American Diabetes Association. Diabetes Care

Proteomic analyses in diverse populations improved risk prediction and identified new drug targets for type 2 diabetes

Journal:	*Diabetes Care*
Manuscript ID	DC23-2145.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Yao, Pang; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Iona, Andri; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Pozarickij, Alfred; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Said, Saredo; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Wright, Neil; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Lin, Kuang; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Millwood, Iona; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Fry, Hannah; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Kartsonaki, Christiana; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Mazidi, Mohsen; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Chen, Yiping; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Bragg, Fiona; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Liu, Bowen; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit

	(CTSU) Yang, Ling (Oxford); University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Liu, Junxi; University of Oxford Nuffield Department of Population Health Avery, Daniel; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Schmidt, Dan; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Sun, Dianjiani; Peking University School of Public Health Department of Epidemiology and Biostatistics Pei, Pei; Peking University Center for Public Health and Epidemic Preparedness & Response Lv, Jun; Peking University Health Science Center, Department of Epidemiology and Biostatistics Yu, Canqing; Peking University, Department of Epidemiology and Biostatistics, School of Public Health Hill, Michael; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Bennett, Derrick; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Walters, Robin; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Li, Liming; Peking University School of Public Health Department of Epidemiology and Biostatistics Clarke, Robert; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Du, Huaidong; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) Chen, Zhengming; University of Oxford Nuffield Department of Population Health, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU)

SCHOLARONE™
Manuscripts

Proteomic analyses in diverse populations improved risk prediction and identified new drug targets for type 2 diabetes

Pang Yao[1], Andri Iona[1], Alfred Pozarickij[1], Saredo Said[1], Neil Wright[1], Kuang Lin[1], Iona Millwood[1,2], Hannah Fry[1,2], Christiana Kartsonaki[1,2], Mohsen Mazidi[1], Yiping Chen[1,2], Fiona Bragg[1], Bowen Liu[1], Ling Yang[1,2], Junxi Liu[1], Daniel Avery[1,2], Dan Schmidt[1,2], Dianjianyi Sun,[3,4,5] Pei Pei[3], Jun Lv[3,4,5], Canqing Yu[3,4,5], Michael Hill[1], Derrick Bennett[1,2], Robin Walters[1,2], Liming Li[3,4,5], Robert Clarke[1], Huaidong Du[1,2], Zhengming Chen[1,2] on behalf of China Kadoorie Biobank Collaborative Group‡

1. Clinical Trial Service Unit & Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
2. Medical Research Council Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
3. Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China
4. Peking University Center for Public Health and Epidemic Preparedness & Response, Beijing, China
5. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China

‡ *Members of the CKB Collaborative Group are shown in the Appendix*

Address for correspondence

Dr Huaidong Du
Nuffield Department of Population Health
University of Oxford
Big Data Institute Building
Old Road Campus, Oxford OX3 7LF, UK
Email: Huaidong.du@ndph.ox.ac.uk
Telephone: +44 (0)1865 743831

Professor Zhengming Chen
Nuffield Department of Population Health
University of Oxford
Big Data Institute Building
Old Road Campus, Oxford OX3 7LF, UK
Email: Zhengming.chen@ndph.ox.ac.uk
Telephone: +44 (0)1865 743839

Running title (<47 characters): New proteins for predicting and treating T2D

Twitter summary (<200 characters): Integrated proteomic and genetic analyses in diverse populations identified novel protein biomarkers for improved risk prediction and development of novel drug targets for type 2 diabetes

Word count: Abstract 250, Text 3643

2 Tables, 2 Figures and 11 eTables / eFigures

(CKB Research Track No.: 2022-0036)

Revised on 5 March 2024

Abbreviations:

AGEN	Asian Genetic Epidemiology Network
AUC	Area under the curve
CAD	Coronary artery disease
cis-pQTL	*cis* protein quantitative trait locus
CKB	China Kadoorie Biobank
CKB	Creatine kinase B-type
CPM	Carboxypeptidase M
CTSD	Cathepsin D
DIAMANTE	Diabetes Meta-Analysis of Trans-Ethnic association studies
ENTR1	Endosome-associated-trafficking regulator 1
ESM1	Endothelial cell-specific molecule 1
FDR	False discovery rate
FGFBP1	Fibroblast growth factor-binding protein 1
GHR	Growth hormone receptor
GLB1	Beta-galactosidase
GO	Gene Ontology
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association study
HbA1c	Hemoglobin A1c
HR	Hazards ratio
HuGE	Human Genetic Evidence
ICD-10	International Classification of Diseases 10th Revision
IHD	Ischemic heart disease
IGFBP	Insulin-like growth factor-binding protein
IGSF9	Protein turtle homolog A
IL18R1	Interleukin-18 receptor 1
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOD	Limit of detection
LPL	Lipoprotein lipase
MI	Myocardial infarction
MR	Mendelian randomization
NPX	Normalized Protein eXpression
NRI	Net reclassification index
OR	Odds ratio
PheWAS	Phenome-wide association study
PON3	Serum paraoxonase/lactonase 3
PRDX2	Peroxiredoxin 2
PRCP	Lysosomal Pro-X carboxypeptidase
RIDA	2-iminobutanoate/2-iminopropanoate deaminase
SD	Standard deviation
SHBG	Sex hormone-binding globulin
SLC2A1	Solute carrier family 2 member 1
T2D	Type 2 diabetes
T2DKP	Type 2 Diabetes Knowledge Portal
UKB	UK Biobank
VNN1	Pantetheinase

CONFIDENTIAL-For Peer Review Only

Abstract (250 words)

Objective: Integrated analyses of plasma proteomics and genetic data in prospective studies can help assess the causal relevance of proteins, improve risk prediction and discover novel protein drug targets for T2D.

Research Design and Methods: We measured plasma levels of 2923 proteins using OLINK Explore among ~2000 randomly selected participants from CKB without prior diabetes at baseline. Cox regression assessed associations of individual protein with incident T2D (n=92 cases). Proteomic-based risk models were developed with discrimination, calibration, reclassification assessed using AUC, calibration plots and NRI, respectively. Two-sample MR analyses using *cis*-pQTLs identified in GWAS of CKB and UKB for specific proteins were conducted to assess their causal relevance for T2D, along with colocalization analyses to examine shared causal variants between proteins and T2D.

Results: Overall 33 proteins were significantly associated (FDR<0.05) with risk of incident T2D, including IGFBP1, GHR and amylase. The addition of these 33 proteins to conventional risk prediction model improved AUC from 0.77 (0.73-0.82) to 0.88 (0.85-0.91) and NRI by 38%, with predicted risks well calibrated with observed risks. MR analyses provided support for the causal relevance for T2D of ENTR1, LPL and PON3, with replication of ENTR1 and LPL in Europeans using different genetic instruments. Moreover, colocalization analyses showed strong evidence (PH4>0.6) of shared genetic variants of LPL and PON3 with T2D.

Conclusion: Proteomic analyses in Chinese adults identified novel associations of multiple proteins with T2D with strong genetic evidence supporting their causal relevance and potential as novel drug targets for prevention and treatment of T2D.

Key words: *Proteomics, Prospective studies, T2D, Genetics, Risk prediction, Drug targets*

Article Highlights (130 / Max 130 words)

- ***Why did we undertake this study?***

  Proteo-genomic analyses in prospective studies can help discover novel protein biomarkers for T2D.

- ***What is the specific question(s) we wanted to answer?***

  We assessed the associations of 2923 proteins with risk of incident T2D, utility of specific proteins for risk prediction and druggability of certain causal proteins.

- ***What did we find?***

  Thirty-three proteins were associated with incident T2D. Adding these 33 proteins to the conventional risk factors substantially improved T2D risk prediction. Genetic analyses identified causal relevance of three proteins, with one (PON3) highly expressed in liver tissue and potential relevant for T2D prevention and treatment.

- ***What are the implications of our findings?***

  These findings highlighted the importance of proteomics in prospective studies to improve risk prediction and discover novel drug targets for T2D.

CONFIDENTIAL-For Peer Review Only

Introduction

Globally type 2 diabetes (T2D) affects >530 million adults,[1] causing substantial risks of premature death and macro- and micro-vascular complications. China has the largest number of people with diabetes (>140 million) in the world and the prevalence is still rising.[1] Several important modifiable risk factors for T2D are established (e.g. adiposity, lack of physical activity and suboptimal diet), which account for about 70% of new cases globally.[2] These risk factors have been widely used, typically in combination with blood glucose and/or HbA1c, to predict risk of T2D and inform prevention and treatment decision in diverse populations.[3, 4] Recently, GWAS of T2D in diverse populations identified >240 common genetic variants, including >180 in East Asian populations.[5, 6] However, the mechanisms underlying many of these associations remain to be elucidated. Plasma proteins play a central role in human biology and represent a primary source of therapeutic targets.[7] Analyses of circulating protein biomarkers, particularly when integrated with genetic data, in population and clinical studies, can help clarify disease aetiology, improve risk prediction and early diagnosis, and discover novel and repurposing therapeutic targets for treatment of T2D and other major diseases.[8-12]

Previous studies of plasma proteins and T2D have highlighted the roles of several specific proteins (e.g. IGFBP1, IGFBP2, GHR and SHBG) in aetiology of T2D.[12-16] Advances in high throughput proteomic assays now enable measurement of several thousand proteins,[17-19] and their application in population and clinical studies of primarily European ancestry populations have identified several novel protein biomarkers for T2D.[12-16] However, little is known about the relevance of protein biomarkers for T2D in non-European ancestry populations including Chinese where the disease rates, distribution of risk factors and genetic architecture differ greatly from European ancestry populations.

CONFIDENTIAL-For Peer Review Only

Moreover, few previous studies undertook detailed genetic analyses (e.g. MR and colocalization analyses) to assess the causal relevance of specific proteins for T2D.[16, 20]

We undertook an integrated analysis of observational and genetic data of ~3000 proteins with incident T2D in ~2000 adults selected from the China Kadoorie Biobank. The present report aims to: (i) identify plasma proteins significantly associated with incident T2D; (ii) assess the utility of selected proteins for prediction of T2D risk; (iii) use *cis*-pQTLs identified in GWAS for proteins to assess their causal relevance for T2D via two-sample MR and separate colocalization analyses; and (iv) clarify the mechanisms of action for specific proteins using PheWAS, tissue expression and other experimental evidence.

Methods

Study population and data collection

Details of the CKB study design, methods, and participants have been previously reported.[21] Briefly, a total of 512,715 participants aged 30-79 years were enrolled from 10 (5 urban, 5 rural) geographically diverse regions in China. At baseline (June 2004-July 2008) and at three subsequent resurveys in a ~5% subset (in 2008, 2014 and 2021, respectively), detailed data were collected on socio-demographic characteristics, smoking, alcohol consumption, diet, physical activity, personal (e.g. IHD, stroke and T2D) and family medical history, along with physical and blood measurements (e.g. blood pressure, BMI, and random blood glucose but not HbA1c).

Follow-up of CKB participants was through linkage to established mortality (cause-specific) and morbidity (including T2D) registries, and to the nationwide health insurance system which records all hospitalised episodes.[21] All disease events and causes of death were ICD-10 coded by trained health workers, blinded to baseline information, and checked and integrated centrally. Ethical approval was obtained from relevant international, national and

regional ethics committees or institutional research boards.[21] All participants provided written informed consent.

