_2(\text{Soma cystatin-C})$  in MEC), physical activity (sport index in ARIC, leisure time mets per week in MEC), current smoking, total cholesterol, HDL-cholesterol, systolic blood pressure, hypertension medication use, BMI. Red dots indicate betas had  $p < 0.05/63$  in Singapore MEC.

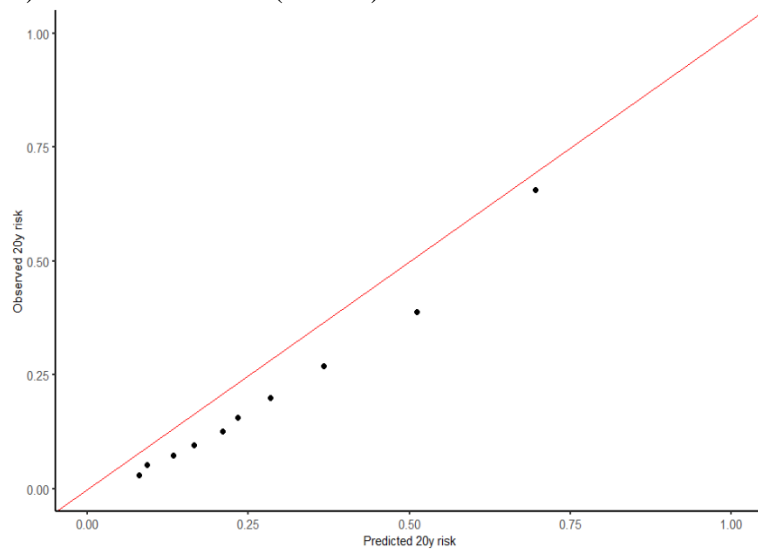


**Supplemental Figure 2.** Calibration curves of the risk model (Model 4 covariates and proteins identified using elastic net with Model 4 adjustment in the 2/3 overall discovery) within discovery (Panel A) overall and within the 1/3 internal validation set (Panel B).

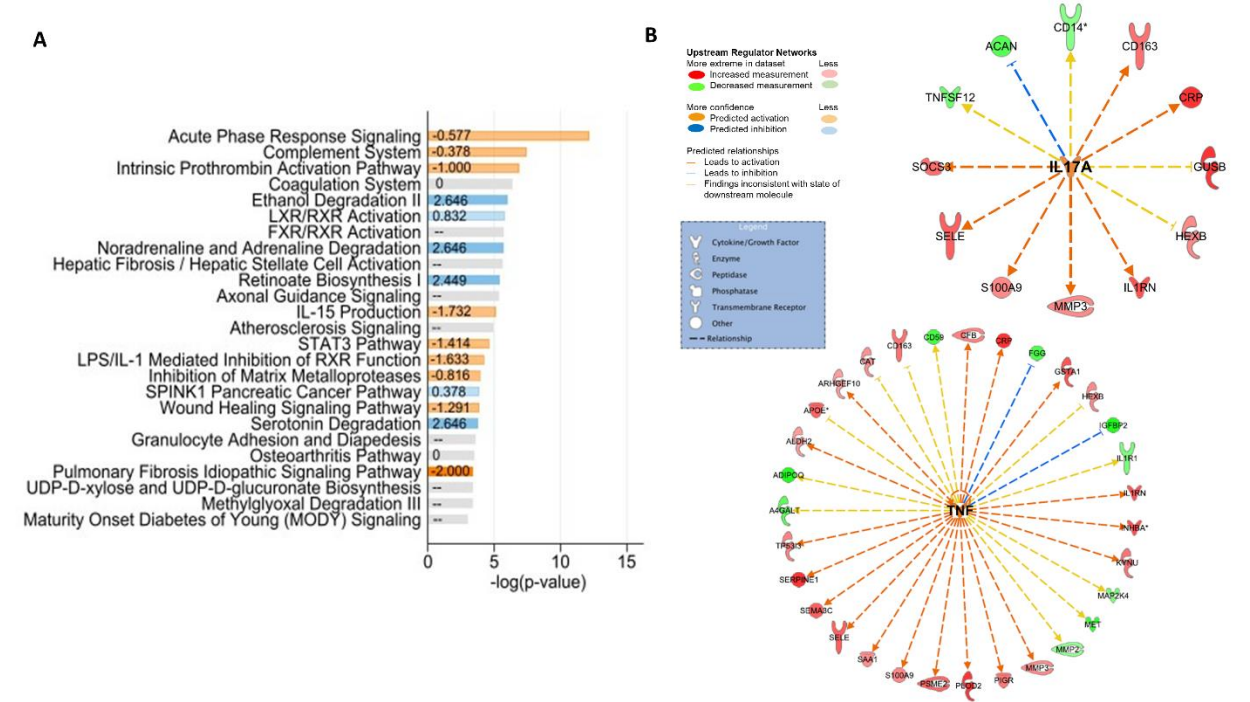
A) Discovery (Overall)



