nature metabolism

LETTERS

https://doi.org/10.1038/s42255-021-00478-5

Check for updates

Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases

Scott C. Ritchie[1,2,3,4] ✉, Samuel A. Lambert[1,2,3,5], Matthew Arnold[3], Shu Mei Teo[1,2], Sol Lim[1,2,3], Petar Scepanovic[1,2,3], Jonathan Marten[1,3], Sohail Zahid[6,7], Mark Chaffin[6], Yingying Liu[8,9], Gad Abraham[1,2,10], Willem H. Ouwehand[4,11,12,13,14], David J. Roberts[12,14,15], Nicholas A. Watkins[12], Brian G. Drew[2,9,16], Anna C. Calkin[2,8,16], Emanuele Di Angelantonio[3,4,5,14,17], Nicole Soranzo[4,13,14], Stephen Burgess[3,18], Michael Chapman[3,5,13], Sekar Kathiresan[19], Amit V. Khera[6,7,20,21], John Danesh[3,4,5,13,14], Adam S. Butterworth[3,4,5,14] and Michael Inouye[1,2,3,4,5,10,22] ✉

Cardiometabolic diseases are frequently polygenic in architecture, comprising a large number of risk alleles with small effects spread across the genome[1–3]. Polygenic scores (PGS) aggregate these into a metric representing an individual's genetic predisposition to disease. PGS have shown promise for early risk prediction[4–7] and there is an open question as to whether PGS can also be used to understand disease biology[8]. Here, we demonstrate that cardiometabolic disease PGS can be used to elucidate the proteins underlying disease pathogenesis. In 3,087 healthy individuals, we found that PGS for coronary artery disease, type 2 diabetes, chronic kidney disease and ischaemic stroke are associated with the levels of 49 plasma proteins. Associations were polygenic in architecture, largely independent of *cis* and *trans* protein quantitative trait loci and present for proteins without quantitative trait loci. Over a follow-up of 7.7 years, 28 of these proteins associated with future myocardial infarction or type 2 diabetes events, 16 of which were mediators between polygenic risk and incident disease. Twelve of these were druggable targets with therapeutic potential. Our results demonstrate the potential for PGS to uncover causal disease biology and targets with therapeutic potential, including those that may be missed by approaches utilizing information at a single locus.

Human genetic studies have identified numerous proteins involved in coronary artery disease (CAD), type 2 diabetes (T2D) and other cardiometabolic diseases through a combination of genome-wide association studies (GWAS), fine-mapping, colocalization and Mendelian randomization by overlaying information at strong cardiometabolic disease loci[9–12]. However, cardiometabolic diseases are polygenic in architecture since they depend on many thousands of variants across the genome, nearly all exerting small lifelong effects[13–17]. These variants are spread across many different pathways and likely exert their effects through multiple levels of regulation, including gene expression, proteins and their interactions, cell morphology and higher-order physiological processes[18–20]. PGS aggregate these small effects into a single number for each individual that captures a fraction of their disease susceptibility. The use of PGS for risk stratification has shown potential clinical utility for disease prevention[21], yet the specific molecular consequences that precede disease risk for these polygenic effects are unknown. For example, proteins that are pathway-level hubs through which polygenic effects converge could be particularly promising targets for pharmaceutical intervention[22,23].

In this study, we demonstrated how PGS can be used to identify proteins with causal roles in disease aetiology. The INTERVAL cohort consists of approximately 50,000 adult blood donors in England[24,25], of which 3,087 participants have linked electronic hospital records, imputed genome-wide genotypes and quantitative levels of 3,438 plasma proteins[26] (Supplementary Data 1 and 2). A schematic of the study is given in Extended Data Fig. 1. The characteristics of the participants are given in Extended Data Fig. 2; participants with a history of any cardiometabolic disease were excluded (Supplementary Table 1), reducing the potential for reverse causality in downstream analysis.

To quantify each participant's relative polygenic risk of atrial fibrillation (AF), CAD, chronic kidney disease (CKD), ischaemic

[1]Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [2]Cambridge Baker Systems Genomics Initiative, Baker Heart & Diabetes Institute, Melbourne, Victoria, Australia. [3]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [4]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK. [5]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK. [6]Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]Department of Medicine, Harvard Medical School, Boston, MA, USA. [8]Lipid Metabolism & Cardiometabolic Disease Laboratory, Baker Heart & Diabetes Institute, Melbourne, Victoria, Australia. [9]Molecular Metabolism & Ageing Laboratory, Baker Heart & Diabetes Institute, Melbourne, Victoria, Australia. [10]Department of Clinical Pathology, University of Melbourne, Parkville, Victoria, Australia. [11]Department of Haematology, University of Cambridge, Cambridge, UK. [12]National Health Service Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK. [13]Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK. [14]National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK. [15]National Institute for Health Research Oxford Biomedical Research Centre, University of Oxford and John Radcliffe Hospital, Oxford, UK. [16]Central Clinical School, Monash University, Melbourne, Victoria, Australia. [17]Centre for Health Data Science, Human Technopole, Milan, Italy. [18]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. [19]Verve Therapeutics, Cambridge, MA, USA. [20]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [21]Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA. [22]The Alan Turing Institute, London, UK. ✉e-mail: sr827@medschl.cam.ac.uk; mi336@medschl.cam.ac.uk

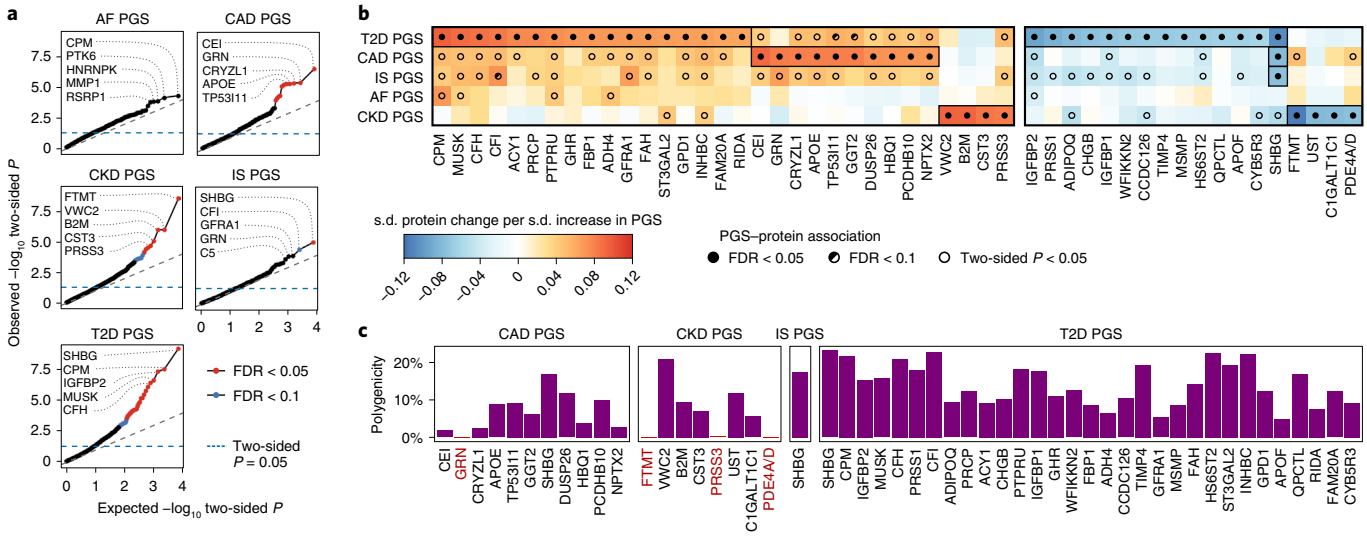


Fig. 1 | Proteins associated with polygenic risk for cardiometabolic disease. a, Quantile–quantile plots of two-sided *P* values from linear regression testing associations between PGS and protein levels in $n = 3{,}087$ INTERVAL participants across all 3,438 tested proteins. Each plot compares the distribution of observed two-sided *P* values (*y* axes) to the distribution of expected two-sided *P* values under the null hypothesis for 3,438 tests (*x* axes) on a $-\log_{10}$ scale. Associations were fitted using linear regression adjusting for age, sex, ten genotype principal components, sample measurement batch and time between blood draw and sample processing. Full summary statistics including exact *P* values are provided in Supplementary Data 3a. The top five proteins by *P* value are labelled. **b**, Heatmaps showing the 49 proteins whose levels were significantly associated with at least one PGS after Benjamini–Hochberg FDR multiple-testing correction (FDR < 0.05) of the two-sided *P* values (statistical tests are as described in **a**). Each heatmap cell shows the s.d. change in protein levels per s.d. increase in PGS. Point estimates for the 49 FDR-significant proteins are detailed in Extended Data Fig. 3. Details about each protein are provided in Extended Data Fig. 4. **c**, Barplots showing the proportion of the genome (%) required to explain each PGS–protein association in $n = 3{,}087$ INTERVAL participants (polygenicity). Proteins are ordered from left to right by strength of PGS–protein association. Highlighted in red are PGS–protein associations that were explained by singular variants regulating protein levels, pQTLs, rather than polygenic. Percentages are detailed in Extended Data Fig. 3.

stroke (IS) and T2D, we applied externally derived genome-wide PGS consisting of 1.8–3.2 million variants. Using PGS, we identified 49 proteins whose levels differed with respect to polygenic risk at a false discovery rate (FDR) of 5% (Fig. 1a,b, Extended Data Figs. 3 and 4 and Supplementary Tables 2 and 3): 31 proteins for the T2D PGS; 11 proteins for the CAD PGS; 8 proteins for the CKD PGS; and 1 protein for the IS PGS. PGS–protein associations included proteins previously associated with cardiometabolic disease, such as cystatin-C (CST3) and beta2-macroglobulin (B2M), which are biomarkers for CKD[27], and fructose-1,6-bisphosphatase 1 (FBP1), which plays a key role in glucose regulation and is a target of T2D drugs[28]. Associated proteins belonged to multiple non-overlapping pathways (Supplementary Information) and many are relatively understudied in the context of their respective diseases (Extended Data Fig. 5) thereby warranting future study.

PGS–protein associations were robust to technical, physiological and environmental confounding. We observed directional consistency and strong correlation of effect sizes when utilizing an orthogonal proteomics technology in independent samples (Extended Data Fig. 6a–c and Supplementary Information). Protein levels and PGS–protein associations were also temporally stable over 2 years of follow-up (Extended Data Fig. 6c,d and Supplementary Information). PGS–protein associations were also robust to circadian and seasonal effects, inclusion of participants with any prevalent cardiometabolic disease and body mass index (BMI), with the exception of six T2D PGS–protein associations that were partially mediated by BMI (Extended Data Fig. 6f,g).

Most PGS–protein associations were not explained by protein quantitative trait loci (pQTLs) (Supplementary Table 4); instead, they were highly polygenic (Fig. 1c). Each protein required a median 12% of the genome to explain its association with a PGS. Only 4 associations could be explained by pQTLs and contributing loci were spread across the genome for the remaining 46 associations (Extended Data Fig. 7).

Three possible scenarios could explain a PGS–protein association[29]: (1) the protein plays a causal role in disease; (2) protein levels are changing in response to disease processes but are not themselves causal (reverse causality); and (3) protein levels are correlated with some other causal factor (confounding) (Fig. 2a). Utilizing a median of 7.7 years of follow-up in nationwide electronic hospital records, we examined whether levels of PGS-associated proteins were associated with risk of onset of the respective cardiometabolic disease, then performed mediation analysis[30–32] to identify the proteins that mediate the PGS–disease associations and thereby play causal roles in disease pathogenesis.

During follow-up of the participants with PGS and plasma proteomics, there were 27 incident T2D events and 15 incident CAD events, enabling us to evaluate the CAD and T2D PGS and their corresponding 42 associated proteins. Ten of 31 (32%) T2D PGS-associated proteins were significantly associated ($P < 0.0012$, Bonferroni correction for the 42 tested proteins) and a further 15 proteins were nominally significantly associated ($P < 0.05$) with increased risk of T2D (Fig. 2b and Extended Data Fig. 3). For the CAD PGS, no proteins were Bonferroni significant and 3 of 11 (27%) proteins were nominally significant. Notably, there was clear directional consistency between the effects of PGS on protein levels and hazard ratios (HRs) for protein levels on incident disease risk (Fig. 2c). Using mediation analysis, we found one protein, insulin-like growth factor-binding protein 2 (IGFBP2), that was a significant mediator ($P < 0.0012$) of polygenic T2D risk (Fig. 2d and Extended Data Fig. 3), indicating a causal role in disease pathogenesis. A further 1 and 14 proteins were nominally significant mediators of polygenic CAD and T2D risk, respectively. Protein–disease