The present study involved 2026 participants selected as a subcohort for a nested case-subcohort study of IHD in CKB.[22] They were randomly selected from a population subset of 69,353 genotyped participants who had no prior history of CVD nor statin use at baseline and were genetically unrelated to each other. Among these 2026 participants, 130 having prevalent diabetes at baseline who were analysed separately for internal replication but excluded from the main analyses. During 11 years of follow-up (up to 1.1.2018), 92 individuals developed incident T2D (ICD10: E10-E14) among the 1896 participants included in the main analyses.

Proteomics assay

Stored baseline plasma samples from participants were retrieved, thawed, and sub-aliquoted to multiple aliquots, with one aliquot (100 μL; for batch 1 assay) shipped on dry ice to the OLINK Biosciences Laboratory at Uppsala, Sweden, and one aliquot (for batch 2 assay) shipped subsequently to OLINK lab at Boston, USA, for multiplex proximity extension assay of proteins. Batch 1 covered 1463 unique proteins first released by the OLINK, while the batch 2 covered a further 1460 unique proteins released subsequently by the OLINK. To minimize inter- and intra-run variation, the samples were randomized across plates and normalized using both an internal control (extension control) and an inter-plate control and then transformed using a pre-determined correction factor.

Details of the OLINK assay performance and validation have been previously reported.[18] The LOD were determined using negative control samples (buffer without antigen). A sample is flagged as having QC warning if incubation control deviates more than a pre-determined value (±0.3) from the median value of all samples on the plate. The pre-

CONFIDENTIAL-For Peer Review Only

processed data were provided in the arbitrary unit Normalized Protein eXpression (NPX) on a log2 scale. The present analyses has a total of 2941 proteins (2923 unique proteins), including 1472 proteins (1463 unique proteins) in batch 1 and 1469 proteins (1460 unique proteins) in batch 2.

Statistical analysis

Plasma protein levels were standardized (i.e. values of each protein were divided by their SD) and analysed as continuous variables. In observational analysis, Cox and logistic regression models were used to estimate adjusted HRs and ORs (and 95% CI) for incident and prevalent diabetes, respectively. All analyses were adjusted for age, age^2, sex, study area, fasting time, ambient temperature, plate ID, education (4 categories: no formal school, primary school, secondary school, high school and above), smoking (3 categories: never, occasional or ex-regular, and regular smoker), alcohol drinking (3 categories: never, occasional or ex-regular, and weekly), physical activity (MET-h), family history of diabetes (binary) and BMI. Sensitivity analyses were conducted by further adjusting dietary variables, and by adjusting the effect of plate using residual from regression of protein values on plate as a covariate in the models. The cross-sectional analysis for random plasma glucose (per SD) was restricted to participants without prior diabetes at baseline. Analyses were also conducted in UKB for diabetes incidence, blood glucose and HbA1c (see **eAppendix**).

For proteins significantly associated with incident diabetes, we further (i) examined the shape of the associations with T2D by quartiles of individual proteins; (ii) assessed the performance of proteomic-based risk prediction of T2D in CKB, with discrimination and calibration utilities assessed using AUC and calibration plots with the Hosmer–Lemeshow test, respectively. Reclassification was measured using both the percentile-based NRI with

CONFIDENTIAL-For Peer Review Only

deciles of relative risk as reference categories, and the continuous NRI. The proteomic risk model was internally validated using 1000 bootstrap method and compared and combined with conventional risk prediction model in Chinese,[23] with further external validation in the UKB; and iii) conducted GO and KEGG enrichment analyses using clusterProfiler (v.4.2.2),[24] to determine which biological functions or processes were significantly enriched based on hypergeometric tests.

For proteins showing significant associations with T2D in observational analyses, a two-sample MR using Wald ratio method,[25, 26] was conducted using (i) *cis*-pQTLs obtained from GWAS of CKB, with lookups in AGEN consortium of East Asian adults including 66,677 diabetes cases;[6] and (ii) *cis*-pQTLs obtained from GWAS of UKB, with lookups in the DIAMANTE Consortium of European descent including 80,154 diabetes cases and 853,816 controls.[5] Colocalization was performed only for those proteins that had 95% credible sets identified by fine mapping in both AGEN T2D and CKB *cis*-pQTL datasets. Fine mapping was performed using susieR (v0.12.16) and colocalization was performed using coloc (v5.2.1) packages in R.

For proteins showing significant genetic associations with T2D, we screened protein expression database of GTEx to study the tissue-specific role of the causal proteins in diabetes. We further searched T2DKP for associations of (i)) *cis*-pQTLs from both CKB and UKB with a range of phenotypes using a *P* value threshold of 5×10^{-8} and (ii) genes with available diseases and traits using a HuGE Scores threshold of 10, indicating stong evidence for the causal relevance of such proteins for diseases or traits.[27]

Figure 1 provides an overview of main analytic approaches. All statistical analyses were performed using R version 4.1.2. Benjamini-Hochberg FDR was used to correct for multiple testing.

CONFIDENTIAL-For Peer Review Only

Results

Among the 1896 participants included in the main analyses, the mean (SD) age was 51.3 (10.4) years, 62.1% were women, and 50.6% were urban residents, which, along with many other baseline characteristics, were similar to those in the overall genotyped CKB cohort (**eTable 1**).

Observational associations of proteins with diabetes

After adjusting for conventional risk factors, 33 proteins were significantly associated at 5% FDR with risk of T2D (batch 1/2: 24/9) (**Figure 2**).The associations were typically log-linear throughout the full ranges of levels of specific proteins examined (**eFigure 1**), although the adjusted HRs (per 1SD higher protein level) varied, from 1.38 to 1.98 for those showing positive associations (23 proteins) and from 0.48 to 0.70 for those showing inverse associations (10 proteins) with T2D (**eFigure 2**). Proteins showing the strongest positive associations were VNN1 (1.98, 95% CI 1.49-2.64), GHR (1.80, 1.35-2.39), PRCP (1.78, 1.32-2.40), CPM (1.68, 1.28-2.22) and IGSF9 (1.67, 1.32-2.11). The proteins showing strongest inverse associations included IGFBP2 (0.48, 0.36-0.65), CKB (0.59, 0.44-0.78), IGFBP1 (0.61, 0.47-0.80), LPL (0.61, 0.48-0.78) and ESM1 (0.62, 0.48-0.69). Six proteins (IGFBP2, VNN1, IGSF9, GLB1, PON3, and RIDA) were significantly associated with incident T2D after applying Bonferroni multiple test correction. Further adjustments for fresh fruit, red meat consumption and a healthy diet score did not alter the results, as were use of the residual from regression of protein values on plate as a covariate in the models.

In internal replication analyses, most of these 33 proteins were significantly associated with blood glucose (29/33; 88%) levels or prevalent diabetes (30/33; 91%) (**eFigure 3**). Moreover, all 33 proteins were externally replicated in UKB (at 5% FDR) for blood glucose, HbA1c, and incident T2D (except one protein), although the effect sizes varied (**eFigure**

CONFIDENTIAL-For Peer Review Only

4). Of these 33 proteins, most proteins had similar HRs in observational analyses of both studies, while the HRs for 3 proteins (VNN1, GHR, IGFBP2) differed somewhat but all directionally concordant with each other.

These 33 proteins were only moderately correlated with each other, with 99.5% of protein pairs having correlation coefficients ranging from -0.7 to +0.7 (**eFigure 5**).

Risk prediction of incident T2D

In CKB the conventional risk prediction model without blood glucose had an AUC of 0.754 (0.710-0.798), increasing to 0.774 (0.730-0.818) with addition of blood glucose (**Table 1**). Proteomic-based model (33 proteins) alone had AUC of 0.824 (0.786-0.862), which significantly outperformed the conventional models. The addition of top 10 or all 33 proteins to conventional risk factors plus blood glucose model yielded an AUC of 0.844 (0.803-0.885) and 0.876 (0.846-0.906), respectively. For NRI the corresponding values were 28% (15-41%) and 38% (24-52%), respectively, using categorical approach, rising to 84% (65%-103%) and 97% (79%-115%) when using continuous approach (**eTable 2**). The observed and predicted risk of T2D showed excellent calibration (the Hosmer–Lemeshow test: χ^2=3.4, *P*=0.90; **eFigure 6**). The application of these same proteins identified in CKB to UKB yielded comparable results for prediction of T2D (**eTables 3-4**).

Enrichment analysis

In enrichment analyses of 33 proteins, hydrolase activity was identified as the top biological pathway (**eFigure 7**). Other pathways such as growth factor binding and insulin-like growth factor binding, were also among the top overrepresented biological pathways. None of the pathways were annotated after correction for multiple testing in similar analyses using KEGG method.

CONFIDENTIAL-For Peer Review Only

Genetic associations

In CKB GWAS, *cis*-pQTL variants were identified for 22 (67%) of the 33 proteins. In two-sample MR analyses involving CKB and AGEN that excluded CKB, 3 proteins (ENTR1, LPL, PON3) were significantly associated at FDR<0.05 with T2D (**Table 2**). Of these 3 proteins, the HRs were less extreme in MR than in observational analyses but were directionally concordant. Moreover, colocalization analyses provided strong support (PH4>0.6) for shared genetic variants of two proteins (LPL and PON3) with T2D. Independent two-sample MR analyses involving 22 *cis*-pQTLs identified in UKB GWAS for these 33 T2D-associated proteins replicated associations for ENTR1 (*P*=0.004) and LPL (*P*=0.01). For PON3, although *cis*-pQTL was identified in UKB GWAS the association (with same direction) was not significant (*P*=0.49).

PheWAS and drug target lookup

In PheWAS analyses of these 3 proteins, *cis*-pQTL for ENTR1 were associated with T2D and several T2D-related traits, including HbA1c, insulin and glucose (**Table 3**). Likewise, *cis*-pQTLs for LPL and PON3 were related to T2D, CVD outcomes (MI and CAD) and CVD risk factors (LDL, TG, ApoA). Within CKB, ENTR1, LPL and PON3 were each significantly associated with glucose in cross-sectional analyses. All three proteins were highly expressed in liver, pancreas and adipose tissues and additional analyses of single-gene KO mouse models identified associations with several lipidaemia-related phenotypes (LPL, ENTR1 and PON3), abnormal liver morphology (ENTR1) and oxidative stress (PON3). Analysis of Open Targets and other databases indicated evidence of drug development for one protein (LPL), including commercially available drug of Ibrolipim, a lipoprotein lipase activator that degrades circulating triglycerides in blood (**Table 2**). However, there were no reports of drug targets or development for ENTR1 and PON3.

CONFIDENTIAL-For Peer Review Only

Discussion

In this study of Chinese adults, we found 33 proteins were significantly associated with risk of incident T2D. The addition of these proteins to the conventional prediction models substantially improved risk prediction of T2D, with comparable performance in Chinese and European populations. Moreover, MR analyses based on *cis*-pQTLs identified in CKB GWAS provided strong support for the causal relevance of three proteins (ENTR1, LPL and PON3) for T2D, with replication of ENTR1 and LPL in Europeans using different *cis*-pQTLs. In colocalization analyses, there was strong evidence of shared causal genetic variants of T2D with two proteins (LPL and PON3). Furthermore, the PheWAS results confirmed the importance of these proteins for T2D or T2D-related traits. Among these three proteins, there was, however, no evidence of any drug development for ENTR1 and PON3.