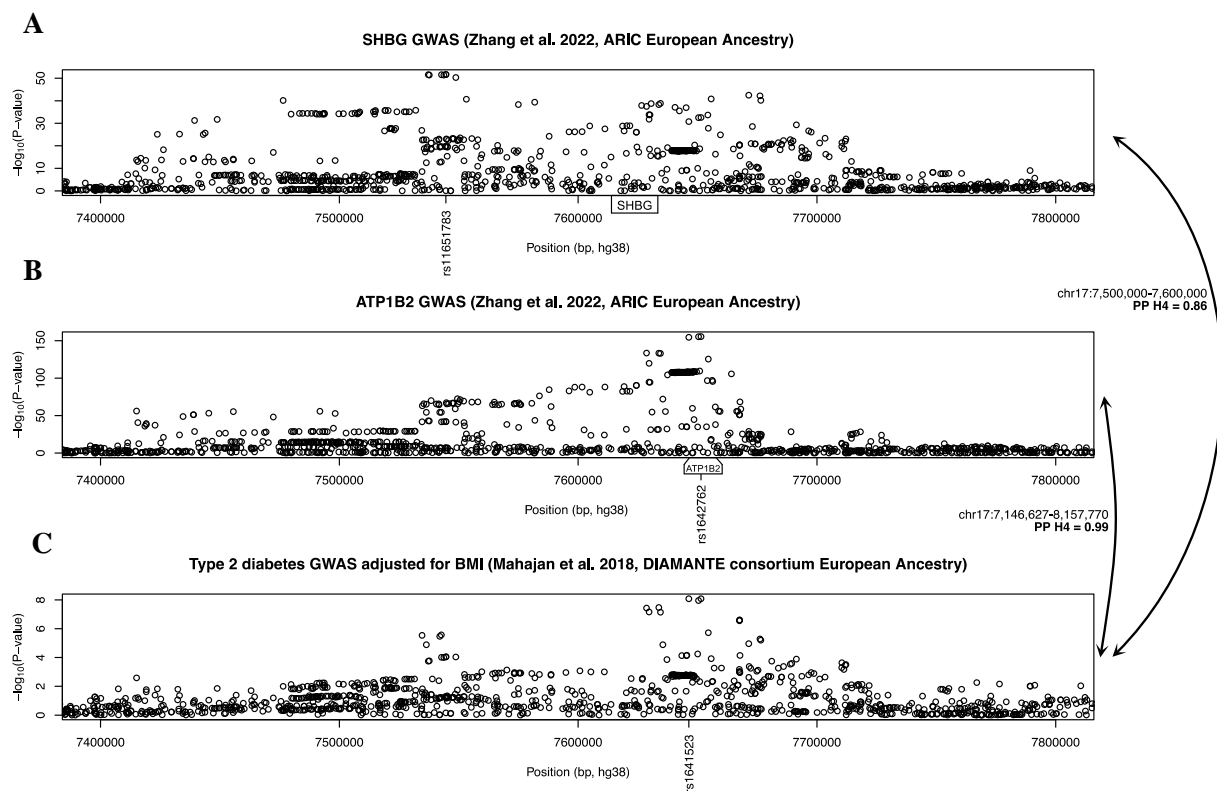
B) Internal Validation (Overall)



**Supplemental Figure 3.** (A) Biological pathways and (B) upstream regulators associated with ~20 year diabetes risk in discovery.



**Supplemental Figure 4.** Regional association plots of SHBG (A), ATP1B2 (B), and type 2 diabetes (C) GWAS on chromosome 17 from 7.4MB to 7.8MB. Gene regions of SHBG and ATP1B2 were marked on the respective x-axis and the genetic variant with the lowest p-value per trait were labeled.



**SUPPLEMENTAL REFERENCES**

1. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, Crews DC, Doria A, Estrella MM, Froissart M, Grams ME, Greene T, Grubb A, Gudnason V, Gutiérrez OM, Kalil R, Karger AB, Mauer M, Navis G, Nelson RG, Poggio ED, Rodby R, Rossing P, Rule AD, Selvin E, Seegmiller JC, Shlipak MG, Torres VE, Yang W, Ballew SH, Couture SJ, Powe NR, Levey AS: New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race. *New England Journal of Medicine* 2021;
2. Baecke JA, Burema J, Frijters JE: A short questionnaire for the measurement of habitual physical activity in epidemiological studies. *Am J Clin Nutr* 1982;36:936-942
3. Selvin E, Steffes MW, Zhu H, Matsushita K, Wagenknecht L, Pankow J, Coresh J, Brancati FL: Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. *The New England journal of medicine* 2010;362:800-811
4. Bland JM, Altman DG: Measurement error proportional to the mean. *BMJ (Clinical research ed)* 1996;313:106