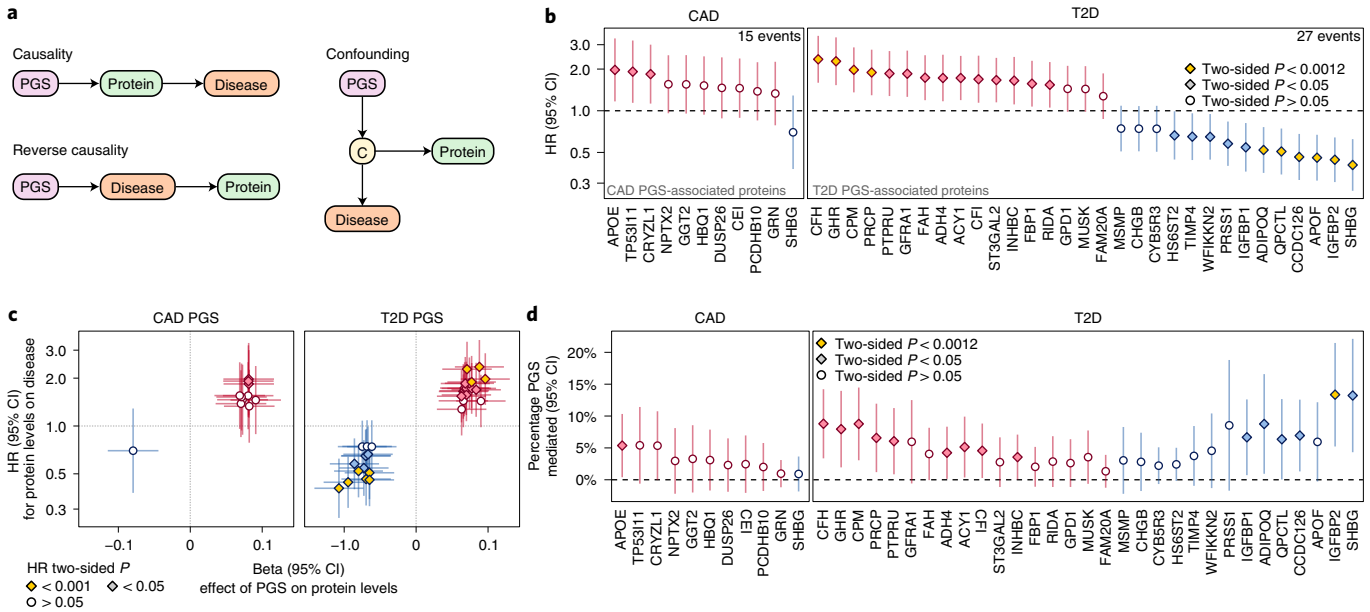


Fig. 2 | PGS-associated proteins influence 7.7-year risk of CAD and T2D. a, Possible models of causality for PGS–protein–disease associations. C, causal disease factor upstream of the protein that induces a correlation between protein levels and disease. **b**, Association of PGS-associated proteins with 7.7-year risk of hospitalization with CAD or T2D in $n = 3{,}087$ INTERVAL participants in Cox proportional hazards models adjusting for age, sex, sample measurement batch and time between blood draw and sample processing. The data shown correspond to the HR for CAD or T2D conferred per s.d. increase in protein levels (points) and its 95% CI (vertical bars). $P < 0.0012$ indicates that the association was significant after Bonferroni correction for the 42 tests. **c**, Comparison of associations between protein levels and the CAD PGS or T2D PGS from Fig. 1b (*x* axes) to HRs for protein levels for 7.7-year risk of hospitalization with CAD or T2D from Fig. 2b (*y* axes). Beta estimates (points; *x* axes) correspond to s.d. change in protein levels per s.d. increase in CAD PGS or T2D PGS in the linear regression described in Fig. 1b. HRs (*y* axes) are as described in **b**. Linear regression and Cox proportional hazards models were fitted in the same $n = 3{,}087$ samples. The point shape and colour correspond to the *P* value in **b**. **d**, Percentage of PGS–disease association mediated by each protein in causal mediation analysis in $n = 3{,}087$ INTERVAL participants adjusting for age, sex, ten genotype principal components, sample measurement batch and time between blood draws. The data shown are the percentage of association between the PGS and hospitalization with CAD or T2D after 7.7 years of follow-up in Cox proportional hazard models mediated by each respective protein (points) and the 95% CI of this percentage (vertical bars). Proteins are ordered from left to right by their HR in Fig. 1b and coloured red where protein levels increased with PGS or blue where protein levels decreased with PGS in Fig. 1b. $P < 0.0012$ indicates that mediation was significant after Bonferroni correction for the 42 tests. Full summary statistics including exact two-sided *P* values for **b**–**d** are detailed in Extended Data Fig. 3.

associations and causal mediation for IGFBP1 and IGFBP2 on polygenic T2D risk were replicated in independent samples where four PGS-associated proteins (IGFBP2, IGFBP1, progranulin (GRN) and tissue inhibitor of metalloproteinases 4 (TIMP4)) were measured with orthogonal proteomics technology (Supplementary Information and Supplementary Table 5).

Since polygenic disease risk is estimated from population-level data, it is unlikely that any single protein explains polygenic risk. In this study, we found that IGFBP2 explained 13.4% of the association between T2D PGS and incident T2D (Extended Data Fig. 3). Across all nominally significant mediators, proteins explained a median of 6.6% of PGS–disease associations (Extended Data Fig. 3), with the 1 CAD PGS mediator (apolipoprotein E (APOE)) explaining 5.4% of CAD polygenic risk–incident CAD association and the 15 T2D PGS mediators explaining 27% of the T2D polygenic risk–incident T2D association.

A complementary approach for causal inference, Mendelian randomization[33], also supported causal effects on T2D for one protein, sex hormone-binding globulin (SHBG) (Extended Data Fig. 9 and Supplementary Tables 6 and 7), which is consistent with our mediation analysis and previous Mendelian randomization analysis of SHBG on T2D[34]. Notably, only 11 (22%) of the proteins associated with PGS could be tested with Mendelian randomization due to a lack of *cis*-pQTLs as genetic instruments, highlighting the complementarity of our PGS–protein association and mediation approach for identifying causal proteins.

Our findings are also consistent with a previous observational study of plasma proteins and T2D risk in the Age, Gene/Environment Susceptibility-Reykjavik Study (AGES-Reykjavik), a cohort of 5,438 older Icelanders with 654 prevalent T2D cases and 112 incident T2D cases in 2,940 participants with 5 years of follow-up and free of T2D at baseline[35]. Of the 31 proteins associated with the polygenic risk of T2D in our healthy, pre-symptomatic cohort, 23 were associated with prevalent T2D and 16 were associated with incident T2D in the AGES-Reykjavik Study (Extended Data Fig. 10a and Supplementary Table 8). Notably, HRs for incident T2D in our INTERVAL analyses were directionally consistent and of similar magnitude to the odds ratios for incident T2D in the AGES-Reykjavik Study (Extended Data Fig. 10b), with a significant overlap between the significant proteins from our causal mediation analysis in INTERVAL and those previously associated with incident T2D (Extended Data Fig. 10a).

Finally, we examined the druggability of proteins mediating polygenic disease risk using the druggable genome[22]. We found that 12 of the 16 proteins mediating the polygenic disease risk were also druggable targets (Table 1). Nine of these were targets of, or interacted with, 76 compounds in the DrugBank database[36] (Supplementary Table 9). These results suggest therapeutic potential for these proteins as modulators of risk for T2D or CAD and indicate high priority targets for further investigation.

Polygenic disease scores are explicitly constructed to maximize risk prediction, typically without consideration of the underlying

Table 1 | Druggable proteins that were nominally significant mediators of polygenic risk

Protein	PGS/ disease	Evidence tier[a]	Small-molecule target[b]	Biological target[c]	ADME[d]	DrugBank compounds[e]	Summary of therapeutic uses for licensed drugs
ADH4	T2D	Tier 1	Y	N	Y	3	Female reproductive disorders, infection control
GHR	T2D	Tier 1	N	Y	N	3	Acromegaly, dwarfism, idiopathic short stature, human immunodeficiency virus weight loss
PRCP	T2D	Tier 1	Y	N	N	0	
SHBG	T2D	Tier 1	Y	Y	N	68	Fertility and reproductive treatments, cancers, mental health, developmental disorders, hypertension, high cholesterol
CPM	T2D	Tier 2	Y	Y	N	0	
IGFBP1	T2D	Tier 2	Y	Y	N	1	Growth failure due to insulin-like growth factor 1 deficiency
IGFBP2	T2D	Tier 2	Y	Y	N	1	Growth failure due to insulin-like growth factor 1 deficiency
ADIPOQ	T2D	Tier 3A	N	Y	N	0	
APOE	CAD	Tier 3A	N	Y	N	5	Zinc deficiency
CFH	T2D	Tier 3A	N	Y	N	5	Zinc deficiency, malnutrition, ear and respiratory infections
CFI	T2D	Tier 3A	N	Y	N	3	Zinc deficiency, malnutrition, ear and respiratory infections
INHBC	T2D	Tier 3A	N	Y	N	0	

List of PGS-associated proteins with nominal evidence ($P < 0.05$) of causal disease effects in mediation analysis (Fig. 2d) that are part of the druggable genome. Full details of each drug and interaction are provided in Supplementary Table 9. Y, yes; N, no. [a]Evidence of druggability in Finan et al.[22]. Tier 1: targets of approved small molecules, biotherapeutic drugs and clinical-phase drug candidates. Tier 2: targets with known bioactive drug-like small-molecule binding partners as well as those with ≥50% identity (over ≥75% of the sequence) with approved drug targets. Tier 3A: secreted or extracellular proteins, proteins with more distant similarity to approved drug targets and members of key druggable gene families not already included in tier 1 or 2, with genes that were in proximity (±50 kb) to a GWAS SNP and had an extracellular location. [b]The protein is targeted, or predicted to be targeted, by a small molecule. [c]The protein is targeted, or predicted to be targeted, by a biotherapeutic (monoclonal antibody/ enzyme or other protein). [d]The protein is involved in absorption, distribution, metabolism or excretion (ADME) of a compound. The information in these preceding columns was obtained from Table S1 in Finan et al.[22]. [e]The number of drugs or compounds in DrugBank database v.5.17 (https://go.drugbank.com/releases/5-1-7) that interact with the protein.

biology. However, PGS also hold additional promise for identifying molecular pathways in the development and progression of disease[8,29]. In this study, we identified 49 plasma proteins significantly associated with PGS for cardiometabolic disease in a healthy pre-disease cohort. Twenty-eight of these proteins were associated with increased risk of future disease and 16 were nominally significant mediators of T2D or CAD, including 12 druggable targets, suggesting that their modulation may potentially attenuate disease risk.

The vast majority of PGS–protein associations were highly polygenic, including for several well-known cardiometabolic disease proteins. This polygenicity was driven by aggregate modest polygenic effects on protein levels from across the genome, which were independent of *cis*- and *trans*-pQTLs and also present for proteins without pQTLs or for which sample sizes have not yet been sufficient for pQTL detection. This highlights the complementarity of PGS to established approaches that utilize information at a single locus, such as Mendelian randomization, colocalization and fine-mapping. However, it is important to recognize that mediation analysis provides weaker evidence of causality than these established single-locus approaches since it is more difficult to rule out confounding (either from measured or unmeasured factors)[37], especially since PGS by design capture horizontal pleiotropy.

Our findings identify new potential targets of cardiometabolic disease that both are supported by human genetic evidence of causality and may be amenable to pharmacological manipulation. Our strongest results were for IGFBP2, a druggable target[22], as a replicable mediator of polygenic risk and incident T2D. IGFBP2 is involved in the regulation of glucose uptake into adipocytes and is associated with increased insulin sensitivity and decreased adipogenesis[38–41]. Increased plasma IGFBP2 levels have been associated with lower T2D risk[35,42,43]; our findings are directionally consistent and indicate that this association is likely causal. Ten additional druggable proteins were found to be new mediators between polygenic risk and incident T2D.

Twelve new protein associations were found for CAD, CKD or T2D. Among these, the strongest evidence was for alcohol dehydrogenase 4 (ADH4), which is involved in a number of metabolic pathways[44] and was found to be both a mediator of polygenic risk and incident T2D and a druggable target. Furthermore, several new associations concerned proteins with sparse literature on their function; for example, crystallin zeta like 1 (CRYZL1) was associated with polygenic risk and incident CAD; however, little is known about CRYZL1 beyond its gene identification[45].

Proteomic data are becoming increasingly available in cohorts of large sample sizes, such as the planned proteomic profiling of UK Biobank participants. Proteomic platforms are also increasing their coverage of the human proteome. Therefore, we anticipate that our PGS mediation analysis approach will enable the identification of further causal proteins for cardiometabolic and other polygenic diseases in future studies.

Overall, this study demonstrates that PGS can be utilized to elucidate new disease biology with therapeutic potential and provides a useful study design for future studies into the molecular drivers of polygenic disease.

Methods

INTERVAL cohort. INTERVAL is a cohort of approximately 50,000 participants nested within a randomized trial studying the safety of varying the frequency of blood donation[24,25]. Participants were blood donors aged 18 years and older (median 44 years of age; 49% women) recruited between June 2012 and June 2014 from 25 centres across England. The collection of their blood samples for research

purposes was done using standard protocols[24]: blood samples for research purposes were collected in 6-ml EDTA tubes using standard venepuncture protocols. The tubes were inverted three times and transferred at ambient temperature to the UK Biocentre for processing. Plasma was extracted into two 0.8-ml plasma aliquots by centrifugation and subsequently stored at −80 °C before use. Participants gave written informed consent and this study was approved by the National Research Ethics Service (no. 11/EE/0538).

Electronic health records were obtained for all INTERVAL participants from the January 2021 release of the National Health Service (NHS) Hospital Episode Statistics database (https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics) for all events up to 8 February 2020, before the onset of the COVID-19 pandemic in England. The median and maximum follow-up times were 6.9 years and 7.7 years, respectively. The earliest available hospital record for any INTERVAL participant was from 25 March 1999, with a maximum retrospective follow-up of 13.6 years. These records came in the form of International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) codes[46] and were subsequently made available to analysts after summarization into 301 end points using the CALIBER rule-based phenotyping algorithms[47] (https://www.caliberresearch.org/portal). ICD-10 codes contributed to each event regardless of whether they coded for primary or non-primary diagnoses in the hospital records.

Genotyping, quality control and imputation of INTERVAL participants were performed as described previously[48]: participants were genotyped using the Affymetrix UK Biobank Axiom array in ten batches. Samples were removed if they had sex mismatch, had extreme heterozygosity, were of non-European ancestry or were duplicate samples. Related samples were removed by excluding one sample from each pair of close relatives (first- or second-degree; identity by descent $\hat{\pi} > 0.187$). Genotyped variants were removed if they were monomorphic, bi-allelic and had Hardy–Weinberg equilibrium $P < 5 \times 10^{-6}$ or call rate <99%. SHAPEIT3 was used to phase variants; imputation to the UK10K/1000 Genomes panel was performed using the Sanger Imputation Server (https://imputation.sanger.ac.uk).

Protein levels in INTERVAL were quantified using SOMAscan assays, processed and quality-controlled as described previously[26]: relative concentrations of 4,034 SOMAscan aptamers were measured in 3,562 INTERVAL participants in two batches by SomaLogic using v.3 of the SOMAscan platform. Aptamer concentrations (relative fluorescence units) were natural log-transformed and then adjusted within each batch for participant age, sex, the first three genetic principal components and time between blood draw and sample processing (<1 or >1 day); the residuals were then inverse rank normal-transformed. In this study, we further adjusted the normalized protein levels used in previous studies for batch number and filtered to 3,793 high-quality aptamers targeting 3,438 proteins after obtaining the latest information about aptamer sensitivity and specificity from SomaLogic. Aptamers were excluded if, in v.4 of the SOMAscan platform, they (1) targeted non-human proteins, (2) measured the fusion construct rather than the target protein or (3) measured a contaminant. A curated information sheet for all 4,034 aptamers is provided in Supplementary Data 1. The distributions of aptamer levels and associations with covariates before and after quality control are given in Supplementary Data 2.

In total, 3,087 INTERVAL participants, without prevalent cardiometabolic disease (see below) and with matched genotype, proteomic and electronic health record data available for the primary analyses, passed quality control.