Previous observational studies of proteomics and T2D have involved primarily European ancestry populations, used different study designs, and included varying number of proteins measured by different assay platforms.[13, 15, 16, 20] Although there were inconsistent findings, several proteins have been consistently associated with T2D, including IGFBP1, IGFBP2, GHR and SHBG.[12, 13] In the present study, observational analyses found that IGFBP1 and IGFBP2 were most strongly associated with T2D. The IGFBPs, which comprise some 15 proteins so far identified, importantly impact on systemic IGF signalling by modulating activity and decay of their binding partners. IGFBP-2, which is mainly released by the liver, directly supports glucose homeostasis by stimulating glucose uptake into adipocytes, and also inhibits adipo-genesis and enhances long-term insulin sensitivity.[28] Consistent with previous findings, GHR was positively associated with T2D risk in the present study. There is evidence that high levels of GHR can accelerate systemic insulin resistance,[29] which may partly explain the observed associations. Several

CONFIDENTIAL-For Peer Review Only

previous studies reported possible protective associations between T2D risk and SHBG, which is a hepatokine that binds to circulating steroid hormones (testosterone, oestradiol) and acts on macrophages and adipocytes to suppress inflammation and lipid accumulation.[30] Although the associations of SHBG with T2D became borderline significant after multiple testing correction (HR per SD higher: 0.69 [0.56-0.85]; 5% FDR *P*=0.057), the present results were consistent with previous study findings. Indeed, in CKB observational analyses, levels of adiponectin were inversely associated with risk of diabetes before multiple testing correction, consistent with previous literature.[31] In genetic analyses, however, we did not found the causal roles of these few known proteins in aetiology of T2D, possibly due to limited study power. The present study also found significant protective associations of pancreatic alpha-amylase (AMY2A, AMY2B) with T2D. Amylase is a digestive enzyme predominantly secreted by the pancreas and salivary glands and acts as a catalyst for carbohydrate hydrolysis, which is one of the viable targets to control T2D. Despite consistent observational finings, few previous studies were able to assess the causal relevance of these proteins in T2D. In our study, we confirmed the causal relevance of AMY2B in UKB using different *cis*-pQTL (OR per SD higher: 0.94 [0.91-0.97]; *P*=0.0003). On the other hand, we found three proteins (PON3, LPL and ENTR1) were significantly associated with T2D in both observational and genetic analyses.

PON3 is expressed primarily in the liver and there is good evidence from animal experiment and epidemiologic studies that PON3 can inhibit oxidative stress, suppress inflammation, improve insulin resistance and abnormal glucolipid metabolism, and protect against atherosclerosis.[32] As in the present study, the inverse associations between PON3 and incident T2D were also reported in two previous prospective studies in Sweden (1026 participants with 146 incident T2D cases) and Germany (1143 participants with 178

CONFIDENTIAL-For Peer Review Only

incident T2D cases),[16, 33] which persisted after adjusting for plasma glucose (marginal significant in the present study), implying a glucose independent association with T2D incidence.[16] However, no previous genetic studies supported its causal relevance for T2D incidence, therefore findings from the present study provides strong and novel support for PON3 as a potential target for improved prevention and treatment of T2D.

LPL (Lipoprotein lipase) is a rate-limiting enzyme that hydrolyzes circulating triglyceride-rich lipoproteins including very low-density lipoproteins and chylomicrons.[34] This enzyme is predominantly located in adipose tissue, muscle and cardiac tissue, and a reduction in LPL activity is associated with an increase in plasma levels of triglycerides, prompting evaluation of target druggability for treatment of dyslipidemia,[34] but its relevance to insulin resistance and glucose metabolism is less clear. Previous MR analyses suggested potential causal effects of LPL on insulin levels and the development of T2D.[33] Moreover, a pharmacological study involving 392,220 Europeans showed that triglyceride-lowering alleles in the LPL were associated with lower risk of T2D, independent of LDL-C lowering genetic mechanisms.[35] These findings provide genetic support for the development of agents that enhance LPL-mediated lipolysis for T2D prevention, which suggests, if further confirmed in other studies, potential opportunities for drug-repurposing for the treatment of T2D.

None of the previous observational studies have examined the associations of plasma levels of ENTR1 with risk of T2D. The *ENTR1* gene encodes the endosome associated trafficking regulator 1, which has a potential role in the transcriptional regulation of the solute carrier family 2 member 1 glucose 40 transporter protein (SLC2A1).[36] Importantly, SLC2A1 is responsible for approximately 30−40% of the glucose uptake in skeletal muscle, with the remainder transported through GLUT4.[36] This may partially explain the strong and apparently causal associations with T2D observed in the present study.

CONFIDENTIAL-For Peer Review Only

However, we could not exclude the possibility that these associations were caused by other pathways and further investigations of ENTR1 as a potential novel target for T2D are warranted.

In recent years, various proteomic-based prediction models for prevalent or incident diabetes have been developed, with varying number of proteins included (3 to 1468) and largely different degree of predictive performance.[9, 12, 14] More recently, UKB developed a ProteinScore for T2D based on 1468 OLINK proteins, which outperformed a polygenic risk score and HbA1C.[9] In the absence of HbA1c, we found that the addition of 33 or even 10 top proteins to conventional risk factors (including blood glucose) significantly improve the risk prediction of incident T2D in Chinese adults. Moreover, the same proteins identified in Chinese adults also yielded comparable results in European populations so could be considered for future clinical application in diverse populations.

The chief strengths of the present study include the large number of proteins assayed, independent replication of the main results internally and externally, use of ancestry-specific genetic instruments to assess causality, exclusion of CKB data from AGEN T2D GWAS summary statistics to minimize potential collider bias resulting from sample overlap, and multiple downstream analyses to assess possible mechanisms underlying these associations. Moreover, we also assessed the utility of proteomic-based risk prediction for T2D in diverse populations, independent and in combination with conventional risk factors. However, the present study also had several limitations. First, the study sample size was modest in CKB, limiting its power to detect more significant associations. Second, we were unable to independently replicate the observational findings in other East Asian populations due to the lack of available data. However, internal replication with prevalence T2D and plasma glucose levels, and external replication with incident T2D, glucose and HbA1c in UKB confirmed the validity of our

observational findings when applying similar multiple test correction. Third, the two-sample MR analyses only involved two thirds of proteins due to lack of overlapping *cis*-pQTLs in publically-available GWAS summary statistics. Fourth, there was no kidney function data collected in CKB among the study participants. In UKB, however, further adjustment for kidney function (blood creatinine) had minor effect on the total number of proteins associated with incident diabetes (1514 with adjustment vs 1541 without adjustment, at FDR<0.05). Future studies with a larger sample size and better genetic instruments, involving perhaps both *cis*- and *trans*-pQTLs, more advanced method of machine learning in risk prediction, and functional analyses are needed to further identify, replicate and clarify the associations of different proteins with T2D in different ancestry populations.

In summary, the present study identified 33 proteins that were significantly associated with T2D, with strong genetic support for the causal relevance of three proteins. With the exception of one protein (LPL), there was no evidence of any drug development for two proteins, particularly PON3, which is highly expressed in liver cells and is a promising drug target for improved prevention and treatment of T2D. Further biological validation using *in vitro* and *in vivo* experiments along with human studies are required to elucidate the underlying mechanisms. The present study highlighted the importance of proteomics in prospective studies of diverse populations to improve risk prediction, enhance understanding of disease aetiology and discover potential novel drug targets for treatment and prevention of T2D as well as other diseases.

CONFIDENTIAL-For Peer Review Only

Acknowledgments: The chief acknowledgment is to the participants, the project staff, and the China CDC and its regional offices for assisting with the fieldwork. We thank Judith Mackay in Hong Kong; Yu Wang, Gonghuan Yang, Zhengfu Qiang, Lin Feng, Maigeng Zhou, Wenhua Zhao, and Yan Zhang in China CDC; Lingzhi Kong, Xiucheng Yu, and Kun Li in the Chinese Ministry of Health; and Sarah Clark, Martin Radley, and Mike Hill in the CTSU, Oxford, for assisting with the planning, conduct and organization of the study.

Funding: The CKB baseline survey and the first re-survey were supported by the Kadoorie Charitable Foundation in Hong Kong. The long-term follow-up and subsequent resurveys have been supported by Wellcome grants to Oxford University (212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z) and grants from the National Natural Science Foundation of China (82192901, 82192904, 82192900) and from the National Key Research and Development Program of China (2016YFC0900500).The UK Medical Research Council (MC_UU_00017/1, MC_UU_12026/2, MC_U137686851), Cancer Research UK (C16077/A29186, C500/A16896) and the British Heart Foundation (CH/1996001/9454), provide core funding to the Clinical Trial Service Unit and Epidemiological Studies Unit at Oxford University for the project. The proteomic assays were supported by BHF (18/23/33512), Novo Nordisk and OLINK. DNA extraction and genotyping were supported by GlaxoSmithKline and the UK Medical Research Council (MC-PC-13049, MC-PC-14135). The trans-ethnic BMI-GS also used data from UKB (Application No.: 50474).

Conflict of interest/Competing interests: None of the authors have any conflicts of interest in relation to this report.

Ethics approval: The China Kadoorie Biobank (CKB) complies with all the required ethical standards for medical research on human subjects. Ethical approvals were granted and maintained by the relevant institutional ethical research committees in the UK and China.

Consent to participate/publication: All participants provided written informed consent.

Data Access Statement: The China Kadoorie Biobank (CKB) is a global resource for the investigation of lifestyle, environmental, blood biochemical and genetic factors as determinants of common diseases. The CKB study group is committed to making the cohort data available to the scientific community in China, the UK and worldwide to advance knowledge about the causes, prevention and treatment of disease. For detailed information on what data is currently available to open access users and how to apply for it, please visit: http://www.ckbiobank.org/site/Data+Access. A research proposal will be requested to ensure that any analysis is performed by *bona fide* researchers. Researchers who are interested in obtaining additional information or data that underlines this paper should contact ckbaccess@ndph.ox.ac.uk. For any data that are not currently available for open access, researchers may need to develop formal collaboration with study group.

Code availability: Custom code was used all statistical analyses in this report.

Author contributions

PY, HD, ZC contributed to the concept and design of the study. PY conducted statistical analyses and drafted the manuscript. PY, AI, AP, SS, NW, KL, IM, HF, CK, MM, YC, FB, BL, LY, JL, DA, DS, DS, PP, JL, CY, MH, DB, RW, LL, RC, HD and ZC were involved in the planning, acquisition and interpretation of data. IM, HF, YC, DA, DS, PP, and MH

provided administrative, technical, or material support. All authors provided critical revision of the manuscript for important intellectual content. PY and ZC are the guarantors of this work and take responsibility for the integrity and accuracy of the data analysis. ZC supervised the work.