Prevalent disease exclusion. The NHS Blood and Transplant blood donation eligibility criteria (https://www.blood.co.uk/who-can-give-blood/) meant that there were built-in exclusions for the INTERVAL cohort for people with a history of major diseases, recent illness or infection. Specifically, for cardiometabolic diseases, the blood donation eligibility criteria excluded individuals who had been diagnosed with AF, had a history of any stroke or a history of major heart disease, including heart failure, coronary thrombosis, myocardial infarction, cardiomyopathy, ischaemic heart disease and arrhythmia, or surgery for non-congenital heart conditions. Use of aspirin or other blood thinners to control elevated blood pressure (hypertension) also made people ineligible to donate blood and participate in the INTERVAL cohort. Individuals with T2D were ineligible unless their T2D was well controlled by diet alone, did not require regular insulin treatment and the individual had not required insulin treatment for at least 4 weeks before attempting blood donation. Extended details on the blood donation criteria eligibility for specific diseases, medications and lifestyle factors can be found at https://my.blood.co.uk/knowledgebase.

In addition to intrinsic exclusion due to the blood donation eligibility criteria, participants were excluded from the analyses if they had any events relating to cardiometabolic disease before baseline assessment. Among the 301 CALIBER end points, we classified 48 as cardiometabolic disease or having potential to introduce reverse causality by modifying risk for incident AF, CAD, CKD, IS or T2D (Supplementary Table 1). In total, 87 participants (2.7%) were excluded, predominantly due to prevalent hypertension ($n = 57$ events; 66% of excluded participants) and prevalent diabetes ($n = 11$ events; 13% of excluded participants), with all others accounting for less than 5% of excluded participants (Supplementary Table 1).

PGS. PGS were derived in a consistent manner by linkage disequilibrium (LD) thinning, at an r^2 threshold of 0.9, the latest GWAS summary statistics for each respective disease (Supplementary Information). The GWAS summary statistics used to derive the AF, CKD and T2D PGS were those published by Nielsen et al.[13] (GCST006414), Wuttke et al.[14] (GCST008065) and Mahajan et al.[15] (GCST007517), respectively. The PGS for CAD and IS used in this study were our previously published CAD[49] and stroke[50] meta-PGS. The CAD PGS was derived from the meta-analysis of three PGS for CAD, including a PGS derived as described above from the GWAS summary statistics published by Nikpay et al.[51]. The IS PGS was derived from the meta-analysis of PGS for IS and its risk factors, including a PGS derived as described above from the GWAS summary statistics for IS published by Malik et al.[16]. Each PGS comprised 1.75–3.23 million single-nucleotide polymorphisms (SNPs) genome-wide and is available to download through the Polygenic Score Catalog[52] (https://www.pgscatalog.org/) with accession nos PGS000727 (AF), PGS000018 (CAD), PGS000728 (CKD), PGS000039 (IS) and PGS000729 (T2D). All PGS were derived from the GWAS summary statistics including only individuals with European ancestry. See Supplementary Information and Extended Data Fig. 8 for details on PGS validation.

The levels of each PGS (sum of dosages × weights) were computed in INTERVAL from probabilistic dosage data using PLINK v.2 (ref. [53]) after mapping PGS variants to those available in the INTERVAL genotype data (Supplementary Information). The levels of each PGS were adjusted for the first ten principal components of the imputed genotype data and standardized to have mean of 0 and s.d. of 1 before downstream statistical analyses.

PGS–protein associations. Each of the five PGS was tested for association with each of the 3,793 aptamers using linear regression (Fig. 1a,b and Extended Data Fig. 3). PGS and proteins were adjusted for covariates and normalized before model fitting (see above). Linear regression coefficients were averaged where multiple high-quality aptamers targeted the same protein (Supplementary Information). FDR correction was subsequently applied across the 3,438 P values (1 per protein) for each PGS separately. Details on aptamer specificity and sensitivity are given in Supplementary Table 2 for the 54 aptamers targeting the 49 PGS-associated proteins; aptamer-specific estimates of PGS on protein levels are detailed in Supplementary Table 3 for the 5 PGS-associated proteins targeted by more than one aptamer (WFIKKN2, GPD1, IGFBP1, IGFBP2 and SHBG).

Polygenicity of PGS–protein associations. To quantify the polygenicity of PGS–protein associations (Fig. 1c and Extended Data Fig. 7), we performed a multistep experiment to determine the proportion of the genome required to explain that association. First, we split the given PGS into separate scores for each of the 1,703 approximately independent LD blocks estimated in Europeans from the 1000 Genomes reference panel by Berisa and Pickrell[54] (https://bitbucket.org/nygcresearch/ldetect-data/src/master/EUR/fourier_ls-all.bed). Next, we tested each of these 1,703 scores for association with the given protein (Supplementary Data 3e). Then, we retested the PGS to protein association, progressively removing independent LD blocks, at each step removing the LD block whose score had the strongest association with the protein. From this, we quantified polygenicity (Fig. 1c) based on the LD blocks needed to be removed from the given PGS to attenuate the PGS–protein association (so that association $P > 0.05$; Supplementary Data 3f) as the sum of removed LD block sizes/sum of all LD block sizes (that is, the proportion of the genome removed). Extended Data Fig. 7 shows the independent LD blocks contributing to the polygenicity of each PGS–protein association.

Independent contributions of PGS and pQTLs to protein levels. Multivariable linear regression models were fitted for each protein on PGS levels and pQTL dosages to estimate their independent contributions to protein levels (Supplementary Table 4). The pQTLs used for each protein were (1) conditionally independent pQTLs mapped in INTERVAL and published by Sun et al.[26], which included both *cis*-pQTLs (within 1 Mb of the encoding gene) and *trans*-pQTLs passing the *trans* significance threshold of $P < 1.5 \times 10^{-11}$; (2) *trans*-pQTLs with $P < 1.5 \times 10^{-11}$ (lead variant only) for proteins not published in Sun et al.[26] (B2M, DUSP26 and FTMT); and (3) hierarchically significant[55,56] *cis*-pQTLs (lead variant only) mapped in this study (Supplementary Data 4 and Supplementary Information) for proteins without *cis*-pQTLs passing the *trans*-pQTL significance threshold above (ACY1, ADIPOQ, APOE, CST3, GPD1, PTPRU, SHBG and UST).

Incident disease associations. PGS and protein levels were tested for association with incident disease using Cox proportional hazards models adjusting for age and sex (Fig. 2b and Extended Data Fig. 8) using the survival package (version 3.2-7) in R. The timescale used was time from baseline to first event of the relevant disease or to the latest available date in the hospital records (8 February 2020). PGS and proteins were adjusted for covariates and normalized before model fitting (see above). Cox model coefficients were averaged where multiple high-quality aptamers targeted the same protein (Supplementary Information).

Incident disease events for AF, CAD, CKD, IS and T2D were defined as the first hospital episode for the closest matching CALIBER phenotype[47]. Incident AF events were defined as any hospital episode with the ICD-10 code I48. Incident IS events were defined as any hospital episode with the ICD-10 code I63 or I69.3.

Incident CAD events were defined as any hospital episode with ICD-10 code I21–I23, I24.1 or I25.2 (CALIBER end point myocardial infarction). The closest matching CALIBER phenotype for T2D was for diabetes more broadly, including ICD-10 codes for any hospital episode for T1D or T2D or complications thereof: E10–E14, G59.0, G63.2, H28.0, H36.0, M14.2, N08.3 or O24.0–O24.3. However, we note that individuals with T1D are not eligible to donate blood and adult onset of T1D is relatively rare compared to T2D[57]. The closest matching CALIBER phenotype for CKD was for end-stage renal disease more broadly, defined as any hospital episode with the ICD-10 codes N16.5, N18.5, T82.4, T86.1, Y60.2, Y61.2, Y84.1, Z49.1, Z49.2, Z94.0 and Z99.2.

Mediation analysis. Mediation analysis was used to identify causal proteins by identifying the PGS-associated proteins that partially mediate the association of PGS on disease (Fig. 2d). This approach uses the counterfactual framework to infer causal effects[30–32] and can be adapted to this setting because the arrow of causality between PGS and any associated phenotype can only flow in one direction since the PGS is fixed at conception (that is, the underlying alleles in each person cannot be modified later in life by protein levels or the development of cardiometabolic disease). In this study, we used the natural effects model developed by Vansteelandt et al.[58], which is available in the medflex R package (version 0.6-7 used in this study)[59], to estimate natural indirect effects (effects of PGS on disease through protein levels) on the log odds scale by imputing unobserved counterfactuals. Standard errors were computed using the robust sandwich estimator[60], from which 95% confidence intervals (CIs) and *P* values were calculated. The percentage of PGS–disease associations mediated by each protein and 95% CIs were subsequently computed as the natural indirect effect and its 95% CI was divided by the total effect estimated by each mediation test. Multiple mediation analysis[61] was performed using the R package mma (version 10.3.2)[62] to quantify the proportion of PGS–disease association mediated by the 15 causal T2D proteins.

Mendelian randomization. Two-sample Mendelian randomization[33] was also performed as an orthogonal approach to identify proteins that may play a causal role in disease (Extended Data Fig. 9 and Supplementary Tables 6 and 7). PGS-associated proteins were tested provided they had three or more independent, as determined by LD ($r^2 < 0.1$), hierarchically significant *cis*-pQTLs after mapping *cis*-pQTLs to the GWAS summary statistics (Supplementary Information) using five different Mendelian randomization methods[63–66], each of which makes use of information across three or more instruments to estimate causal effects, with each method differentially robust to different sources of bias, to obtain a consensus (median) estimate of causal effects of protein levels on disease risk (Supplementary Information). Hierarchically significant *cis*-pQTLs and tagging variants (LD $r^2 > 0.1$) were excluded where they encoded changes to protein structure[67] (for example, missense mutations) and therefore potentially reflected differences in aptamer binding affinity rather than regulation of protein levels (Supplementary Information). Aptamers were also excluded if they had similar affinity for/comparable binding to multiple proteins or differential binding to specific isoforms (Supplementary Table 3 and Supplementary Information).

In total, 11 of the 49 PGS-associated proteins could be tested. GWAS summary statistics were obtained from Nelson et al.[17] for CAD (GCST004787), Wuttke et al.[14] for CKD (GCST008065), Malik et al.[16] for IS (GCST006906) and Mahajan et al.[15] for T2D (GCST007518). In all cases, we used the GWAS summary statistics for the samples of recent European ancestry. For T2D, we used the BMI-adjusted GWAS summary statistics to avoid false positive causal estimates arising where pQTLs influence T2D risk through BMI rather than through the tested protein (horizontal pleiotropy). We considered there to be a significant causal effect where $P < 0.05$ along with no significant evidence that causal effects were due to associations of the pQTLs with some other causal risk factor (horizontal pleiotropy; Egger intercept[66] $P > 0.05$). Analysis was performed using the R package MendelianRandomization (version 0.5.0)[68]. Colocalization analysis[69] was also performed where the *cis*-pQTL instruments had $P < 1 \times 10^{-6}$ in the respective GWAS (Supplementary Table 7 and Supplementary Information).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data used in this study are publicly available or deposited in a public repository. The INTERVAL cohort data are available via the European Genome-phenome Archive with study accession no. EGAS00001002555. Dataset access is subject to approval by an independent data access committee since the data contain potentially identifying and sensitive patient information. Response times from the data access committee are typically within 1 week. All other data used in this study are publicly available without restriction. The PGS used in this study are available to download through the Polygenic Score Catalog (https://www.pgscatalog.org/) with accession nos PGS000727 (AF), PGS000018 (CAD), PGS000728 (CKD), PGS000039 (IS) and PGS000729 (T2D). The GWAS summary statistics used to generate new PGS for CKD, T2D and AF in this study are available to download through the GWAS Catalog (https://www.ebi.ac.uk/gwas/) with study accession nos GCST008065 (for the CKD GWAS published by Wuttke et al.[14]), GCST007517 (for the T2D GWAS published by Mahajan et al.[15]) and GCST006414 (for the AF GWAS published by Nielsen et al.[13]). The additional GWAS summary statistics used for Mendelian randomization analysis are also available through the GWAS Catalog with study accession nos GCST004787 (for the CAD GWAS published by Nelson et al.[17]), GCST006906 (for the IS GWAS published by Malik et al.[16]) and GCST007518 (for the T2D GWAS adjusted for BMI published by Mahajan et al.[15]). Full pQTL summary statistics published by Sun et al.[26] for all SomaLogic SOMAscan aptamers are available to download from https://www.phpc.cam.ac.uk/ceu/proteins/. The DrugBank database is publicly available to download at https://www.drugbank.ca/releases/latest. Summary statistics for all statistical tests are available in Supplementary Data 3; the additional *cis*-pQTLs mapped in this study are provided in Supplementary Data 4.

Code availability

The code used to generate the results of this study, along with a detailed list of software and versions, is available on GitHub (https://github.com/sritchie73/cardiometabolic_prs_plasma_proteome/), which is permanently archived by Zenodo[70] at https://doi.org/10.5281/zenodo.4762747.

Received: 14 May 2021; Accepted: 14 September 2021;
Published online: 08 November 2021

References

1. Purcell, S. M. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
2. Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
3. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
4. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).
5. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
6. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
7. McCarthy, M. I. & Mahajan, A. The value of genetic risk scores in precision medicine for diabetes. *Expert Rev. Precis. Med. Drug Dev.* **3**, 279–281 (2018).
8. *International Common Disease Alliance Recommendations and White Paper* v.1.0 (ICDA Organizing Committee and Working Groups, 2020); https://drive.google.com/file/d/16SVJ5lbneN9hB9E03PZMhpescAN527HO/view
9. Erdmann, J., Kessler, T., Munoz Venegas, L. & Schunkert, H. A decade of genome-wide association studies for coronary artery disease: the challenges ahead. *Cardiovasc. Res.* **114**, 1241–1257 (2018).
10. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
11. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
12. Zheng, J. et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).
13. Nielsen, J. B. et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).
14. Wuttke, M. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
15. Mahajan, A. et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
16. Malik, R. et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524–537 (2018).
17. Nelson, C. P. et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
18. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
19. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife* **10**, e58615 (2021).
20. Liu, X., Li, Y. I. & Pritchard, J. K. *Trans* effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034.e6 (2019).
21. Sun, L. et al. Polygenic risk scores in cardiovascular risk prediction: a cohort study and modelling analyses. *PLoS Med.* **18**, e1003498 (2021).
22. Finan, C. et al. The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, eaag1166 (2017).