Open Access Statement

This research was funded in whole, or in part, by the Wellcome Trust (212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z). For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Federation ID. IDF Diabetes Atlas, 10th edn. *Brussels, Belgium: International Diabetes Federation*. 2021;
2. O'Hearn M, Lara-Castor L, Cudhea F, et al. Incident type 2 diabetes attributable to suboptimal diet in 184 countries. *Nat Med*. 2023/04/01 2023;29(4):982-995. doi:10.1038/s41591-023-02278-8
3. Abbasi A, Peelen LM, Corpeleijn E, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ : British Medical Journal*. 2012;345:e5900. doi:10.1136/bmj.e5900
4. Edlitz Y, Segal E. Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards. *eLife*. 2022/06/22 2022;11:e71862. doi:10.7554/eLife.71862
5. Mahajan A, Spracklen CN, Zhang W, et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat Genet*. May 2022;54(5):560-572. doi:10.1038/s41588-022-01058-3
6. Spracklen CN, Horikoshi M, Kim YJ, et al. Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature*. Jun 2020;582(7811):240-245. doi:10.1038/s41586-020-2263-3
7. Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*. 2017/01/01 2017;16(1):19-34. doi:10.1038/nrd.2016.230
8. Suhre K, McCarthy MI, Schwenk JM. Genetics meets proteomics: perspectives for large population-based studies. *Nature Reviews Genetics*. 2021/01/01 2021;22(1):19-37. doi:10.1038/s41576-020-0268-2
9. Gadd DA, Hillary RF, Kuncheva Z, et al. Blood protein levels predict leading incident diseases and mortality in UK Biobank. *medRxiv*. 2023:2023.05.01.23288879. doi:10.1101/2023.05.01.23288879
10. Sun BB, Chiou J, Traylor M, et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature*. 2023/10/04 2023;doi:10.1038/s41586-023-06592-6
11. Dhindsa RS, Burren OS, Sun BB, et al. Rare variant associations with plasma protein levels in the UK Biobank. *Nature*. 2023/10/04 2023;doi:10.1038/s41586-023-06547-x
12. 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(4):733-741. doi:10.2337/dc22-1830
13. Elhadad MA, Jonasson C, Huth C, et al. Deciphering the Plasma Proteome of Type 2 Diabetes. *Diabetes*. Dec 2020;69(12):2766-2778. doi:10.2337/db20-0296
14. Huth C, von Toerne C, Schederecker F, et al. Protein markers and risk of type 2 diabetes and prediabetes: a targeted proteomics approach in the KORA F4/FF4 study. *Eur J Epidemiol*. 2019/04/01 2019;34(4):409-422. doi:10.1007/s10654-018-0475-8
15. Yuan S, Xu F, Li X, et al. Plasma proteins and onset of type 2 diabetes and diabetic complications: Proteome-wide Mendelian randomization and colocalization analyses. *Cell Reports Medicine*. 2023;4(9)doi:10.1016/j.xcrm.2023.101174
16. Molvin J, Pareek M, Jujic A, et al. Using a Targeted Proteomics Chip to Explore Pathophysiological Pathways for Incident Diabetes– The Malmö Preventive Project. *Scientific Reports*. 2019/01/22 2019;9(1):272. doi:10.1038/s41598-018-36512-y
17. Gold L, Walker JJ, Wilcox SK, Williams S. Advances in human proteomics at high scale with the SOMAscan proteomics platform. *N Biotechnol*. Jun 15 2012;29(5):543-9. doi:10.1016/j.nbt.2011.11.016

CONFIDENTIAL-For Peer Review Only

18. Assarsson E, Lundberg M, Holmquist G, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One*. 2014;9(4):e95192. doi:10.1371/journal.pone.0095192
19. Ferkingstad E, Sulem P, Atlason BA, et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet*. 2021/12/01 2021;53(12):1712-1721. doi:10.1038/s41588-021-00978-w
20. Chen ZZ, Gao Y, Keyes MJ, et al. Protein Markers of Diabetes Discovered in an African American Cohort. *Diabetes*. Apr 1 2023;72(4):532-543. doi:10.2337/db22-0710
21. Chen Z, Chen J, Collins R, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol*. Dec 2011;40(6):1652-66. doi:10.1093/ije/dyr120
22. Yao P, Iona A, Kartsonaki C, et al. Conventional and genetic associations of adiposity with 1463 proteins in relatively lean Chinese adults. *Eur J Epidemiol*. 2023/09/07 2023;doi:10.1007/s10654-023-01038-9
23. Xu S, Coleman RL, Wan Q, et al. Risk prediction models for incident type 2 diabetes in Chinese people with intermediate hyperglycemia: a systematic literature review and external validation study. *Cardiovasc Diabetol*. 2022/09/13 2022;21(1):182. doi:10.1186/s12933-022-01622-5
24. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*. 2021/08/28/ 2021;2(3):100141. doi:https://doi.org/10.1016/j.xinn.2021.100141
25. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*. Sep 15 2014;23(R1):R89-98. doi:10.1093/hmg/ddu328
26. Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. *Int J Epidemiol*. Jun 2016;45(3):908-15. doi:10.1093/ije/dyw127
27. Costanzo MC, von Grotthuss M, Massung J, et al. The Type 2 Diabetes Knowledge Portal: An open access genetic resource dedicated to type 2 diabetes and related traits. *Cell Metab*. 2023/04/04/ 2023;35(4):695-710.e6. doi:https://doi.org/10.1016/j.cmet.2023.03.001
28. Wittenbecher C, Ouni M, Kuxhaus O, et al. Insulin-Like Growth Factor Binding Protein 2 (IGFBP-2) and the Risk of Developing Type 2 Diabetes. *Diabetes*. 2018;68(1):188-197. doi:10.2337/db18-0620
29. Liu J, Nie C, Xue L, et al. Growth hormone receptor disrupts glucose homeostasis via promoting and stabilizing retinol binding protein 4. *Theranostics*. 2021;11(17):8283-8300. doi:10.7150/thno.61192
30. Bourebaba N, Ngo T, Śmieszek A, Bourebaba L, Marycz K. Sex hormone binding globulin as a potential drug candidate for liver-related metabolic disorders treatment. *Biomed Pharmacother*. 2022/09/01/ 2022;153:113261. doi:https://doi.org/10.1016/j.biopha.2022.113261
31. Li S, Shin HJ, Ding EL, van Dam RM. Adiponectin Levels and Risk of Type 2 Diabetes: A Systematic Review and Meta-analysis. *JAMA*. 2009;302(2):179-188. doi:10.1001/jama.2009.976
32. Liu Y, Zhu D, Dong G, Zeng Y, Jiang P, Xiao Y. Liver paraoxonase 3 expression and the effect of liraglutide treatment in a rat model of diabetes. *Adv Clin Exp Med*. Feb 2021;30(2):157-163. doi:10.17219/acem/130605
33. Luo H, Bauer A, Nano J, et al. Associations of plasma proteomics with type 2 diabetes and related traits: results from the longitudinal KORA S4/F4/FF4 Study. *Diabetologia*. 2023/09/01 2023;66(9):1655-1668. doi:10.1007/s00125-023-05943-2

34. Liu Y, Li H, Wang S, Yin W, Wang Z. Ibrolipim attenuates early-stage nephropathy in diet-induced diabetic minipigs: Focus on oxidative stress and fibrogenesis. *Biomed Pharmacother*. 2020/09/01/ 2020;129:110321. doi:https://doi.org/10.1016/j.biopha.2020.110321

35. Lotta LA, Stewart ID, Sharp SJ, et al. Association of Genetically Enhanced Lipoprotein Lipase–Mediated Lipolysis and Low-Density Lipoprotein Cholesterol–Lowering Alleles With Risk of Coronary Disease and Type 2 Diabetes. *JAMA Cardiology*. 2018;3(10):957-966. doi:10.1001/jamacardio.2018.2866

36. Farries G, Bryan K, McGivney CL, et al. Identification of expression quantitative trait loci in the skeletal muscle of Thoroughbreds reveals heritable variation in expression of genes relevant to cofactor metabolism. *bioRxiv*. 2019:713669. doi:10.1101/713669

Table 1. Predictive values of conventional risk factors, RPG, and 33 proteins for incident T2D, separately and combined

Model	AUC	NRI (95% CI)
Base model [a]	0.754 (0.710-0.798)	
Random plasma glucose (RPG)	0.646 (0.591-0.700)	
33 proteins	0.824 (0.786-0.862)	
Base model + RPG	0.774 (0.730-0.818)	14% (3-25%)[b]
Base model + 33 proteins	0.874 (0.844-0.904)	36% (22-50%)[b]
RPG + 33 proteins	0.829 (0.791-0.868)	43% (32-54%)[c]
Base model + RPG + 33 proteins	0.876 (0.846-0.906)	38% (24-52%)[d]
Base model + RPG + top 10 proteins[e]	0.844 (0.803-0.885)	28% (19-37%)[d]

[a] Predictors in the base model included age, sex, study area, fasting time, education, smoking, alcohol consumption, physical activity, family history of diabetes and BMI.
[b] Reference: Base model
[c] Reference: RPG
[d] Reference: Base model + RPG
[e] Ordered by *P* value

CONFIDENTIAL-For Peer Review Only

Table 2. Genetic effect estimates, colocalization, PheWAS results, and relevant drug targets of three proteins showing genetic effects on T2D

Protein	Full name	Two-sample MR			PH4	PheWAS associations	Drug (indication)
		cis-pQTL	OR (95% CI) per SD higher	*P-value*			
ENTR1	Endosome-associated-trafficking regulator 1	rs1051957	1.26 (1.18-1.34)	1.3E-11	0.01	T2D, HbA1c, glucose, insulin	–
LPL	Lipoprotein lipase	rs17411113	0.91 (0.85-0.97)	0.0098	0.87	T2D, MI, CAD, TG, VLDL, ApoA	Ibrolipim (lipid-lowering)
PON3	Serum paraoxonase/ lactonase 3	rs1053275	0.94 (0.89-0.98)	0.0047	0.65	ApoA, LDL	–

CONFIDENTIAL-For Peer Review Only

Figure legends

Figure 1. Overview of study design, analytic approaches and key findings

Figure 2. Associations of 1-SD higher levels of 2941 proteins with incident diabetes in observational analyses

Models were adjusted for age, age^2, sex, study area, fasting time, ambient temperature, plate ID, education, smoking, alcohol consumption, physical activity, family history of diabetes and BMI. Red, blue and grey dots denote significant positive, significant inverse and non-significant associations, respectively.

CONFIDENTIAL-For Peer Review Only

Figure legends

Figure 1. Overview of study design, analytic approaches and key findings

Figure 2. Associations of 1-SD higher levels of 2941 proteins with incident diabetes in observational analyses

Models were adjusted for age, age^2, sex, study area, fasting time, ambient temperature, plate ID, education, smoking, alcohol consumption, physical activity, family history of diabetes and BMI. Red, blue and grey dots denote significant positive, significant inverse and non-significant associations, respectively.

CONFIDENTIAL-For Peer Review Only

Proteomic analyses in diverse populations improved risk prediction and identified new drug targets for type 2 diabetes

Supplementary Material

Contents

CONFIDENTIAL-For Peer Review Only

Members of the China Kadoorie Biobank collaborative group

International Steering Committee: Junshi Chen, Zhengming Chen (PI), Robert Clarke, Rory Collins, Liming Li (PI), Chen Wang, Jun Lv, Richard Peto, Robin Walters.

International Co-ordinating Centre, Oxford: Daniel Avery, Derrick Bennett, Ruth Boxall, Sushila Burgess, Ka Hung Chan, Yiping Chen, Zhengming Chen, Johnathan Clarke, Robert Clarke, Huaidong Du, Ahmed Edris Mohamed, Hannah Fry, Simon Gilbert, Pek Kei Im, Andri Iona, Maria Kakkoura, Christiana Kartsonaki, Hubert Lam, Kuang Lin, James Liu, Mohsen Mazidi, Iona Millwood, Sam Morris, Qunhua Nie, Alfred Pozarickij, Paul Ryder, Saredo Said, Dan Schmidt, Becky Stevens, Iain Turnbull, Robin Walters, Baihan Wang, Lin Wang, Neil Wright, Ling Yang, Xiaoming Yang, Pang Yao.