23. Ghoussaini, M. et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).
24. Moore, C. et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).
25. Di Angelantonio, E. et al. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet* **390**, 2360–2371 (2017).
26. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
27. Jovanović, D., Krstivojević, P., Obradović, I., Durdević, V. & Dukanović, L. Serum cystatin C and beta2-microglobulin as markers of glomerular filtration rate. *Ren. Fail.* **25**, 123–133 (2003).
28. van Poelje, P. D., Dang, Q. & Erion, M. D. Fructose-1,6-bisphosphatase as a therapeutic target for type 2 diabetes. *Drug Discov. Today Ther. Strateg.* **4**, 103–109 (2007).
29. Holmes, M. V. & Davey Smith, G. Can Mendelian randomization shift into reverse gear? *Clin. Chem.* **65**, 363–366 (2019).
30. Imai, K., Keele, L. & Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* **15**, 309–334 (2010).
31. Imai, K., Keele, L., Tingley, D. & Yamamoto, T. Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *Am. Polit. Sci. Rev.* **105**, 765–789 (2011).
32. Hernán, M. A. A definition of causal effect for epidemiological research. *J. Epidemiol. Community Health* **58**, 265–271 (2004).
33. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
34. Ding, E. L. et al. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. *N. Engl. J. Med.* **361**, 1152–1163 (2009).
35. Gudmundsdottir, V. et al. Circulating protein signatures and causal candidates for type 2 diabetes. *Diabetes* **69**, 1843–1853 (2020).
36. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
37. VanderWeele, T. J. Mediation analysis: a practitioner's guide. *Annu. Rev. Public Health* **37**, 17–32 (2016).
38. Russo, V. C., Azar, W. J., Yau, S. W., Sabin, M. A. & Werther, G. A. IGFBP-2: the dark horse in metabolism and cancer. *Cytokine Growth Factor Rev.* **26**, 329–346 (2015).
39. Assefa, B. et al. Insulin-like growth factor (IGF) binding protein-2, independently of IGF-1, induces GLUT-4 translocation and glucose uptake in 3T3-L1 adipocytes. *Oxid. Med. Cell. Longev.* **2017**, 3035184 (2017).
40. Wheatcroft, S. B. et al. IGF-binding protein-2 protects against the development of obesity and insulin resistance. *Diabetes* **56**, 285–294 (2007).
41. Hedbacker, K. et al. Antidiabetic effects of *IGFBP2*, a leptin-regulated gene. *Cell Metab.* **11**, 11–22 (2010).
42. Rajpathak, S. N. et al. Insulin-like growth factor axis and risk of type 2 diabetes in women. *Diabetes* **61**, 2248–2254 (2012).
43. Wittenbecher, C. et al. Insulin-like growth factor binding protein 2 (IGFBP-2) and the risk of developing type 2 diabetes. *Diabetes* **68**, 188–197 (2019).
44. Yin, S.-J., Chou, C.-F., Lai, C.-L., Lee, S.-L. & Han, C.-L. Human class IV alcohol dehydrogenase: kinetic mechanism, functional roles and medical relevance. *Chem. Biol. Interact.* **143–144**, 219–227 (2003).
45. Kim, M. Y. et al. Identification of a zeta-crystallin (quinone reductase)-like 1 gene (*CRYZL1*) mapped to human chromosome 21q22.1. *Genomics* **57**, 156–159 (1999).
46. *International Statistical Classification of Diseases and Related Health Problems: Instruction Manual* (World Health Organization, 2004).
47. Kuan, V. et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit. Health* **1**, e63–e77 (2019).
48. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429.e19 (2016).
49. Inouye, M. et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
50. Abraham, G. et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat. Commun.* **10**, 5819 (2019).
51. Nikpay, M. et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
52. Lambert, S. A. et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nature* **53**, 420–425 (2021).
53. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
54. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).
55. Peterson, C. B., Bogomolov, M., Benjamini, Y. & Sabatti, C. TreeQTL: hierarchical error control for eQTL findings. *Bioinformatics* **32**, 2556–2558 (2016).
56. Huang, Q. Q., Ritchie, S. C., Brozynska, M. & Inouye, M. Power, false discovery rate and winner's curse in eQTL studies. *Nucleic Acids Res.* **46**, e133 (2018).
57. Atkinson, M. A., Eisenbarth, G. S. & Michels, A. W. Type 1 diabetes. *Lancet* **383**, 69–82 (2014).
58. Vansteelandt, S., Bekaert, M. & Lange, T. Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiol. Methods* **1**, 130–158 (2012).
59. Steen, J., Loeys, T., Moerkerke, B. & Vansteelandt, S. medflex: an R package for flexible mediation analysis using natural effect models. *J. Stat. Softw.* **76**, 1–46 (2017).
60. Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22 (1986).
61. Yu, Q., Fan, Y. & Wu, X. General multiple mediation analysis with an application to explore racial disparities in breast cancer survival. *J. Biom. Biostat.* **5**, 1–9 (2014).
62. Yu, Q. & Li, B. mma: an R package for mediation analysis with multiple mediators. *J. Open Res. Softw.* **5**, 11 (2017).
63. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
64. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
65. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **46**, 1985–1998 (2017).
66. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
67. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
68. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).
69. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
70. Ritchie, S. sritchie73/cardiometabolic_prs_plasma_proteome: Nature revisions round 3. *Zenodo* https://doi.org/10.5281/zenodo.4762747 (2021).
71. Rasmussen, K. L. Plasma levels of apolipoprotein E, *APOE* genotype and risk of dementia and ischemic heart disease: a review. *Atherosclerosis* **255**, 145–155 (2016).
72. Sofat, R. et al. Circulating apolipoprotein E concentration and cardiovascular disease risk: meta-analysis of results from three studies. *PLoS Med.* **13**, e1002146 (2016).
73. Nikpay, M., Soubeyrand, S., Tahmasbi, R. & McPherson, R. Multiomics screening identifies molecular biomarkers causally associated with the risk of coronary artery disease. *Circ. Genom. Precis. Med.* **13**, e002876 (2020).
74. Ruttmann, E. et al. γ-Glutamyltransferase as a risk factor for cardiovascular disease mortality: an epidemiological investigation in a cohort of 163 944 Austrian adults. *Circulation* **112**, 2130–2137 (2005).
75. Lee, D. S. et al. Gamma glutamyl transferase and metabolic syndrome, cardiovascular disease, and mortality risk: the Framingham Heart Study. *Arterioscler. Thromb. Vasc. Biol.* **27**, 127–133 (2007).
76. Kojima, Y. et al. Progranulin expression in advanced human atherosclerotic plaque. *Atherosclerosis* **206**, 102–108 (2009).
77. Pugeat, M. et al. Interrelations between sex hormone-binding globulin (SHBG), plasma lipoproteins and cardiovascular risk. *J. Steroid Biochem. Mol. Biol.* **53**, 567–572 (1995).
78. Sutton-Tyrrell, K. et al. Sex-hormone-binding globulin and the free androgen index are related to cardiovascular risk factors in multiethnic premenopausal and perimenopausal women enrolled in the Study of Women Across the Nation (SWAN). *Circulation* **111**, 1242–1249 (2005).
79. Liu, P. Y., Death, A. K. & Handelsman, D. J. Androgens and cardiovascular disease. *Endocr. Rev.* **24**, 313–340 (2003).
80. Li, G.-S. et al. Do the mutations of *C1GALT1C1* gene play important roles in the genetic susceptibility to Chinese IgA nephropathy? *BMC Med. Genet.* **10**, 101 (2009).
81. Yoshida, T. et al. Association of gene polymorphisms with chronic kidney disease in high- or low-risk subjects defined by conventional risk factors. *Int. J. Mol. Med.* **23**, 785–792 (2009).

82. Foster, M. C., Yang, Q., Hwang, S.-J., Hoffmann, U. & Fox, C. S. Heritability and genome-wide association analysis of renal sinus fat accumulation in the Framingham Heart Study. *BMC Med. Genet.* **12**, 148 (2011).
83. Madsen, T. E. et al. Circulating SHBG (sex hormone-binding globulin) and risk of ischemic stroke: findings from the WHI. *Stroke* **51**, 1257–1264 (2020).
84. Baumeier, C. et al. Caloric restriction and intermittent fasting alter hepatic lipid droplet proteome and diacylglycerol species and prevent diabetes in NZO mice. *Biochim. Biophys. Acta* **1851**, 566–576 (2015).
85. Ngo, D. et al. Proteomic profiling reveals novel biomarkers and pathways in type 2 diabetes risk. *JCI Insight* **6**, e144392 (2021).
86. Spranger, J. et al. Adiponectin and protection against type 2 diabetes mellitus. *Lancet* **361**, 226–228 (2003).
87. Lau, W., Andrew, T. & Maniatis, N. High-resolution genetic maps identify multiple type 2 diabetes loci at regulatory hotspots in African Americans and Europeans. *Am. J. Hum. Genet.* **100**, 803–816 (2017).
88. Suckale, J. & Solimena, M. The insulin secretory granule as a signaling hub. *Trends Endocrinol. Metab.* **21**, 599–609 (2010).
89. Kim-Muller, J. Y. et al. Aldehyde dehydrogenase 1a3 defines a subset of failing pancreatic β cells in diabetic mice. *Nat. Commun.* **7**, 12631 (2016).
90. Voight, B. F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
91. Guevara-Aguirre, J. et al. Growth hormone receptor deficiency is associated with a major reduction in pro-aging signaling, cancer, and diabetes in humans. *Sci. Transl. Med.* **3**, 70ra13 (2011).
92. Rajwani, A. et al. Increasing circulating IGFBP1 levels improves insulin sensitivity, promotes nitric oxide production, lowers blood pressure, and protects against atherosclerosis. *Diabetes* **61**, 915–924 (2012).
93. Xu, S., Lind, L., Zhao, L., Lindahl, B. & Venge, P. Plasma prolylcarboxypeptidase (angiotensinase C) is increased in obesity and diabetes mellitus and related to cardiovascular dysfunction. *Clin. Chem.* **58**, 1110–1115 (2012).
94. Grarup, N., Sandholt, C. H., Hansen, T. & Pedersen, O. Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia* **57**, 1528–1541 (2014).
95. Dwinovan, J., Colella, A. D., Chegeni, N., Chataway, T. K. & Sokoya, E. M. Proteomic analysis reveals downregulation of housekeeping proteins in the diabetic vascular proteome. *Acta Diabetol.* **54**, 171–190 (2017).
96. Lopez, P. H. et al. Mice lacking sialyltransferase ST3Gal-II develop late-onset obesity and insulin resistance. *Glycobiology* **27**, 129–139 (2017).
97. Kato, N. Insights into the genetic basis of type 2 diabetes. *J. Diabetes Investig.* **4**, 233–244 (2013).
98. Levey, A. S. & Coresh, J. Chronic kidney disease. *Lancet* **379**, 165–180 (2012).
99. Levey, A. S. et al. A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150**, 604–612 (2009).

Acknowledgements

Participants in the INTERVAL randomized controlled trial were recruited with the active collaboration of NHS Blood and Transplant (www.nhsbt.nhs.uk), which has supported fieldwork and other elements of the trial. DNA extraction and genotyping were co-funded by the National Institute for Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC) (no. BRC-1215-20014). Olink Proteomics assays were funded by Biogen. SomaLogic assays were funded by Merck and the NIHR Cambridge BRC (no. BRC-1215-20014). The academic coordinating centre for INTERVAL was supported by core funding from the NIHR Blood and Transplant Research Unit in Donor Health and Genomics (no. NIHR BTRU-2014-10024), UK Medical Research Council (MRC) (no. MR/L003120/1), British Heart Foundation (nos SP/09/002, RG/13/13/30194 and RG/18/13/33946) and the NIHR Cambridge BRC (no. BRC-1215-20014). A complete list of the investigators and contributors to the INTERVAL trial is provided in ref. [25]. The academic coordinating centre thanks blood donor centre staff and blood donors for participating in the INTERVAL trial. This work was supported by Health Data Research UK, which is funded by the UK MRC, Engineering and Physical Sciences Research Council (EPSRC), Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. This study was also supported by the Victorian Government's Operational Infrastructure Support programme. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery operated by the University of Cambridge Research Computing Service (https://www.hpc.cam.ac.uk/high-performance-computing), provided by Dell EMC and Intel using tier-2 funding from the EPSRC (capital grant no. EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk). This work uses data provided by patients and collected by the NHS and Public Health England as part of their care and support. Data on hospital episode statistics, mortality and cancer registration were obtained from NHS Digital (data sharing agreement reference no. DARS-NIC-156334-711SX). S.C.R. and J.M. were funded by the NIHR Cambridge BRC (no. BRC-1215-20014). S.A.L. is supported by a Canadian Institutes of Health Research postdoctoral fellowship (no. MFE-171279). G.A. was supported by a National Health and Medical Research Council of Australia Early Career Fellowship (no. 1090462). S.B. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (no. 204623/Z/16/Z). A.V.K. was supported by grants from the National Human Genome Research Institute (award nos 1K08HG010155 and 5UM1HG008895), an institutional grant from the Broad Institute of MIT and Harvard (variant2function) and a Hassenfeld Scholar Award from Massachusetts General Hospital. J.D. holds a British Heart Foundation Professorship and an NIHR Senior Investigator Award. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The views expressed in this manuscript are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Author contributions

S.C.R., S.Z., M. Chaffin, B.G.D., A.C.C., N.S., S.K., A.V.K., A.S.B. and M.I. conceptualized the study. S.C.R., M.A., Y.L. and A.S.B. curated the data. S.C.R. carried out the formal analysis. J.D. and M.I. acquired the funding. S.C.R., M.A., Y.L., B.G.D., A.C.C. and M. Chaffin carried out the investigation. S.C.R., S.A.L., S.M.T., S.L., P.S., J.M., G.A. and M.I. devised the methodology. A.V.K., S.K., A.S.B. and M.I. administered the project. W.H.O., D.J.R., N.A.W., B.G.D., A.C.C., E.D.A., M. Chapman, J.D., A.S.B. and M.I. curated the resources. S.C.R. managed the software. G.A., B.G.D., A.C.C., E.D.A., S.K., A.S.B. and M.I. supervised the study. S.C.R., S.A.L., S.M.T., S.L., P.S., J.M., G.A., S.B. and A.V.K. validated the data. S.C.R. visualized the data. S.C.R. and M.I. wrote the original manuscript draft. S.C.R., S.A.L., S.M.T., S.L., P.S., J.M., G.A., B.G.D., A.C.C., N.S., S.B., A.V.K., J.D., A.S.B. and M.I. reviewed and edited the manuscript draft.

Competing interests

Several authors are now employed by or run pharmaceutical companies. All significant contributions to this study were made before these roles and the named companies had no role in the study. M.A. is an employee of AstraZeneca. P.S. is an employee of Roche. J.M. is an employee of Genomics PLC. G.A. is an employee of CSL Limited. S.K. is the chief executive officer of Verve Therapeutics. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42255-021-00478-5.

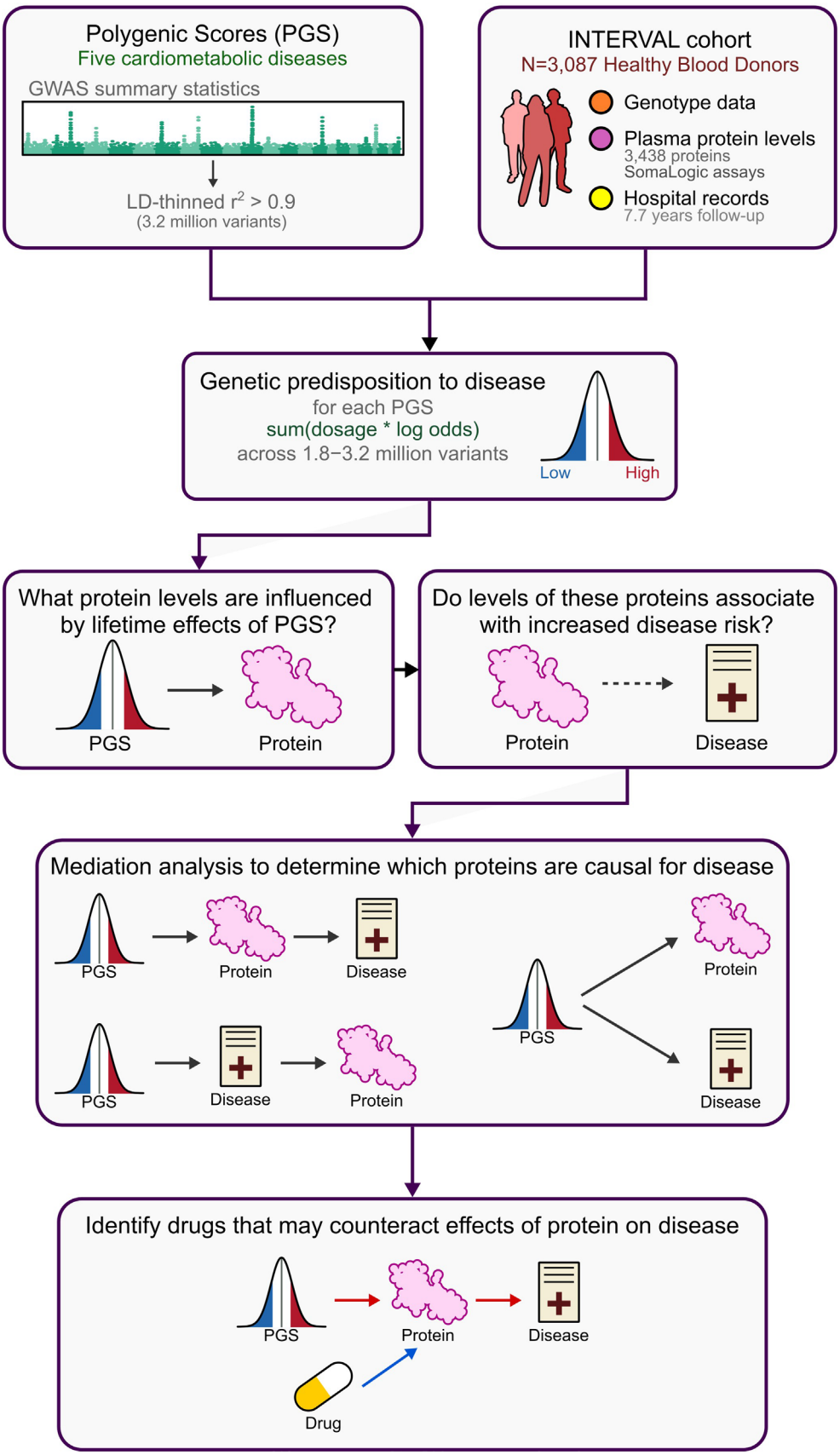
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42255-021-00478-5.

Correspondence and requests for materials should be addressed to Scott C. Ritchie or Michael Inouye.

Peer review information *Nature Metabolism* thanks Matthew Nelson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Isabella Samuelson.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Study schematic. Overview of the study design.

	Cohort characteristics
Participants	N=3,087
Women	N=1,528 (49%)
Age (years)	Median: 44.0 (Range: 18.0−75.6, IQR: 30.5−54.7)
Weight (kilograms; kg)	Median: 76.6 (Range: 49.4−177.0, IQR: 66.7−88.0)
Height (meters; m)	Median: 1.73 (Range: 1.07−2.41, IQR: 1.65−1.80)
BMI (kg/m^2)	Median: 25.5 (Range: 13.1−81.5, IQR: 23.1−28.5)

Extended Data Fig. 2 | Cohort characteristics. IQR: interquartile range. Body mass index (BMI) was computed from self-reported height and weight.

	Association with PGS for coronary artery disease					Association with incident coronary artery disease (N=15 events)			Causal effect of PGS on disease risk through protein in mediation analysis		
Protein	Beta	95% CI	P-value	FDR	Polygenicity	HR	95% CI	P-value	% PGS Mediated	95% CI	P-value
APOE	0.081	[0.046, 0.12]	$6x10^{-6}$	0.004	8.8%	1.97	[1.17, 3.33]	0.011	5.40%	[0.4%, 10%]	0.034
TP53I11	0.081	[0.046, 0.12]	$6x10^{-6}$	0.004	9.2%	1.92	[1.14, 3.22]	0.014	5.40%	[-0.6%, 11%]	0.078
CRYZL1	0.081	[0.046, 0.12]	$6x10^{-6}$	0.004	2.6%	1.84	[1.13, 3.00]	0.015	5.30%	[-0.1%, 11%]	0.052
NPTX2	0.069	[0.034, 0.10]	$1x10^{-4}$	0.040	2.8%	1.55	[0.959, 2.52]	0.073	3.00%	[-2.2%, 8.1%]	0.26
GGT2	0.081	[0.045, 0.12]	$7x10^{-6}$	0.004	6.2%	1.55	[0.954, 2.52]	0.077	3.30%	[-2.0%, 8.6%]	0.22
HBQ1	0.071	[0.036, 0.11]	$7x10^{-5}$	0.026	3.9%	1.52	[0.937, 2.47]	0.090	3.10%	[-1.7%, 7.9%]	0.21
CEI	0.091	[0.056, 0.13]	$4x10^{-7}$	0.001	1.9%	1.46	[0.89, 2.38]	0.14	2.40%	[-2.1%, 7.0%]	0.29
DUSP26	0.073	[0.037, 0.11]	$6x10^{-5}$	0.024	12%	1.46	[0.88, 2.43]	0.14	2.30%	[-1.9%, 6.5%]	0.28
PCDHB10	0.070	[0.035, 0.11]	$9x10^{-5}$	0.031	10%	1.38	[0.85, 2.24]	0.19	2.00%	[-1.8%, 5.8%]	0.3
SHBG	-0.079	[-0.11, -0.044]	$1x10^{-5}$	0.005	17%	0.70	[0.38, 1.29]	0.25	0.90%	[-1.8%, 3.6%]	0.53
GRN	0.082	[0.047, 0.12]	$5x10^{-6}$	0.004	0.1%	1.33	[0.78, 2.26]	0.29	1.00%	[-1.2%, 3.1%]	0.38
	Association with PGS for type 2 diabetes					Association with incident type 2 diabetes (N=27 events)					
IGFBP2	-0.095	[-0.13, -0.061]	$5x10^{-8}$	$6x10^{-5}$	15%	0.44	[0.31, 0.64]	$2x10^{-5}$	13.4%	[5.2%, 21%]	0.001
CFH	0.088	[0.054, 0.12]	$4x10^{-7}$	$3x10^{-4}$	21%	2.35	[1.59, 3.48]	$2x10^{-5}$	8.80%	[3.4%, 14%]	0.002
CPM	0.096	[0.062, 0.13]	$3x10^{-8}$	$6x10^{-5}$	22%	1.97	[1.36, 2.87]	$4x10^{-4}$	8.80%	[3.1%, 14%]	0.003
SHBG	-0.11	[-0.14, -0.073]	$8x10^{-10}$	$3x10^{-6}$	23%	0.41	[0.26, 0.62]	$4x10^{-5}$	13.2%	[4.3%, 22%]	0.004
GHR	0.071	[0.036, 0.11]	$6x10^{-5}$	0.014	11%	2.28	[1.53, 3.39]	$5x10^{-5}$	7.90%	[1.9%, 14%]	0.009
CCDC126	-0.070	[-0.10, -0.035]	$7x10^{-5}$	0.014	10%	0.46	[0.32, 0.68]	$8x10^{-5}$	6.90%	[1.3%, 13%]	0.016
PRCP	0.077	[0.043, 0.11]	$7x10^{-6}$	0.003	12%	1.89	[1.29, 2.76]	$1x10^{-3}$	6.60%	[1.2%, 12%]	0.017
PTPRU	0.074	[0.040, 0.11]	$2x10^{-5}$	0.006	18%	1.86	[1.28, 2.70]	0.001	6.10%	[0.9%, 11%]	0.022
IGFBP1	-0.073	[-0.11, -0.038]	$3x10^{-5}$	0.009	18%	0.54	[0.36, 0.81]	0.003	6.70%	[0.7%, 13%]	0.027
ADIPOQ	-0.080	[-0.11, -0.046]	$4x10^{-6}$	0.002	9.3%	0.52	[0.36, 0.76]	$8x10^{-4}$	8.80%	[0.9%, 17%]	0.028
ACY1	0.077	[0.043, 0.11]	$8x10^{-6}$	0.003	9.1%	1.72	[1.19, 2.49]	0.004	5.10%	[0.4%, 9.9%]	0.035
CFI	0.083	[0.049, 0.12]	$2x10^{-6}$	$1x10^{-3}$	23%	1.69	[1.14, 2.50]	0.009	4.60%	[0.3%, 8.8%]	0.037
ADH4	0.070	[0.035, 0.10]	$7x10^{-5}$	0.014	6.5%	1.72	[1.17, 2.54]	0.006	4.20%	[0.1%, 8.3%]	0.043
QPCTL	-0.065	[-0.099, -0.030]	$2x10^{-4}$	0.027	17%	0.51	[0.35, 0.74]	$5x10^{-4}$	6.40%	[0.0%, 13%]	0.048
INHBC	0.066	[0.032, 0.100]	$2x10^{-4}$	0.022	22%	1.65	[1.11, 2.45]	0.013	3.60%	[0.0%, 7.1%]	0.049
APOF	-0.065	[-0.099, -0.031]	$2x10^{-4}$	0.026	5.0%	0.46	[0.31, 0.67]	$6x10^{-5}$	5.90%	[-0.3%, 12%]	0.061
GFRA1	0.068	[0.034, 0.10]	$9x10^{-5}$	0.015	5.3%	1.85	[1.26, 2.72]	0.002	6.00%	[-0.6%, 13%]	0.074
PRSS1	-0.086	[-0.12, -0.052]	$1x10^{-6}$	$5x10^{-4}$	18%	0.58	[0.40, 0.84]	0.004	8.50%	[-1.7%, 19%]	0.1
FAH	0.068	[0.033, 0.10]	$1x10^{-4}$	0.018	14%	1.73	[1.19, 2.52]	0.004	4.10%	[-0.1%, 8.2%]	0.054
ST3GAL2	0.067	[0.032, 0.10]	$1x10^{-4}$	0.021	19%	1.66	[1.15, 2.41]	0.007	2.80%	[-1.1%, 6.7%]	0.17
FBP1	0.070	[0.036, 0.10]	$7x10^{-5}$	0.014	8.6%	1.57	[1.072, 2.30]	0.021	2.00%	[-1.1%, 5.2%]	0.2
WFIKKN2	-0.070	[-0.10, -0.036]	$6x10^{-5}$	0.014	12%	0.65	[0.44, 0.947]	0.025	4.60%	[-1.3%, 10%]	0.11
RIDA	0.063	[0.029, 0.097]	$3x10^{-4}$	0.036	7.5%	1.54	[1.054, 2.25]	0.026	2.90%	[-1.1%, 6.8%]	0.16
TIMP4	-0.069	[-0.10, -0.035]	$8x10^{-5}$	0.015	19%	0.65	[0.44, 0.959]	0.030	3.70%	[-0.9%, 8.4%]	0.12
HS6ST2	-0.067	[-0.10, -0.033]	$1x10^{-4}$	0.021	23%	0.66	[0.44, 0.991]	0.045	2.40%	[-0.2%, 5.0%]	0.07
GPD1	0.066	[0.031, 0.10]	$2x10^{-4}$	0.024	12%	1.44	[0.982, 2.11]	0.062	2.60%	[-1.1%, 6.4%]	0.18
MUSK	0.090	[0.056, 0.12]	$3x10^{-7}$	$2x10^{-4}$	16%	1.44	[0.982, 2.10]	0.062	3.50%	[-0.6%, 7.7%]	0.097
CHGB	-0.075	[-0.11, -0.041]	$2x10^{-5}$	0.005	10%	0.74	[0.51, 1.081]	0.12	2.80%	[-1.8%, 7.4%]	0.23
CYB5R3	-0.062	[-0.096, -0.027]	$4x10^{-4}$	0.049	9.1%	0.74	[0.51, 1.086]	0.12	2.20%	[-0.7%, 5.1%]	0.13
MSMP	-0.068	[-0.10, -0.034]	$1x10^{-4}$	0.017	8.7%	0.74	[0.51, 1.087]	0.13	3.00%	[-2.2%, 8.3%]	0.26
FAM20A	0.063	[0.029, 0.098]	$3x10^{-4}$	0.036	12%	1.27	[0.87, 1.86]	0.21	1.30%	[-1.2%, 3.9%]	0.31
	Association with PGS for chronic kidney disease					Association with incident chronic kidney disease (N=0 events)					
VWC2	0.094	[0.056, 0.13]	$1x10^{-6}$	0.002	21%						
B2M	0.094	[0.056, 0.13]	$1x10^{-6}$	0.002	9.4%						
CST3	0.085	[0.047, 0.12]	$1x10^{-5}$	0.010	7.0%						
UST	-0.081	[-0.12, -0.042]	$4x10^{-5}$	0.023	12%						
C1GALT1C1	-0.079	[-0.12, -0.041]	$5x10^{-5}$	0.026	5.6%						
FTMT	-0.11	[-0.15, -0.076]	$4x10^{-9}$	$1x10^{-5}$	0.1%						
PRSS3	0.082	[0.043, 0.12]	$3x10^{-5}$	0.020	0.2%						
PDE4D/A	-0.076	[-0.11, -0.038]	$9x10^{-5}$	0.040	0.1%						
	Association with PGS for ischaemic stroke					Association with incident ischaemic stroke (N=3 events)					
SHBG	-0.076	[-0.11, -0.042]	$1x10^{-5}$	0.041	17%						