National Co-ordinating Centre, Beijing: Xiao Han, Can Hou, Qingmei Xia, Chao Liu, Jun Lv, Pei Pei, Dianjanyi Sun, Canqing Yu,.

10 Regional Co-ordinating Centres:

Guangxi Provincial CDC: Naying Chen, Duo Liu, Zhenzhu Tang. Liuzhou CDC: Ningyu Chen, Qilian Jiang, Jian Lan, Mingqiang Li, Yun Liu, Fanwen Meng, Jinhuai Meng, Rong Pan, Yulu Qin, Ping Wang, Sisi Wang, Liuping Wei, Liyuan Zhou. **Gansu** Provincial CDC: Caixia Dong, Pengfei Ge, Xiaolan Ren. Maiji CDC: Zhongxiao Li, Enke Mao, Tao Wang, Hui Zhang, Xi Zhang. **Hainan** Provincial CDC: Jinyan Chen, Ximin Hu, Xiaohuan Wang. Meilan CDC: Zhendong Guo, Huimei Li, Yilei Li, Min Weng, Shukuan Wu. **Heilongjiang** Provincial CDC: Shichun Yan, Mingyuan Zou, Xue Zhou. Nangang CDC: Ziyan Guo, Quan Kang, Yanjie Li, Bo Yu, Qinai Xu. **Henan** Provincial CDC: Liang Chang, Lei Fan, Shixian Feng, Ding Zhang, Gang Zhou. Huixian CDC: Yulian Gao, Tianyou He, Pan He, Chen Hu, Huarong Sun, Xukui Zhang. **Hunan** Provincial CDC: Biyun Chen, Zhongxi Fu, Yuelong Huang, Huilin Liu, Qiaohua Xu, Li Yin. Liuyang CDC: Huajun Long, Xin Xu, Hao Zhang, Libo Zhang. **Jiangsu** Provincial CDC: Jian Su, Ran Tao, Ming Wu, Jie Yang, Jinyi Zhou, Yonglin Zhou. Suzhou CDC: Yihe Hu, Yujie Hua, Jianrong Jin, Fang Liu, Jingchao Liu, Yan Lu, Liangcai Ma, Aiyu Tang, Jun Zhang. **Qingdao** CDC: Liang Cheng, Ranran Du, Ruqin Gao, Feifei Li, Shanpeng Li, Yongmei Liu, Feng Ning, Zengchang Pang, Xiaohui Sun, Xiaocao Tian, Shaojie Wang, Yaoming Zhai, Hua Zhang, Licang CDC: Wei Hou, Silu Lv, Junzheng Wang. **Sichuan** Provincial CDC: Xiaofang Chen, Xianping Wu, Ningmei Zhang, Weiwei Zhou. Pengzhou CDC: Xiaofang Chen, Jianguo Li, Jiaqiu Liu, Guojin Luo, Qiang Sun, Xunfu Zhong. **Zhejiang** Provincial CDC: Weiwei Gong, Ruying Hu, Hao Wang,Meng Wang, Min Yu. Tongxiang CDC: Lingli Chen, Qijun Gu, Dongxia Pan, Chunmei Wang, Kaixu Xie, Xiaoyi Zhang.

CONFIDENTIAL-For Peer Review Only

eAppendix

Covariates selection in CKB

In our exploratory analyses, ambient temperature (temperature recorded on the date of blood collection) was associated with levels of ~40% proteins. Given the substantial variation of ambient temperature among 10 study areas (60℃ difference: from -25℃ to 35℃) in CKB, it is important to consider it as a covariate in the model.

The UK Biobank sample population

UK Biobank (UKB) is a population-based cohort of around 500,000 individuals aged between 40-69 years that were recruited between 2006 and 2010. Genome-wide genotyping, exome sequencing, electronic health record linkage, whole-body magnetic resonance imaging, blood and urine biomarkers and physical and anthropometric measurements are available. More information regarding the full measurements can be found at: https://biobank.ndph.ox.ac.uk/showcase/. The UK Biobank Pharma Proteomics Project (UKB-PPP) is a precompetitive consortium of 13 biopharmaceutical companies funding the generation of blood-based proteomic data from UKB volunteer samples (**eTable 3**).

Proteomics in the UK Biobank

The UKB-PPP sample includes 54,306 UKB participants and 2923 unique proteins measured using The Olink technology uses Proximity Extension Assay. A randomised subset of 46,673 individuals were selected from baseline UKB, with 6,385 individuals selected by the UKB-PPP consortium members and1,268 individuals included that participated in a COVID-19 study. The randomised samples have been shown to be highly representative of the wider UKB population and they are used in the current analyses, whereas the consortium selected individuals were enriched for 122 diseases. Details on sample selection for UKB-PPP, in addition to processing and quality control information for the Olink assay are provided in elsewhere (Nature 2023; *DOI: 10.1038/s41586-023-06592-6*).

Electronic health data linkage in the UK Biobank

Electronic health linkage to NHS records was used to collate incident diagnoses. Death information was sourced from the death registry data available through the UK Biobank.

CONFIDENTIAL-For Peer Review Only

Cancer outcomes were sourced from the cancer registry (ICD codes), whereas non-cancer diseases were sourced from first occurrence traits available in the UK Biobank. The first occurrence traits integrate GP (read2/3), ICD (9/10) with self-report and ICD codes present on the death registry to identify the earliest date of diagnoses. These data sources are linked to 3-digit ICD trait codes.

Statistical analyses

Plasma protein levels were standardized (i.e. values of each protein were divided by their SD) and analysed as continuous variables. In observational analysis, Cox regression models were used to estimate adjusted HRs (and 95% CI) for incident. All analyses were adjusted for age, age^2, sex, assess center, fasting time, education, smoking status, alcohol intake frequency, physical activity and BMI. All analyses were restricted to participants without prior diabetes at baseline. All statistical analyses were performed using R version 4.1.2. Benjamini-Hochberg FDR was used to correct for multiple testing.

CONFIDENTIAL-For Peer Review Only

eTable 1. Baseline characteristics of participants in proteomic subcohort and genotyped cohort in CKB

Characteristics [a]	Proteomic subcohort (n=1896)	Genotyped CKB (n=79,159)
Demographic and lifestyle factors		
Age, years, mean (SD)	51.3 (10.4)	51.7 (10.6)
Women, %	62.1	60.2
Urban residents, %	50.6	45.4
Education ≥high school, %	21.4	21.5
Ever regular smoker in men, %	74.9	74.0
Ever regular smoker in women, %	3.3	3.1
Ever regular drinker in men, %	40.1	36.9
Ever regular drinker in women, %	2.7	2.4
Physical activity, MET-h/day, mean (SD)	21.3 (12.3)	21.3 (13.9)
Medical history and health status, %		
Self-rated poor health	13.4	10.2
Chronic kidney disease	1.4	1.3
Cancer	0.6	0.5
Family history of diabetes [b]	16.1	15.8
Anthropometry and blood pressure, mean (SD)		
BMI, kg/m^2	23.9 (3.2)	23.6 (3.4)
WC, cm	80.3 (8.7)	80.1 (9.7)
SBP, mmHg	131 (20)	130 (21)
RPG, mmol/L	6.1 (2.5)	6.0 (2.3)
Fasting time, hours	5.1 (4.3)	5.2 (4.2)

[a] Adjusted for age, sex and study area, as appropriate.
[b] Family history of diabetes represents diabetes history of any first-degree relative
Abbreviations: SD=Standard deviation; BMI=Body mass index; SBP=Systolic blood pressure; MET=Metabolic equivalent of task; RPG=Random plasma glucose; WC=Waist circumference

CONFIDENTIAL-For Peer Review Only

eTable 2. Continuous NRI of prediction models of conventional risk factors, RPG, and 33 proteins for incident T2D, separately and combined

Model	NRI (95% CI)
Base model [a]	
Random plasma glucose (RPG)	
33 proteins	
Base model + RPG	44% (23-65%)[b]
Base model + 33 proteins	106% (88-124%)[b]
RPG + 33 proteins	98% (80-116%)[c]
Base model + RPG + 33 proteins	97% (79-115%)[d]
Base model + RPG + top 10 proteins [e]	84% (65-103%)[d]

a Predictors in the base model included age, sex, study area, fasting time, education, smoking, alcohol consumption, physical activity, family history of diabetes and BMI.
b Reference: Base model
c Reference: RPG
d Reference: Base model + RPG
e Ordered by P value

CONFIDENTIAL-For Peer Review Only

eTable 3. Baseline characteristics of participants in UKB

Characteristics	UKB-PPP Randomized baseline (n=46595)	UKB full cohort
Demographic and lifestyle factors		
Age, years, mean (SD)	56.7 (8.1)	56.5 (8.1)
Women, %	54.3	54.4
Townsend Deprivation Index	-1.23 (3.15)	-1.29 (3.09)
Smoking		
Never	54.3%	54.8%
Previous	34.8%	34.6%
Current	10.9%	10.6%
Ethnic background		
Asian/Asian British	2.0%	2.0%
Black/Black British	1.7%	1.6%
Chinese	0.3%	0.3%
Mixed	0.7%	0.6%
White	94.4%	94.6%
Other ethnic group	1.0%	0.9%
ABO blood group		
O	43.2%	43.3%
A	43.4%	43.5%
B	9.7%	9.6%
AB	3.7%	3.6%
Anthropometry, mean (SD)		
BMI, kg/m^2	27.4 (4.8)	27.4 (4.8)
Biochemistry		
Alanine aminotransferase (U/l)	23.41 (14.28)	23.55 (14.18)
Alkaline phosphatase (U/l)	83.74 (26.3)	83.67 (26.46)
Aspartate aminotransferase (U/l)	26.27 (11.09)	26.23 (10.66)
Cholesterol (mmol/l)	5.69 (1.15)	5.69 (1.14)
Creatinine (umol/l)	72.28 (18.91)	72.31 (18.55)
Gamma glutamyltransferase (U/l)	37.29 (41.11)	37.39 (42.09)
Glucose (mmol/l)	5.13 (1.25)	5.12 (1.24)
Glycated haemoglobin (HbA1c) (mmol/mol)	36.20 (6.86)	36.13 (6.78)
HDL-cholesterol (mmol/l)	1.45 (0.38)	1.45 (0.38)
Triglycerides (mmol/l)	1.75 (1.03)	1.75 (1.03)

CONFIDENTIAL-For Peer Review Only

eTable 4. Predictive values of conventional risk factors, RPG and proteins, separately and combined, for incident T2D in UKB

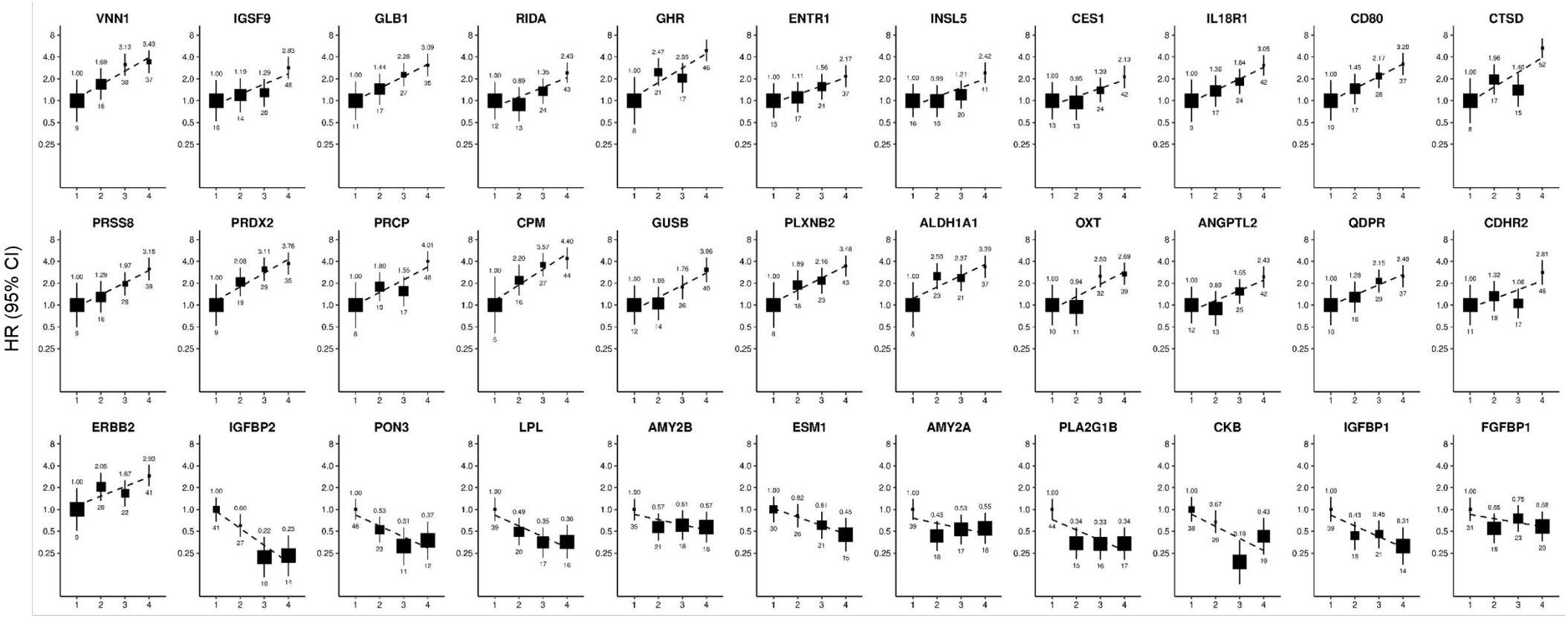
Model	AUC (95% CI)
Base model [a]	0.796 (0.787-0.805)
Random plasma glucose (RPG)	0.738 (0.726-0.749)
33 proteins	0.890 (0.840-0.857)
Base model + RPG	0.852 (0.843-0.861)
Base model + 33 proteins	0.879 (0.870-0.887)
RPG + 33 proteins	0.882 (0.873-0.891)
Base model + RPG + 33 proteins	0.893 (0.884-0.906)
HbA1c	0.892 (0.884-0.899)
Base model + RPG + HbA1c	0.917 (0.910-0.923)
RPG + HbA1c + 33 proteins	0.931 (0.924-0.938)
Base model + RPG + HbA1c + 33 proteins	0.937 (0.930-0.944)

[a] Adjusted for age, sex, study area, fasting time, education, smoking, alcohol consumption, physical activity, family history of diabetes and BMI.

CONFIDENTIAL-For Peer Review Only

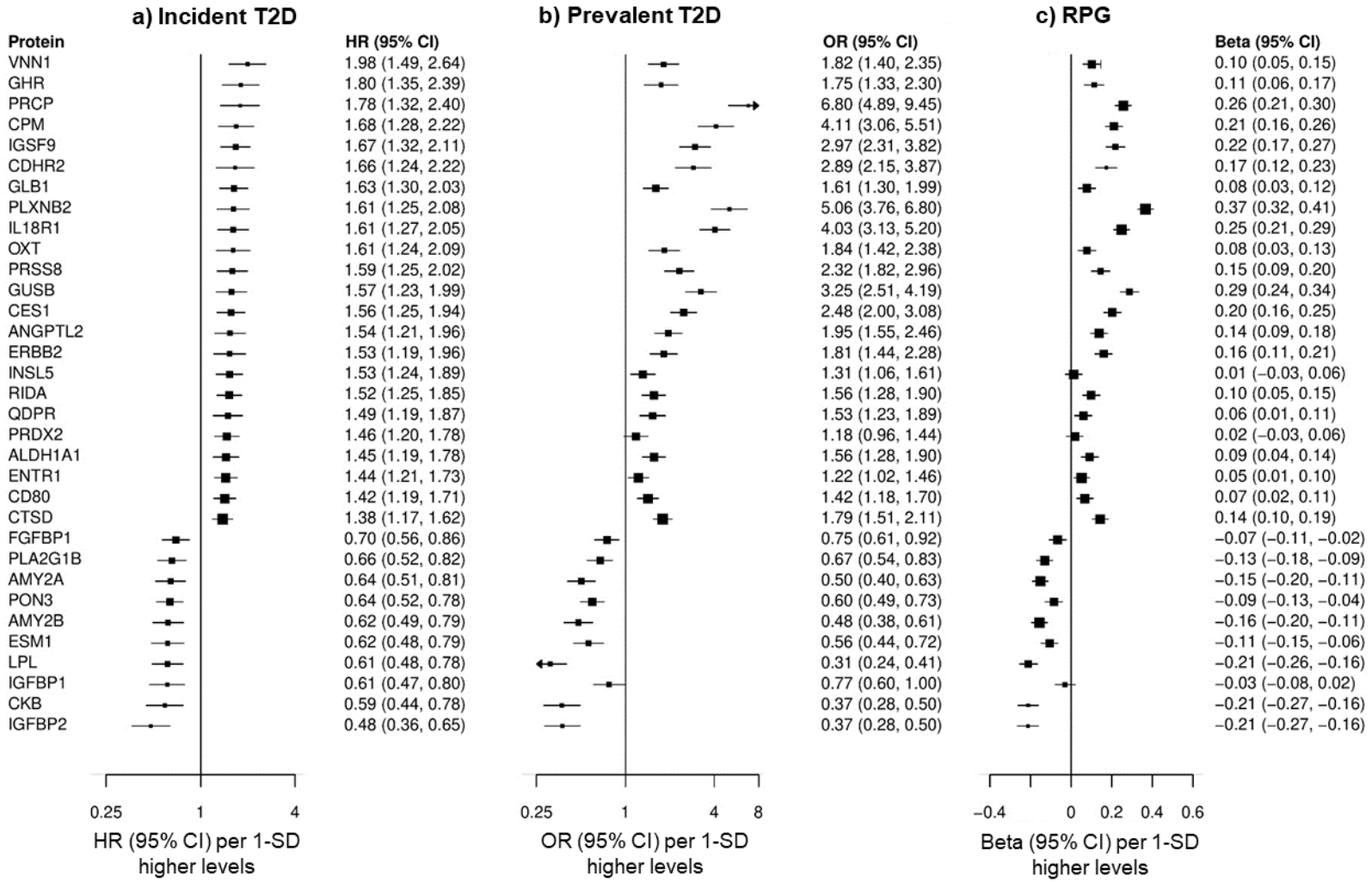
eFigure 1. Adjusted HRs for risk of diabetes by quartiles of 33 significant proteins

The sizes of the data markers are proportional to the inverse of the variance of the log HRs. The numbers above the 95% CI are point estimates for HRs, and the numbers below are numbers of diabetes cases for each category. The models were adjusted for sex, age, age², region, fasting time, ambient temperature, plate ID, education, smoking, alcohol consumption, physical activity, family history of diabetes and BMI.



CONFIDENTIAL-For Peer Review Only

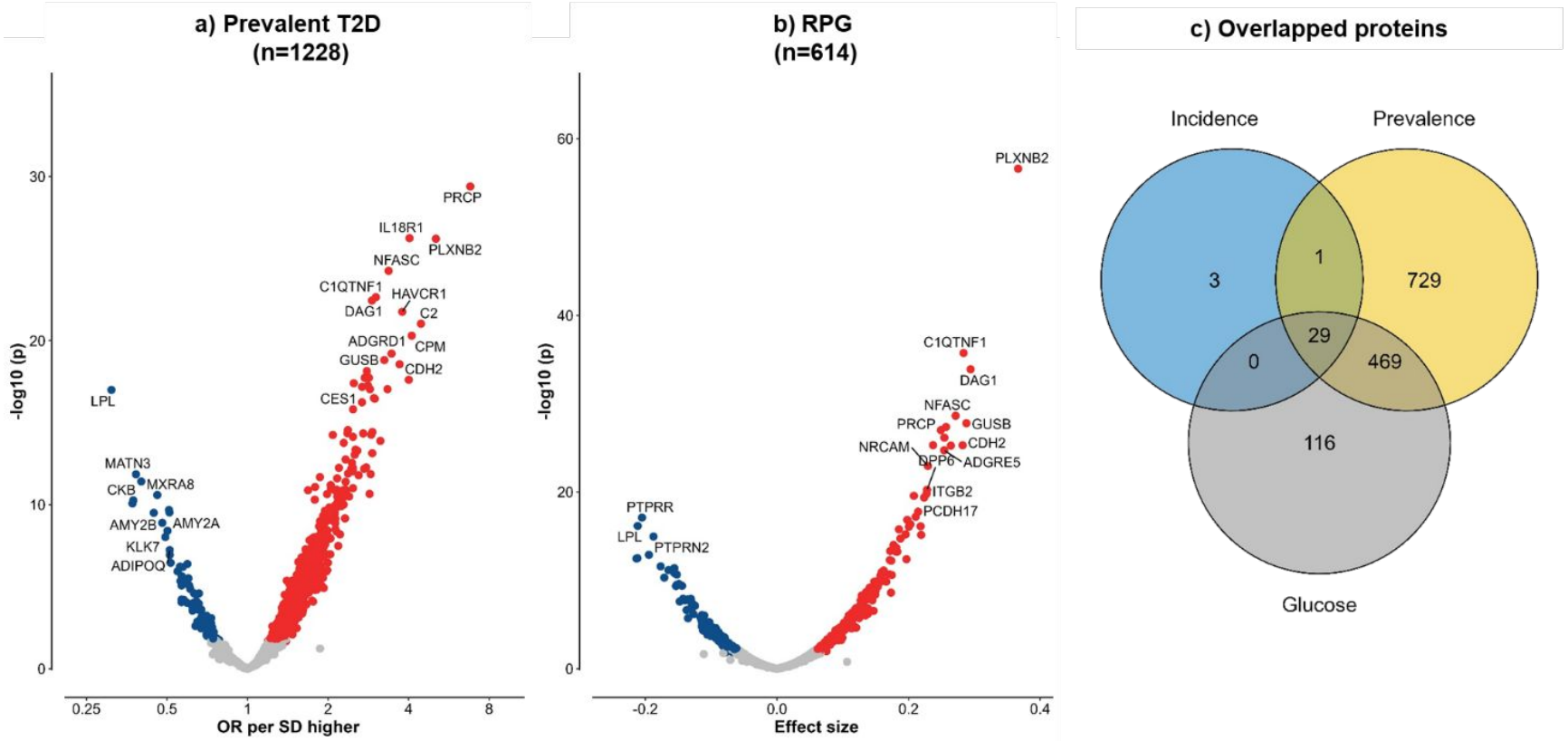
eFigure 2. Associations of 1-SD higher levels of 33 significant proteins with a) incident T2D, b) prevalent T2D and c) RPG, respectively