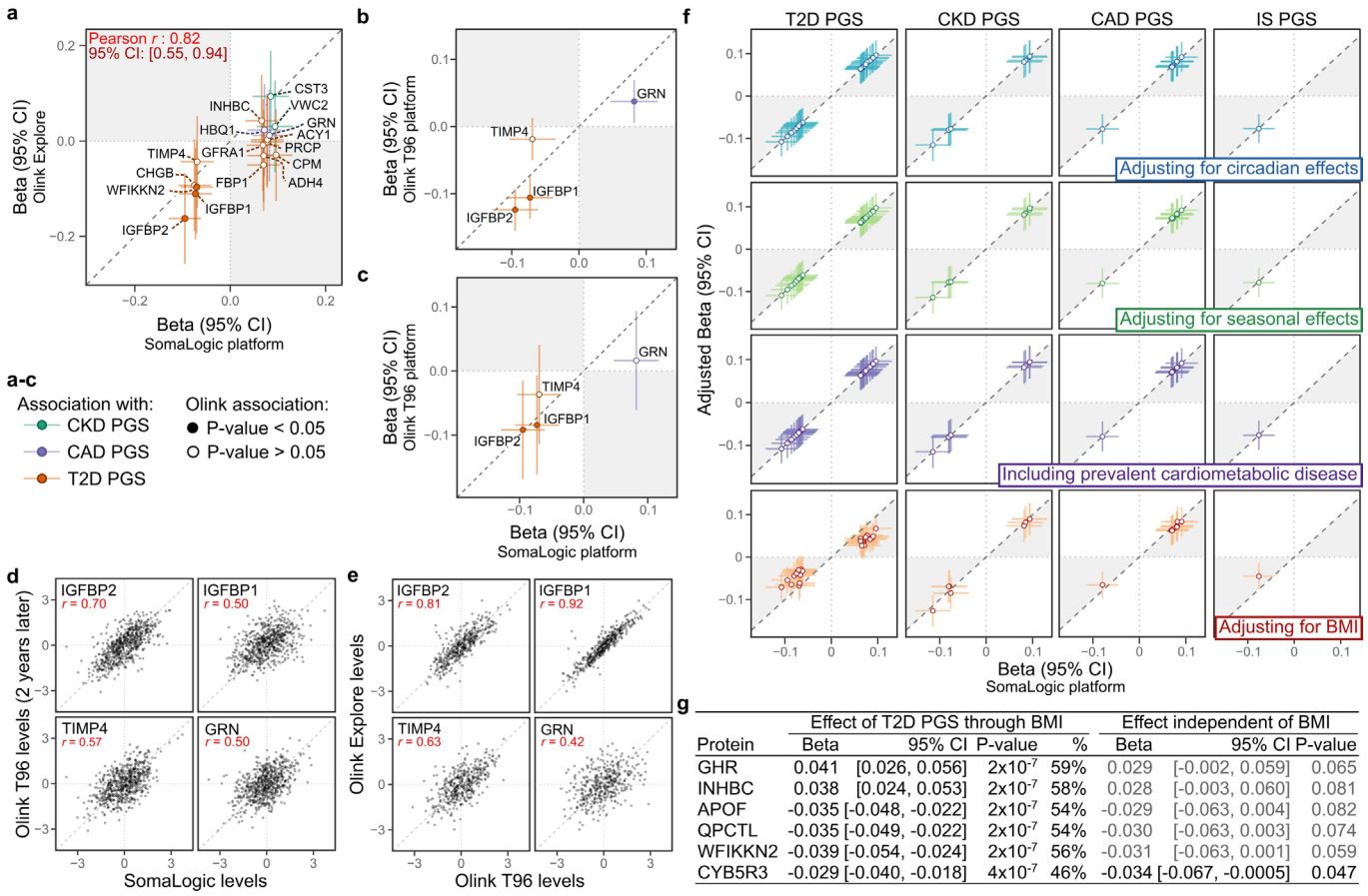
Extended Data Fig. 3 | Summary statistics for PGS to protein to disease associations. Beta: standard deviation change in protein levels per standard deviation increase in PGS (from Fig. 1b) in linear regression adjusting for age, sex, 10 genotype PCs, sample measurement batch, and time between blood draw and sample processing. FDR: Benjamini-Hochberg false discovery rate corrected P-value. FDR correction was applied separately for each PGS to all 3,438 P-values from linear regression of each of the 3,438 measured proteins on the respective PGS. Polygenicity: proportion of the genome (%) required to explain the PGS to protein association (from Fig. 1c). HR: hazard ratio for 7.7 year risk of hospitalisation with the respective disease conferred per standard deviation increase in protein levels (from Fig. 2b) in cox proportional hazard models using follow-up as time scale and adjusting for age, sex, sample measurement batch, and time between blood draw and sample processing. Associations highlighted in red indicate significant associations after Bonferroni correction for the 42 tests ($P < 0.0012$). Associations dulled in grey indicate $P > 0.05$. % PGS Mediated: Percentage of total association between the respective PGS and 7.7 year risk of hospitalisation with the respective disease mediated by the respective protein (from Fig. 2d). Highlighted in red indicates mediation was significant after Bonferroni correction for the 42 tests ($P < 0.0012$). Entries dulled in grey indicate $P > 0.05$. Linear regression, polygenicity, cox proportional hazard models, and mediation analysis were all performed in the same n = 3,087 independent INTERVAL participants. In each instance, 95% CI corresponds to the 95% confidence interval of the respective point estimate. All P-values are two-sided. 95% confidence intervals and P-values could not be formulated for the polygenicity tests. For proteins measured by more than one SomaLogic aptamer (GPD1, IGFBP1, IGFBP2, SHBG, and WFIKKN2) effect sizes were averaged and two-sided P-values were obtained from averaged Z-scores, and aptamer-specific summary statistics are detailed in Supplementary Table 3.

Protein	UniProt	Gene	Chr	Start	PGS	Aptamer	Aptamer target
ACY1	Q03154	*ACY1*	3	52,017,300	T2D	3343-1	Aminoacylase-1
ADH4	P08319	*ADH4*	4	100,044,832	T2D	8325-37	Alcohol dehydrogenase 4
ADIPOQ	Q15848	*ADIPOQ*	3	186,560,463	T2D	3554-24	Adiponectin
APOE*	P02649	*APOE*	19	45,409,039	CAD	2418-55	Apolipoprotein E, isoforms E3 and E4
APOF	Q13790	*APOF*	12	56,754,355	T2D	12370-30	Apolipoprotein F
B2M	P61769	*B2M*	15	45,003,685	CKD	3485-28	Beta-2-microglobulin
C1GALT1C1	Q96EU7	*C1GALT1C1*	X	119,759,529	CKD	5735-54	C1GALT1-specific chaperone 1
CCDC126	Q96EE4	*CCDC126*	7	23,636,998	T2D	6388-21	Coiled-coil domain-containing protein 126
CEI	Q86SI9	*C5orf38*	5	2,752,058	CAD	6378-2	Protein CEI
CFH	P08603	*CFH*	1	196,621,008	T2D	4159-130	Complement factor H
CFI	P05156	*CFI*	4	110,661,848	T2D	2567-5	Complement factor I
CHGB	P05060	*CHGB*	20	5,891,974	T2D	8235-48	Secretogranin-1
CPM	P14384	*CPM*	12	69,244,955	T2D	7768-10	Carboxypeptidase M
CRYZL1	O95825	*CRYZL1*	21	34,961,647	CAD	9207-60	Quinone oxidoreductase-like protein 1
CST3	P01034	*CST3*	20	23,608,534	CKD	2609-59	Cystatin-C
CYB5R3	P00387	*CYB5R3*	22	43,013,846	T2D	7215-18	NADH-cytochrome b5 reductase 3
DUSP26	Q9BV47	*DUSP26*	8	33,448,848	CAD	8967-6	Dual specificity protein phosphatase 26
FAH	P16930	*FAH*	15	80,445,233	T2D	11424-4	Fumarylacetoacetase
FAM20A	Q96MK3	*FAM20A*	17	66,531,257	T2D	6433-57	Pseudokinase FAM20A
FBP1	P09467	*FBP1*	9	97,365,415	T2D	7206-20	Fructose-1,6-bisphosphatase 1
FTMT	Q8N4E7	*FTMT*	5	121,187,650	CKD	8048-9	Ferritin, mitochondrial
GFRA1	P56159	*GFRA1*	10	117,816,436	T2D	3314-74	GDNF family receptor alpha-1
GGT2	P36268	*GGT2*	22	21,562,261	CAD	6334-9	Inactive gamma-glutamyltranspeptidase 2
GHR	P10912	*GHR*	5	42,423,577	T2D	2948-58	Growth hormone receptor
GPD1	P21695	*GPD1*	12	50,497,602	T2D	13697-51 11081-1	Glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic
GRN	P28799	*GRN*	17	42,422,491	CAD	4992-49	Granulins
HBQ1	P09105	*HBQ1*	16	230,333	CAD	7965-25	Hemoglobin subunit theta-1
HS6ST2	Q96MM7	*HS6ST2*	X	131,760,038	T2D	13524-25	Heparan-sulfate 6-O-sulfotransferase 2
IGFBP1	P08833	*IGFBP1*	7	45,927,959	T2D	13741-36 2771-35	Insulin-like growth factor-binding protein 1
IGFBP2	P18065	*IGFBP2*	2	217,498,127	T2D	2570-72 8469-41	Insulin-like growth factor-binding protein 2
INHBC	P55103	*INHBC*	12	57,828,543	T2D	6408-2	Inhibin beta C chain
MSMP	Q1L6U9	*MSMP*	9	35,752,987	T2D	8080-24	Prostate-associated microseminoprotein
MUSK	O15146	*MUSK*	9	113,430,935	T2D	11547-84	Muscle, skeletal receptor tyrosine-protein kinase
NPTX2	P47972	*NPTX2*	7	98,246,597	CAD	6521-35	Neuronal pentraxin-2
PCDHB10	Q9UN67	*PCDHB10*	5	140,571,952	CAD	9963-19	Protocadherin beta-10
PDE4D/A*	Q08499 P27815	*PDE4D* *PDE4A*	5 19	58,264,865 10,527,449	CKD	5255-22	Combined levels of cAMP-specific 3',5'-cyclic phosphodiesterase 4D and 4A
PRCP	P42785	*PRCP*	11	82,535,409	T2D	5722-78	Lysosomal Pro-X carboxypeptidase
PRSS1	P07477	*PRSS1*	7	142,457,319	T2D	3049-61	Trypsin-1
PRSS3	P35030	*PRSS3*	9	33,750,464	CKD	3479-71	Trypsin-3
PTPRU	Q92729	*PTPRU*	1	29,563,028	T2D	8337-65	Receptor-type tyrosine-protein phosphatase U
QPCTL	Q9NXS2	*QPCTL*	19	46,195,741	T2D	8866-53	Glutaminyl-peptide cyclotransferase-like protein
RIDA	P52758	*RIDA*	8	99,114,567	T2D	14636-25	Ribonuclease UK114
SHBG	P04278	*SHBG*	17	7,517,382	CAD, IS, T2D	7909-37 4929-55	Sex hormone-binding globulin
ST3GAL2	Q16842	*ST3GAL2*	16	70,413,338	T2D	6281-51	CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 2
TIMP4	Q99727	*TIMP4*	3	12,194,568	T2D	6462-12	Metalloproteinase inhibitor 4
TP53I11	O14683	*TP53I11*	11	44,907,454	CAD	13022-20	Tumor protein p53-inducible protein 11
UST	Q9Y2C2	*UST*	6	149,068,063	CKD	8364-74	Uronyl 2-sulfotransferase
VWC2	Q2TAL6	*VWC2*	7	49,813,257	CKD	11121-56	Brorin
WFIKKN2	Q8TEU8	*WFIKKN2*	17	48,912,011	T2D	3235-50 13408-23	WAP, Kazal, immunoglobulin, Kunitz and NTR domain-containing protein 2

Extended Data Fig. 4 | Information about each PGS associated protein. Aptamer: Sequence ID for the SomaLogic aptamer(s) targeting the protein. A * next to the protein name indicates the aptamer(s) binds to specific isoforms of the listed protein or binds to multiple proteins; see Aptamer target column. Extended details on aptamer sensitivity and specificity can be found in Supplementary Table 2.