CONFIDENTIAL-For Peer Review Only

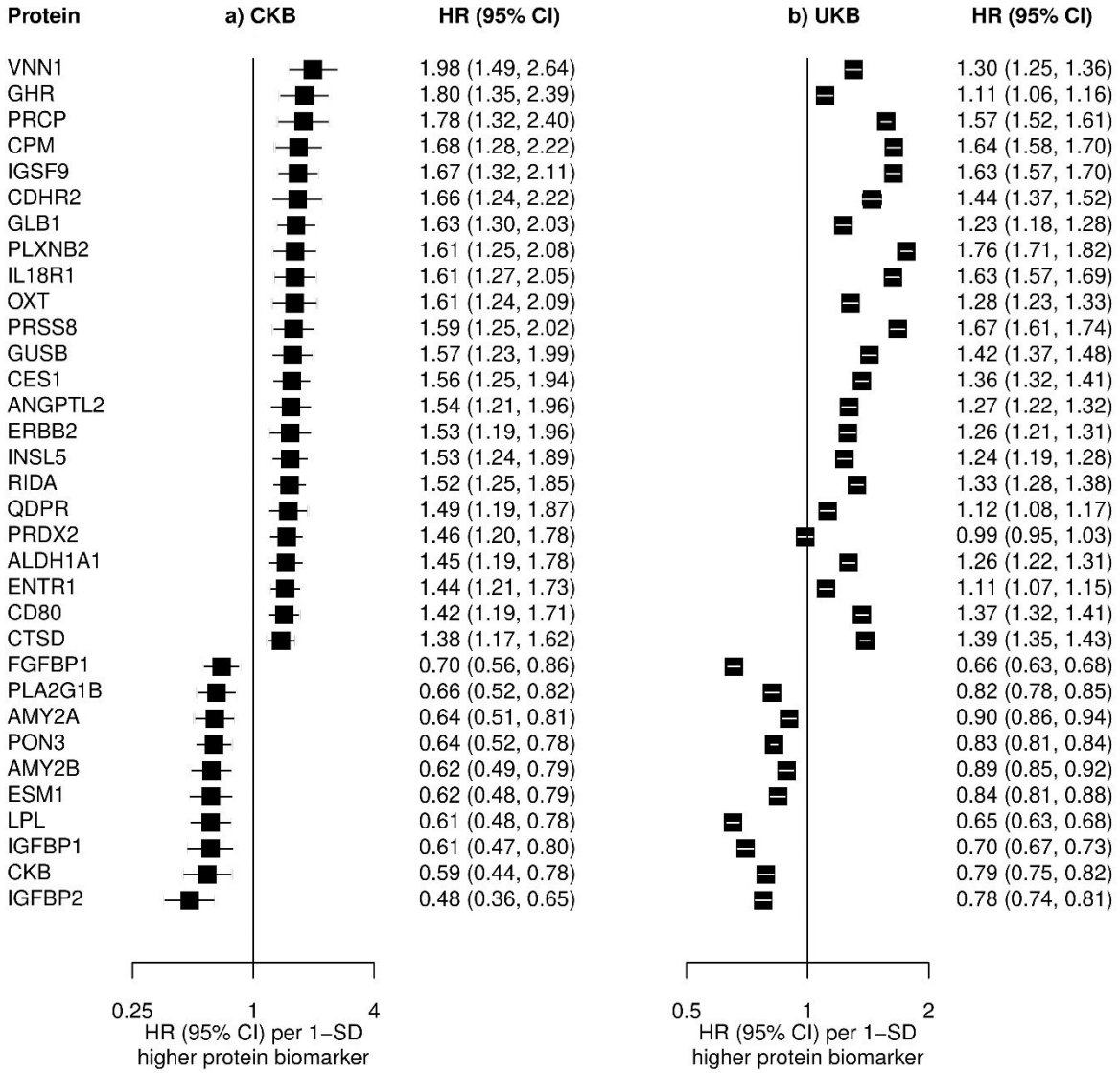
eFigure 3. Associations of 1-SD higher levels of 2941 proteins with a) prevalent diabetes, b) RPG levels and c) number of proteins overlapped with incident T2D in observational analyses

Models were adjusted for age, age square, sex, study area, fasting time, ambient temperature, plate ID, education, smoking, alcohol consumption, physical activity, family history of diabetes and BMI. Red, blue and grey dots denote positive significant, inverse significant and non-significant associations, respectively.



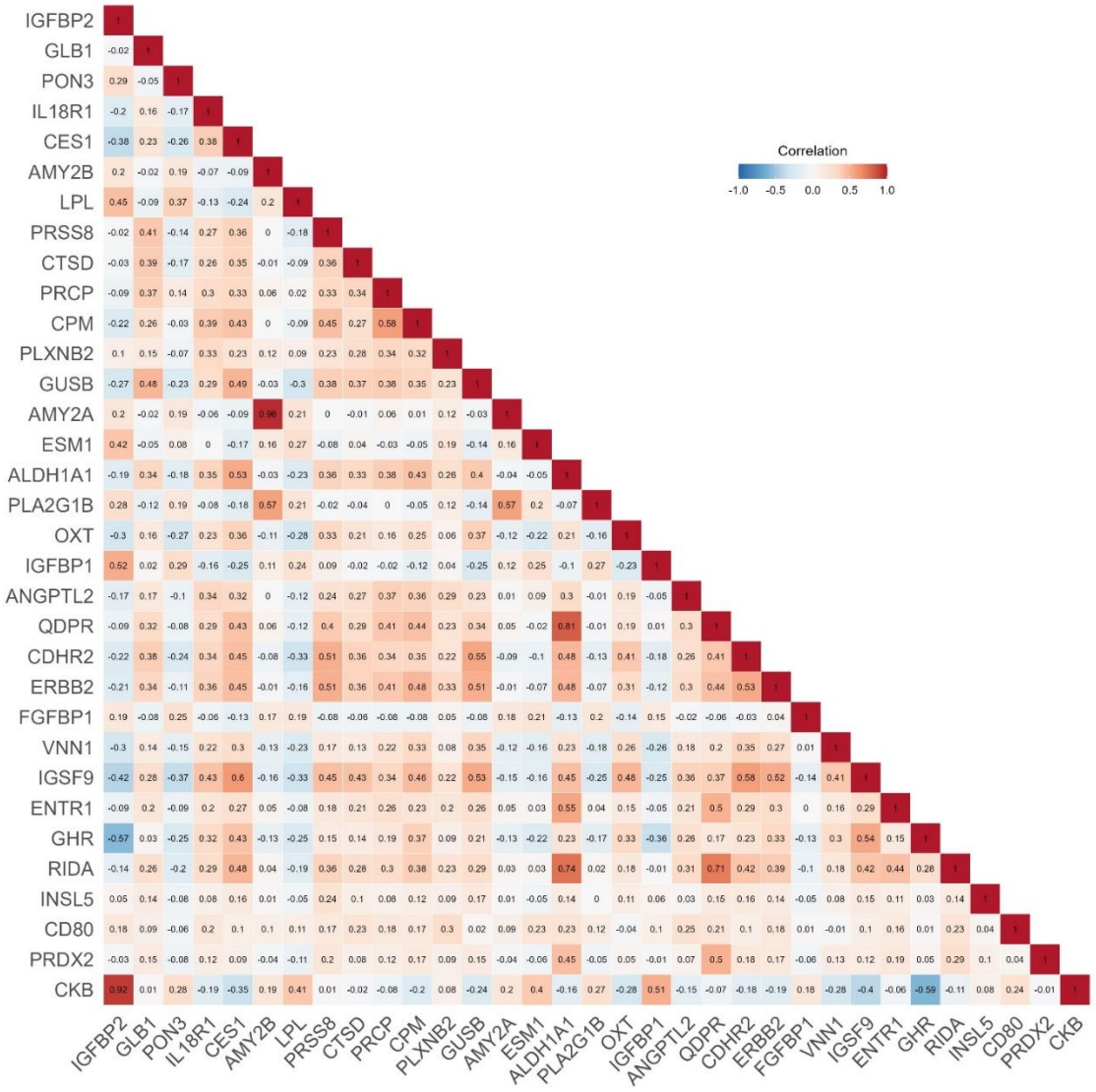
eFigure 4. Adjusted HRs for T2D associated with 1-SD higher levels of 33 significant proteins (OLINK batch 1) in a) CKB and b) UKB, respectively

Models were adjusted for age, age^2, sex, study area, fasting time, ambient temperature (CKB only), plate ID, education, smoking, alcohol consumption, physical activity, family history of diabetes and BMI. The comparison was conducted among 33 proteins from OLINK 2 batches.



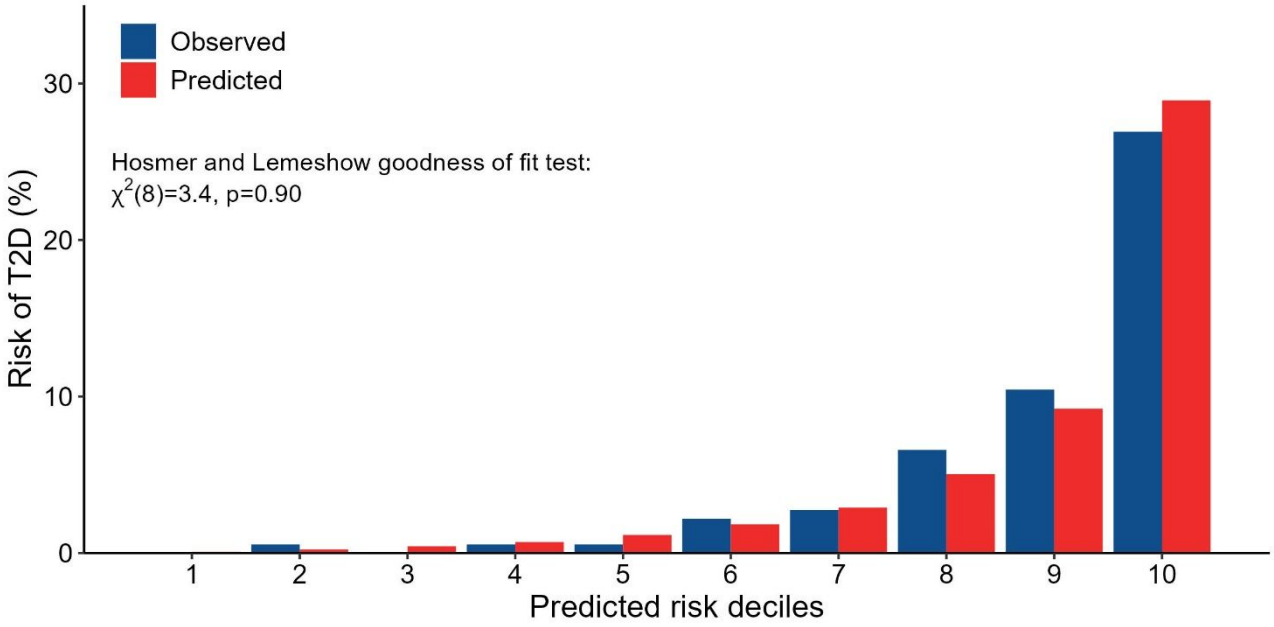
CONFIDENTIAL-For Peer Review Only

eFigure 5. Correlation matrix of 33 proteins significantly associated with risk of incident T2D