	Coronary artery disease
Disease association previously observed	APOE[71,72], CEI[73], GGT2[74,75], GRN[76], SHBG[77–79]
No reported association	CRYZL1, DUSP26, HBQ1, NPTX2, PCDBH10, TP53I11
	Chronic kidney disease
Disease association previously observed	B2M[27], C1GALT1C1[80], CST3[27], PDE4A[81], PDE4D[81], VWC2[82]
No reported association	FTMT, PRSS3, UST
	Ischaemic stroke
Disease association previously observed	SHBG[83]
	Type 2 diabetes
Disease association previously observed	ACY1[84,85], ADIPOQ[85,86], APOF[35], CCDC126[87], CFH[85], CFI[85], CHGB[88], CPM[35], CYB5R3[89], FAH[90], FBP1[28], GFRA1[35,85], GHR[35,91], GPD1[35], HS6ST2[35], IGFBP1[92], IGFBP2[35,42,43], INHBC[35], MSMP[35], PRCP[35,93], PRSS1[35], PTPRU[35], QPCTL[94], RIDA[35,95], SHBG[34], ST3GAL2[96], TIMP4[97], WFIKKN2[35,85]
No reported association	MUSK, ADH4, FAM20A

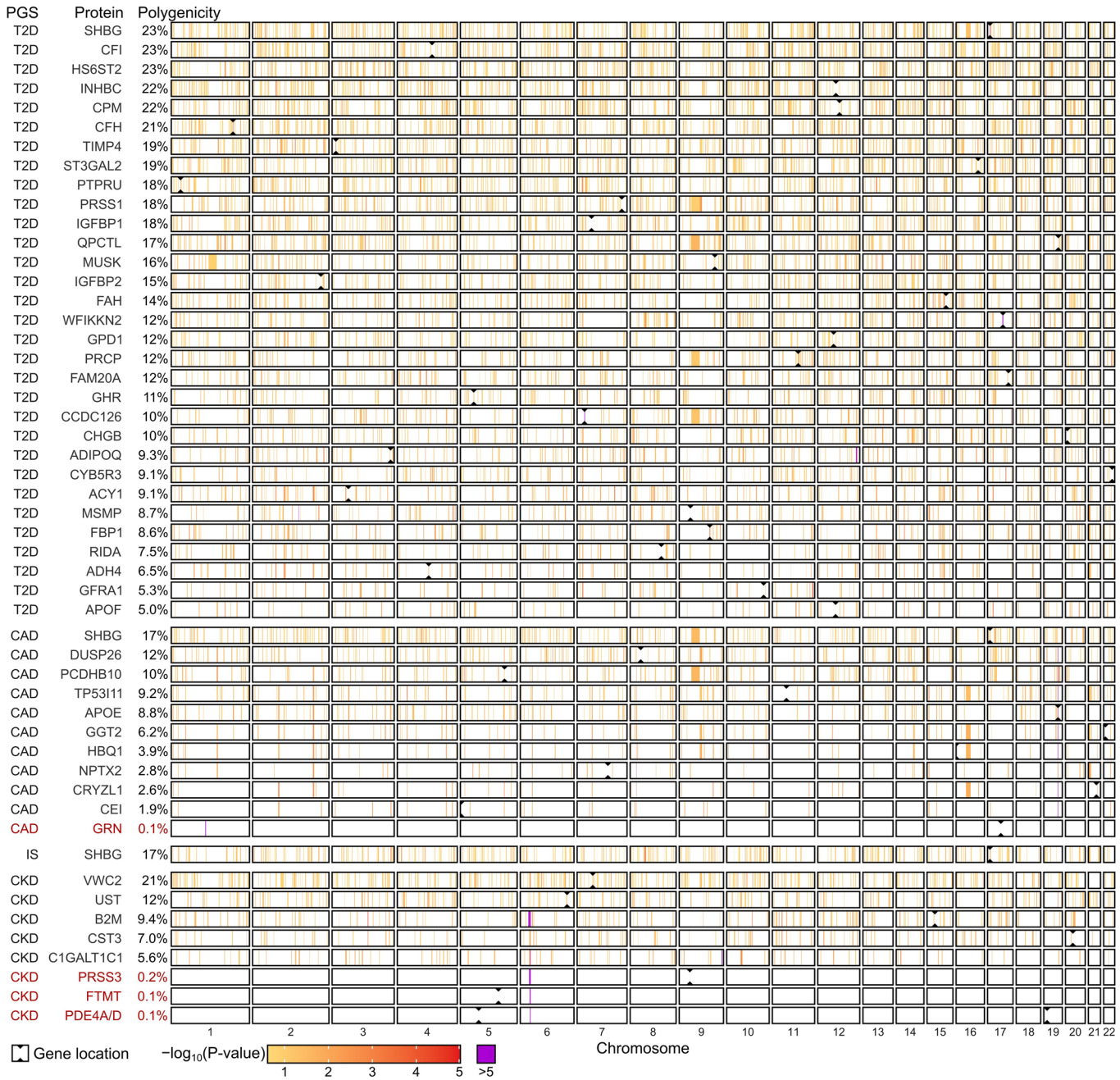
Extended Data Fig. 5 | Previous evidence for PGS-associated proteins in disease. Citations provided where association with the respective disease has been previously observed[71–97].



g

	Effect of T2D PGS through BMI				Effect independent of BMI		
Protein	Beta	95% CI	P-value	%	Beta	95% CI	P-value
GHR	0.041	[0.026, 0.056]	$2x10^{-7}$	59%	0.029	[-0.002, 0.059]	0.065
INHBC	0.038	[0.024, 0.053]	$2x10^{-7}$	58%	0.028	[-0.003, 0.060]	0.081
APOF	-0.035	[-0.048, -0.022]	$2x10^{-7}$	54%	-0.029	[-0.063, 0.004]	0.082
QPCTL	-0.035	[-0.049, -0.022]	$2x10^{-7}$	54%	-0.030	[-0.063, 0.003]	0.074
WFIKKN2	-0.039	[-0.054, -0.024]	$2x10^{-7}$	56%	-0.031	[-0.063, 0.001]	0.059
CYB5R3	-0.029	[-0.040, -0.018]	$4x10^{-7}$	46%	-0.034	[-0.067, -0.0005]	0.047

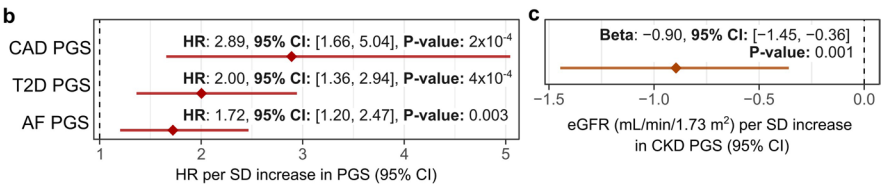
Extended Data Fig. 6 | Robustness of PGS to protein associations. a-c) Robustness and longitudinal stability of PGS to protein associations to proteomics technology. **d-e)** Robustness and longitudinal stability of protein levels to proteomics technology. **f)** Robustness of PGS-protein associations to environmental and physiological confounding. **g)** Mediation of PGS-protein associations through body mass index (BMI) for six proteins associated with T2D PGS. **a)** Compares PGS-protein associations from Fig. 1b in n = 3,087 INTERVAL participants in which protein levels were measured with the SomaLogic platform (x-axis) to PGS-protein associations tested in an independent set of n = 418 INTERVAL participants in which protein levels were measured with the Olink Explore platform (y-axis). In total 1,463 proteins were quantified by the Olink Explore platform, including 907 quantified by the SomaLogic platform, and among these 16 of the 49 PGS-associated proteins from Fig. 1b. **b)** Compares PGS-protein associations from Fig. 1b (x-axis) to PGS-protein associations tested in an independent set of n = 3,848 INTERVAL participants in which protein levels were measured with the Olink T96 platform (y-axis). In total 265 proteins were quantified by the Olink T96 platform, including 224 quantified by the SomaLogic platform, and among these 4 of the 49 PGS-associated proteins from Fig. 1b. **c)** Compares PGS-protein associations tested in n = 646 INTERVAL participants in which protein levels were measured with both the SomaLogic platform (x-axis) and, after two-years of follow-up, the Olink T96 platform (y-axis). **a-c)** Data shown correspond to the beta estimates from linear regression (points) and their 95% confidence interval (bars), indicating standard deviation change in protein levels per standard deviation increase in the respective PGS (denoted by colour). Solid points indicate two-sided P-value < 0.05 for the test on the y-axis. Linear regression on both axes were adjusted for age (at protein measurement), sex, 10 genotype PCs, and platform-specific technical covariates. Full summary statistics including exact P-values are detailed in Supplementary Data 3,b for linear regression tests on y-axes, and in Supplementary Data 3,a for linear regression tests on x-axes. **d)** Compares protein levels quantified by the SomaLogic platform (x-axes) to protein levels quantified by the Olink T96 platform (y-axes) after two years of follow-up in n = 646 INTERVAL participants. **e)** Compares protein levels quantified by the Olink T96 platform (x-axes) to protein levels quantified by the Olink Explore platform (y-axes) in n = 418 INTERVAL participants. **f)** Compares PGS-protein associations from Fig. 1b in n = 3,087 INTERVAL participants (x-axes) to PGS-protein associations (1) additionally adjusted for circadian effects (time of day of blood draw), (2) additionally adjusted for seasonal effects (date of blood draw), (3) when including 87 additional participants with prevalent cardiometabolic disease (n = 3,174 on y-axis), and (4) when adjusting for BMI (n = 3,072 participants with non-missing BMI on y-axis). All associations were testing using linear regression adjusting for age, sex, 10 genotype PCs, sample measurement batch, and time between blood draw and sample measurement in addition to the covariates noted above. Data shown correspond to the beta estimates from linear regression (points) and their 95% confidence interval (bars), indicating standard deviation change in protein levels per standard deviation increase in the respective PGS (denoted by colour). Full summary statistics including exact P-values in these sensitivity analyses are detailed in Supplementary Data 3,c. **g)** For the six proteins whose association with T2D PGS was attenuated by adjustment for BMI (P > 0.05; Extended Data Fig. 6f) gives, from mediation analysis, the estimated effect of T2D PGS on the protein levels through BMI (standard deviation change in protein levels through BMI per standard deviation increase in T2D PGS), percentage of T2D PGS to protein levels mediated by BMI, and the estimated effect of T2D PGS on protein levels independent of BMI in n = 3,072 INTERVAL participants. All P-values are two-sided.



Extended Data Fig. 7 | Polygenicity of PGS to protein associations. Linkage disequilibrium (LD) blocks contributing to each PGS to protein association in polygenicity tests. Briefly, each PGS was partitioned into 1,703 approximately independent LD blocks[54] then tested for association with each protein in linear regression adjusting for age, sex, 10 genotype PCs, sample measurement batch, and time between blood draw and sample processing in 3,087 INTERVAL participants. Full summary statistics including exact two-sided P-values for these tests are detailed in Supplementary Data 3,e. Next, to obtain the set of LD blocks contributing to each PGS to protein association, LD blocks were sequentially removed from the PGS in ascending order by association P-value (two-sided) until the association between resulting PGS and protein levels were attenuated (two-sided $P > 0.05$). Full summary statistics including exact two-sided P-values for these tests are detailed in Supplementary Data 3,f. The polygenicity of PGS to protein association (% of genome) shown on the left (and in Fig. 1c) was subsequently computed based on the sum of lengths of all contributing LD blocks (in base pairs) as a proportion of the genome. Here, associations ($-\log_{10}$ two-sided P-values) between protein levels and LD blocks contributing to the PGS to protein association are shown. Regions in white contain LD blocks that did not contribute to the PGS to protein association. PGS to protein associations listed in red are those explained by pQTLs (*cis* and/or *trans*) rather than polygenic.

a

Incident Disease	Events	Men	Age of onset
Atrial fibrillation	33	25	64.2 (59.2–69.8)
Type 2 diabetes	27	18	55.5 (47.7–63.3)
Coronary artery disease	15	12	57.2 (58.8–65.3)
Ischaemic stroke	3	1	73.1 (68.8–75.6)
Chronic kidney disease	0	–	–
Any of the above:	74	54	62.1 (53.5–67.7)

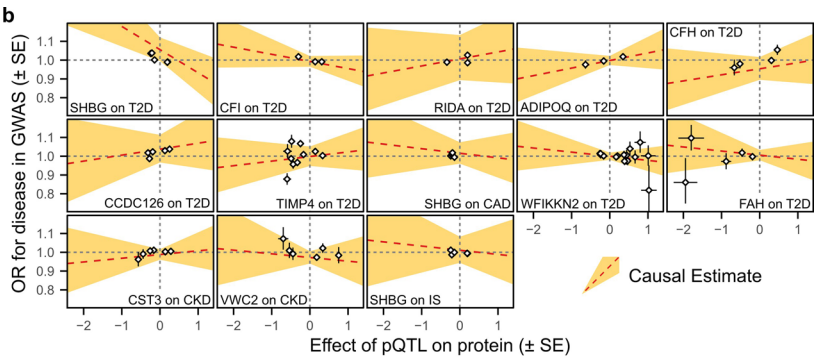


Extended Data Fig. 8 | Incident disease and PGS validity. a) Incident disease events over the 7.7 year of follow-up in the n = 3,087 INTERVAL participants. Endpoint: incident disease definition available in INTERVAL for the relevant PGS, as defined by CALIBER phenotyping algorithms. Age of onset: median age of first hospitalisation with the respective endpoint. Numbers in brackets gives the interquartile range. **b)** Hazard ratio (HR) (points) and 95% confidence interval (95% CI) (horizontal bar) for 7.7 year risk of hospitalisation with the respective endpoint per standard deviation increase in the respective PGS in cox proportional hazards models using follow-up as time scale and adjusting for age, sex, 10 genotype PCs, sample measurement batch, and time between blood draw and sample processing in n = 3,087 INTERVAL participants. P-values are two-sided. **c)** Association between CKD PGS with estimated glomerular filtration rate (eGFR), a marker of renal function used in chronic kidney disease diagnosis: decreased eGFR is indicative of reduced renal function[98]. EGFR was computed from serum creatinine in n = 3,307 participants using the CKD-EPI equation[99]. Association was fit with linear regression adjusting for age and sex, and 10 genotype PCs. The point corresponds to the change in eGFR per standard deviation increase in CKD PGS, and the horizontal bar corresponds to the 95% CI. P-values are two-sided.