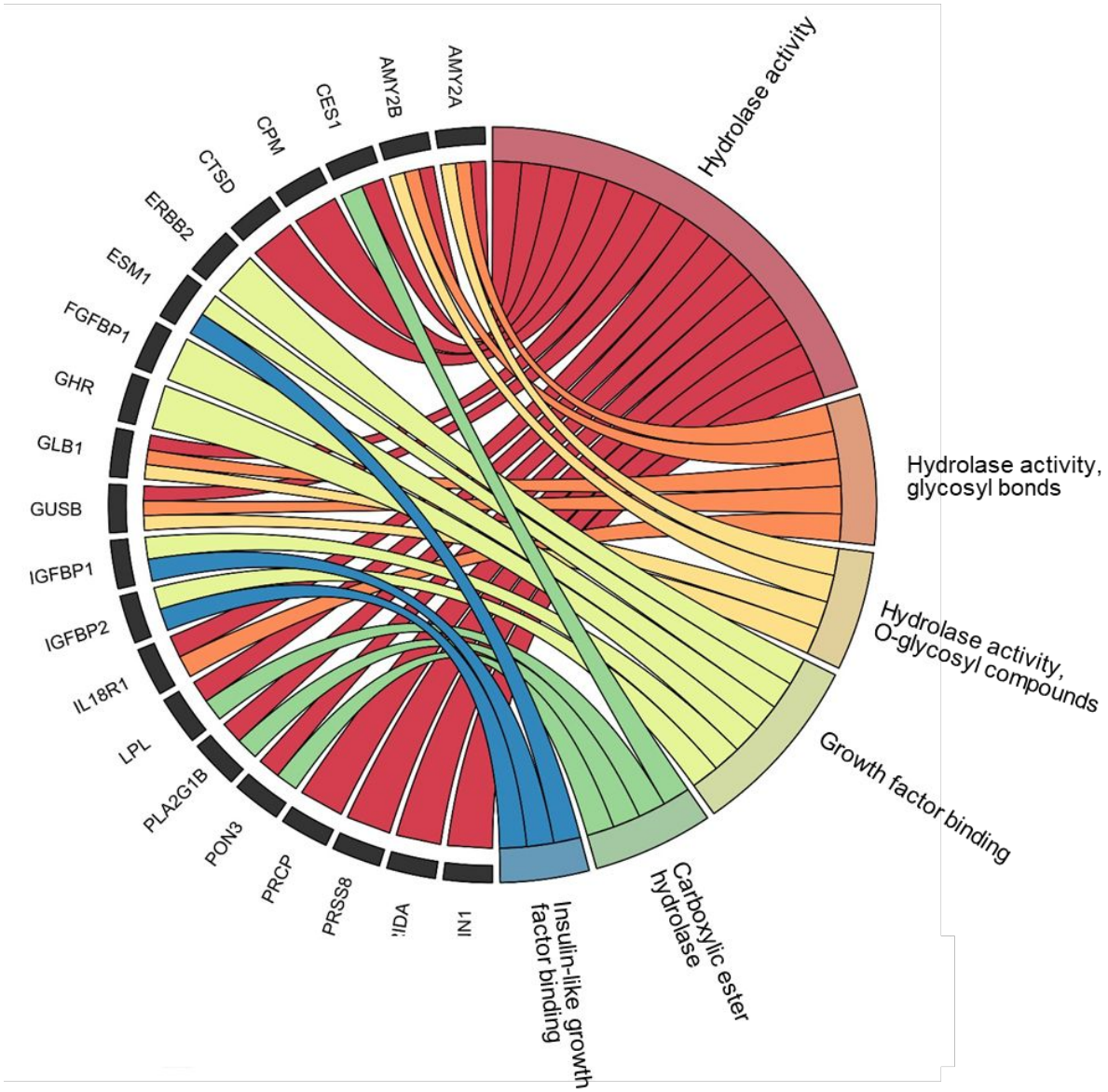
eFigure 6. Calibration plot of risk prediction models for T2D

The predictors in the model included age, sex, study area, fasting time, education, smoking, alcohol consumption, physical activity, family history of diabetes, BMI, RPG, and 33 proteins



CONFIDENTIAL-For Peer Review Only

eFigure 7. Chord diagrams of enriched GO molecular functions for 33 proteins significantly associated with risk of T2D



CONFIDENTIAL-For Peer Review Only

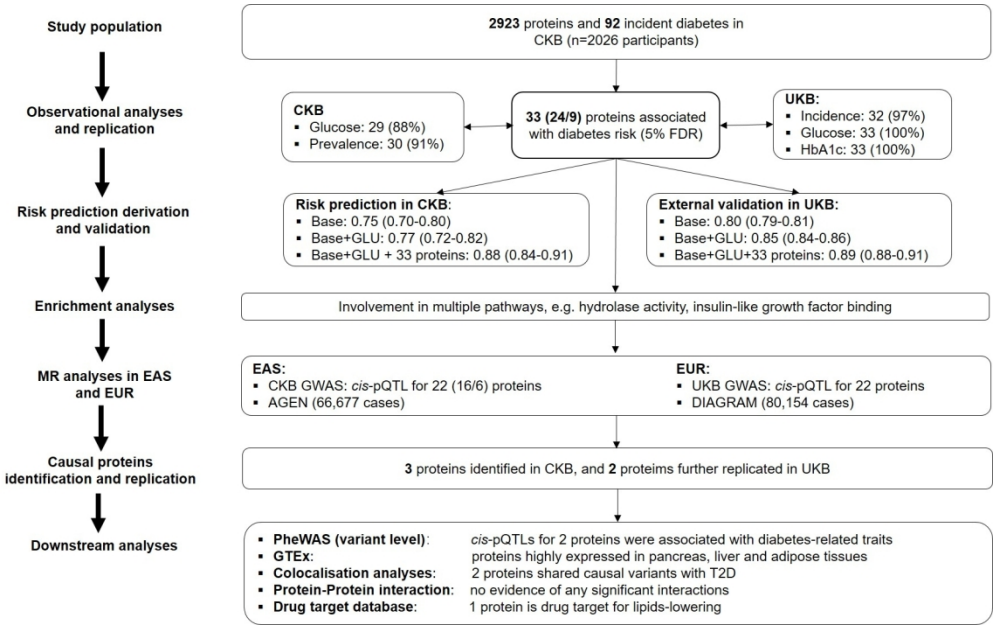


Figure 1

301x187mm (150 x 150 DPI)

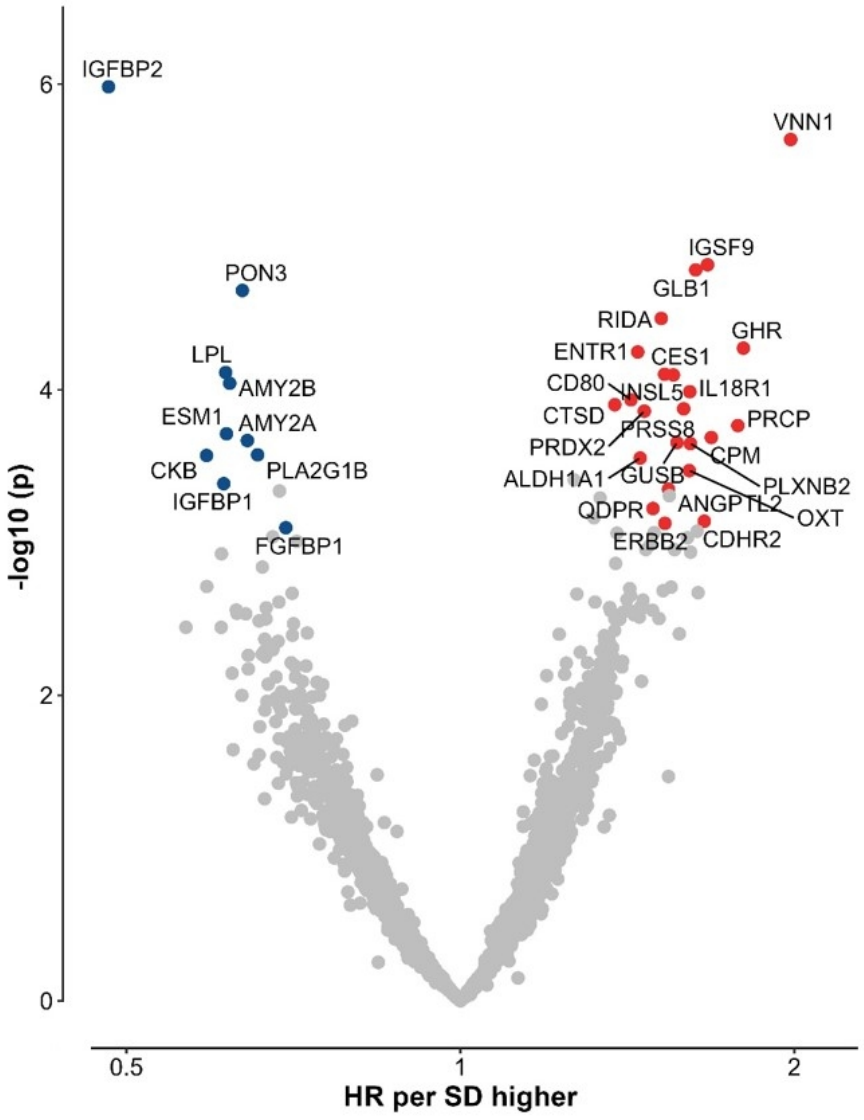
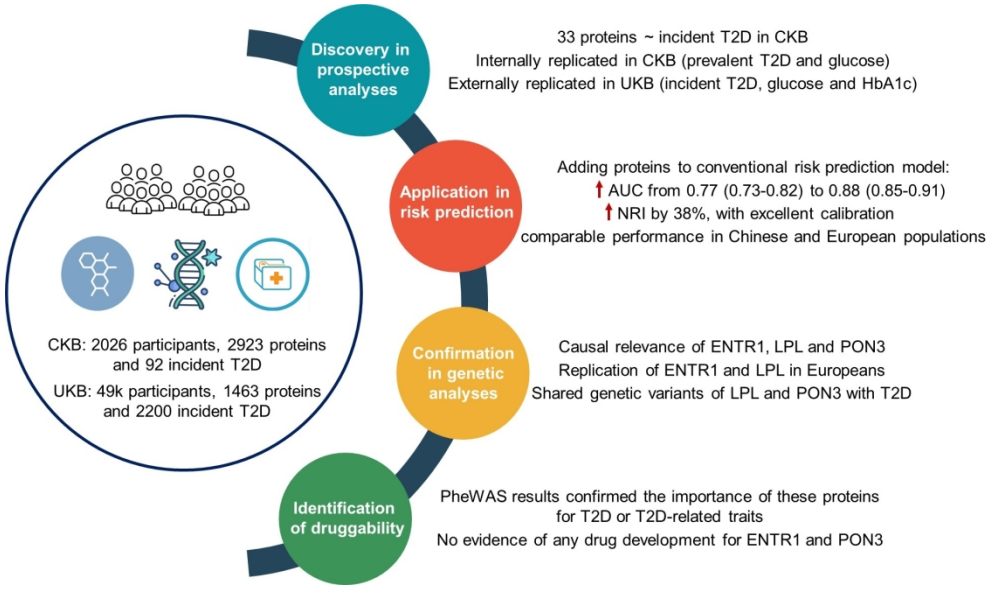


Figure 2

110x151mm (150 x 150 DPI)



338x196mm (150 x 150 DPI)