a

Protein	Outcome	Causal Estimate OR	95% CI	P-value	Pleiotropy P-value
SHBG	T2D	0.88*	[0.83, 0.931]	$4x10^{-4}$	0.073
CFI	T2D	0.962	[0.921, 1.002]	0.061	0.63
RIDA	T2D	1.037*	[0.986, 1.086]	0.17	0.92
ADIPOQ	T2D	1.040	[0.983, 1.097]	0.18	0.86
CFH	T2D	1.032*	[0.982, 1.085]	0.21	0.15
CCDC126	T2D	1.032	[0.969, 1.11]	0.34	0.31
TIMP4	T2D	1.024	[0.982, 1.068]	0.34	0.95
SHBG	CAD	0.976*	[0.922, 1.034]	0.43	0.59
WFIKKN2	T2D	0.978*	[0.938, 1.025]	0.46	0.94
FAH	T2D	0.978	[0.921, 1.030]	0.53	0.69
CST3	CKD	1.020*	[0.957, 1.087]	0.55	0.40
VWC2	CKD	0.980	[0.908, 1.056]	0.60	0.10
SHBG	IS	0.978*	[0.912, 1.060]	0.72	0.76

b

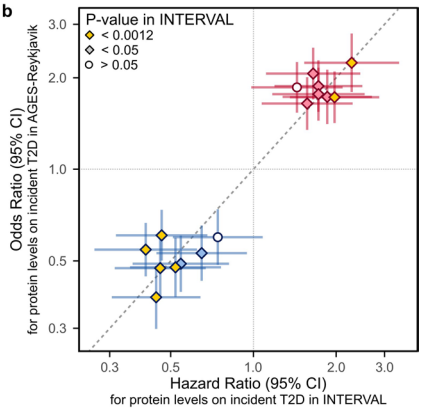


Extended Data Fig. 9 | Mendelian randomisation analysis. a) Causal effects of protein levels on disease risk estimated through two-sample Mendelian randomisation analysis of pQTL summary statistics and disease GWAS summary statistics. OR: consensus estimate of the odds ratio conferred per standard deviation increase in protein levels across five Mendelian randomisation methods. * Estimated causal effect is directionally consistent with PGS-protein associations in Fig. 1b. 95% CI: 95% confidence interval. P-value: Two-sided P-value obtained by averaging Z-scores across five Mendelian randomisation methods. Entries are greyed out where $P > 0.05$, and red where $P < 0.0038$ (Bonferroni correction for 13 tests). Pleiotropy P-value: two-sided P-value for the intercept term in Egger regression, indicating where $P < 0.05$ confounding of the causal estimate by horizontal pleiotropy. Full summary statistics including exact P-values are detailed in Supplementary Table 6. **b)** Dose response curves showing the estimated causal effect of changes in protein levels on disease risk for each protein and disease. Points on each plot show the *cis*-pQTLs used as genetic instruments for each test. On the x-axes, points show the standard deviation change in protein levels per copy of the minor allele in the pQTL summary statistics, and horizontal bars show +/- the standard error. On the y-axes, points show odds ratio conferred per copy of the minor allele in the GWAS summary statistics, and vertical bars indicate show +/- the standard error. Effect sizes, standard errors, and exact two-sided P-values from pQTL and GWAS summary statistics are detailed in Supplementary Table 7. The slope of the orange dashed line corresponds to the estimated causal effect (consensus Odds Ratio from **a**). The yellow ribbon shows the 95% confidence interval for the estimated causal effect (slope), accounting also for the 95% confidence interval for the intercept term in Egger regression.

a

AGES-Reykjavik / INTERVAL	Proteome-wide signficant for incident T2D	Proteome-wide signficant for prevalent T2D	
Associated with T2D PGS	16 ($P = 3x10^{-17}$)	23 ($P = 7x10^{-13}$)	31
P < 0.05 for incident T2D	14 ($P = 8x10^{-16}$)	21 ($P = 9x10^{-14}$)	25
P < 0.05 in causal mediation	11 ($P = 1x10^{-14}$)	13 ($P = 3x10^{-9}$)	15
	520	99	3,250

b



Extended Data Fig. 10 | Overlap of results with proteome-wide T2D associations in AGES-Reykjavik. a) Contingency table tabulating the overlap in results from our study detailed in Extended Data Fig. 3 (rows) with proteome-wide significant associations with incident and prevalent T2D in AGES-Reykjavik in Gudmundsdottir *et al*. 2020[35] (columns). One-sided P-values from Fisher's exact tests are given in each cell testing whether the overlap is greater than expected by chance. Row totals and column totals indicate the number of proteins in each row and column group, and the total overlap in proteins present in both studies (3,250) is given in the bottom right. **b)** For the 16 of 31 proteins nominally associated with T2D PGS in INTERVAL (Fig. 2b) and proteome-wide significant for incident T2D in AGES-Reykjavik, compares hazard ratios (points; x-axis) for incident T2D in INTERVAL (N = 27 cases over 7.7 years of follow-up in 3,087 participants) to odds ratios (points; y-axis) for incident T2D in AGES-Reykjavik (N = 112 cases after 5 years of follow-up in 2,940 participants). Cox proportional hazards models in INTERVAL were fit with follow-up as time scale, adjusting for age, sex, 10 genotype PCs, sample measurement batch, and time between blood draw and sample processing. Logistic regression in AGES-Reykjavik were fit adjusting for age and sex[35]. Horizontal and vertical bars correspond to the 95% confidence intervals of the hazard ratios and odds ratios respectively. Two-sided $P < 0.0012$ indicates association with incident T2D in INTERVAL from Fig. 2b was significant after Bonferroni correction for the 42 tested protein to disease associations. Summary statistics including exact two-sided P-values from both analyses are given in Supplementary Table 8.

nature portfolio

Corresponding author(s): Dr. Scott C. Ritchie
Dr. Michael Inouye

Last updated by author(s): Sep 6, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection: No software was used

Data analysis: Code used for data analysis are available on GitHub at https://github.com/sritchie73/cardiometabolic_prs_plasma_proteome/ which is permanently archived by Zenodo at doi: 10.5281/zenodo.4762747.

The following software and versions were used to run these scripts:

- Scientific Linux release 7.7 (Nitrogen) (HPC operating system)
- slurm version 19.05.5 (HPC queue manager and job submission system)
- GNU bash version 4.2.46(2) (shell environment used to run bash scripts)
- PLINK v1.90b6.10 64-bit (17 Jun 2019) (www.cog-genomics.org/plink/1.9/), aliased as plink1.9 in the scripts.
- PLINK v2.00a2LM AVX2 Intel (24 Jul 2019) (www.cog-genomics.org/plink/2.0/), aliased as plink2 in the scripts.
- R versions 3.6 and 4.0.3, along with R packages:
  - data.table version 1.12.8, 1.13.2
  - foreach version 1.4.4, 1.5.1
  - doMC version 1.3.5, 1.3.7
  - XML version 3.98-1.20, 3.99-0.5
  - biomaRt version 2.40.3, 2.46.0 (Bioconductor package)
  - openxlsx version 4.1.0.1, 4.2.3
  - ggplot2 version 3.3.0, 3.3.2
  - MendelianRandomization version 0.4.1, 0.5.0
  - ggrepel version 0.8.1, 0.8.2
  - ggrastr version 0.1.7, 0.2.1 (github package, https://github.com/VPetukhov/ggrastr)

- ggnewscale version 0.3.0, 0.4.3
- RColorBrewer version 1.1-2
- pheatmap version 1.0.12 (development version, https://github.com/raivokolde/pheatmap)
- impute version 1.57.0, 1.64.0 (Bioconductor package)
- WGCNA version 1.68, 1.69
- RNOmni version 0.7.1, 1.0.0
- cowplot version 1.0.0, 1.1.0
- lubridate version 1.7.9.2
- R.utils version 2.10.1
- powerMediation version 0.3.2
- scales version 1.1.1
- mma version 10.3.2
- survival version 3.2-7
- mediation version 4.5.0
- coloc version 3.2-1
- medflex version 0.6-7
- gridExtra version 2.3
- seriation version 1.2-9
- httr version 1.4.2
- jsonlite version 1.7.1
- xml2 version 1.3.2

- The BGEN software suite (https://www.well.ox.ac.uk/~gav/bgen_format/software.html) including:
  - bgenix version 1.1.4
  - qctool version 2.0.5, alised as qctool2 in the scripts
  - ldstore version 1.1
- SQLite version 3.30.1, aliased as sqlite3 in the scripts.

For R and R packages, version 3.6 was primarily used for the main pipeline and scripts run under src/pubs/cardiometabolic_proteins/ and src/pubs/cardiometabolic_proteins/review1/, while R version 4.0.3 was used for scripts run under src/pubs/cardiometabolic_proteins/review2/ and src/pubs/cardiometabolic_proteins/review3/. For R packages where two versions are listed, the first is the version used in R version 3.6 and the second is the version used in R 4.0.3.

Inkscape version 0.92.3 was used to layout and annotate figures from the figure components generated within the R scripts. Microsoft Office Professional Plus 2016 was used to draft the manuscript (Microsoft Word) and curate supplemental tables (Microsoft Excel) on Windows 10 Enterprise edition.

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

Data

Policy information about availability of data

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

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

All data used in this study is publicly available or deposited in a public repository. INTERVAL cohort data are available via the European Genotype-phenome Archive (EGA) with study accession EGAS00001002555 (https://www.ebi.ac.uk/ega/studies/EGAS00001002555). Dataset access is subject to approval by an independent Data Access Committee as they contain potentially identifying and sensitive patient information. Response times from the data access committee are typically within one week. All other data used in this study is publicly available without restriction. The PGS used in this study are available to download through the Polygenic Score Catalog (https://www.pgscatalog.org/) with accession numbers PGS000727 (atrial fibrillation), PGS000018 (coronary artery disease), PGS000728 (chronic kidney disease), PGS000039 (ischaemic stroke), and PGS000729 (type 2 diabetes). GWAS summary statistics used to generate new PGS for chronic kidney disease, type 2 diabetes, and atrial fibrillation in this study are available to download through the GWAS Catalog (https://www.ebi.ac.uk/gwas/) with study accessions GCST008065 (for chronic kidney disease GWAS published by Wuttke et al. in 201914), GCST007517 (for type 2 diabetes GWAS published by Mahajan et al. in 201815), GCST006414 (for atrial fibrillation GWAS published by Nielsen et al. in 201813). Additional GWAS summary statistics used for Mendelian randomisation analysis are also available through the GWAS Catalog with study accessions GCST004787 (for coronary artery disease GWAS published by Nelson et al. 201717), GCST006906 (for ischaemic stroke GWAS published by Malik et al. 201816) and GCST007518 (for type 2 diabetes GWAS adjusted for BMI published by Mahajan et al. 201815). Full pQTL summary statistics published by Sun et al. 201826 for all SomaLogic SOMAscan aptamers are available to download from https://www.phpc.cam.ac.uk/ceu/proteins/. The DrugBank database is publicly available to download at https://www.drugbank.ca/releases/latest. Summary statistics for all statistical tests are available in Supplementary Data 3 and additional cis-pQTLs mapped in this study are provided in Supplementary Data 4.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Unless otherwise noted, the sample size for all statistical tests was 3,087 independent samples (exceptions are in Extended Data Fig. 6). In all cases sample size was determined by the available data, after exclusion of samples with prevalent cardiometabolic disease. No sample size calculations were performed here.
Data exclusions	Samples with prevalent cardiometabolic disease were excluded to reduce potential for reverse causality. Sensitivity analysis to additional inclusion of these samples is performed in Extended Data Fig. 6f.
Replication	Replication of PGS to protein associations (Fig. 1) were assessed using independent technology (Olink proximity extension assays) in an independent set of INTERVAL participants (Extended Data Fig. 6a,b) for proteins quantifiable on the Olink Explore platform (16 proteins in 418 independent samples; Extended Data Fig. 6a) or on the Olink T96 platform (4 proteins in 3,848 independent samples; Extended Data Fig. 6b). In the 418 independent samples with 16/49 PGS-associated proteins measured on the Olink Explore platform we observed a strong correlation in effect sizes and effect size direction, but the number of samples was too small to reliably detect the associations at $P < 0.05$ (Extended Data Fig. 6a). In the 3,848 independent samples with 4/49 PGS-associated proteins measured on the Olink T96 platform 3 of 4 PGS to protein associations replicated (Extended Data Fig. 6b). Replication of protein to incident disease associations and mediation analyses were also performed in this independent set of 3,848 samples for the 4/49 PGS-associated proteins measured on the Olink T96 platform (Supplementary Table 5). Among the 3/4 proteins for which the PGS to protein associated replicated, both proteins with significant incident disease associations in the discovery cohort replicated, as did significant results in mediation analysis. Replication of protein to incident disease associations was additionally examined by intersecting our results with summary statistics from protein-wide association scan for T2D performed in the AGES-Reykjavik cohort by Gudmundsdottier et al. 2020 (Extended Data Fig. 8, Supplementary Table 8). Among the 31 proteins associated with T2D PGS, 23 were associated with prevalent T2D in Gudmundsdottier et al. 2020 at proteome-wide significance (Supplementary Table 8) all with consistent direction of effect with the PGS to protein associations. Among the 25/31 proteins significantly associated with incident T2D in our INTERVAL analyses, 14 were also associated with incident T2D in Gudmundsdottier et al. 2020 at proteome-wide significance, all with consistent direction of effects. All attempted replication experiments are reported here and in detail in the Main Text, Supplementary Notes, and listed Extended Data Figures and Supplementary Tables.
Randomization	The selection of INTERVAL participants for SomaLogic protein measurement was randomised, and performed in a previous study (Sun et al. 2018, ref. 26). For Olink protein measurements (Extended Data Fig. 6a-e), participants were selected to be over 50 years of age, and to have limited overlap with the group of participants with SomaLogic protein measurements, but otherwise selected at random. Selection of INTERVAL participants for Olink T96 measurements was done several years prior to the commencement of this study. For the Olink Explore platform, participants for measurement were selected at random from those with Olink T96 measurements. For serum creatinine, participants were selected at random for Metabolon measurement several years prior to the commencement of this study. All participants in these INTERVAL sub-cohorts contributed to the analyses, after exclusion of samples with withdrawn consent or without consent for electronic health record linkage (n=1 sample) and exclusion of samples with cardiometabolic disease prior to protein quantification (to prevent reverse causality, as noted above). All experiments either used the samples and sub-cohorts described here, or publicly available summary statistics from previous publications (e.g. GWAS and pQTL summary statistics).
Blinding	In all cases, investigators were blinded to group allocation and sample randomisation both during data collection and analysis. Group allocation and random sample selection was performed by an independent data management team prior to data collection and analysis.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	The n=3,087 INTERVAL participants were aged 18-75 with median age of 44.0 years (49% women) and median BMI of 25.5 (Extended Data Fig. 2). All participants had matched genotype, proteomics, and electronic health records. Participants with history of cardiometabolic disease prior to baseline assessment (Supplementary Table 1) were excluded from the analysis. Among these n=3,087 participants, over the 7.7 years of follow-up, 33 went on to develop atrial fibrillation (24% women), 27 to develop type 2 diabetes (33% women), 15 to develop coronary artery disease (20% women), 3 to develop ischaemic stroke (67% women), and 0 to develop chronic kidney disease (Extended Data Fig. 8a).
Recruitment	Participants were randomly recruited from regular blood donors from 25 blood donation centres from across the UKB as part of a clinical trial to study the safety of decreasing lenght of time between repeated blood donations. Blood donation eligibility criteria meant participants were healthy and had no history of major illness (such as cardiovascular disease or cancers) or recent acute illness (i.e. from infectious disease). Selection bias was also present as people who volunteer to donate blood and/or clinical trials bias towards healthy, and as such the study cohort is much healthier and younger than the general UK population. The impact of this on the study results is that the study participants are at low risk of the studied diseases, which is reflected in the rarity of incident disease events over the 7.7 years of follow-up, reducing power to detect true positive associations in all experiments.
Ethics oversight	Participants gave informed consent and this study was approved by the National Research Ethics Service (11/EE/0538)

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