Exome-wide association study in 54,698 south Asians identifies novel type 2 diabetes associations with *RNF19A*, *HNF4A*, and dissects role of coding variants in GP2 and *CDKAL1*

Sam Hodgson [1]*, Vi Bui [1]*, Margherita Bigossi[1], Daniel Stow[1], Cyrielle Maroteau[2], Alice Williamson[3,4], Siqi Hu[5], Alexandra M Blee[5], Adem Y. Dawed[2], Julia Carrasco-Zanini Sanchez[3], Viswanathan Baskar[6], Jebarani Saravanan[6], Benjamin M Jacobs[1], Georgios Kalantzis[7], Stuart Rison[1], Klaudia Walter[7], Erwan Pennarun[1], Genes & Health Research Team, Hilary Martin[7], Inês Barroso[8], Venkatesan Radha[6], Rajendra Pradeepa[6], Claudia Langenberg[3,4], V Mohan[6,9], Ranjit Mohan Anjana[6,9], David A van Heel[10], Amit R. Majithia[5], Yalda Jamshidi[2], Sarah Finer[1,11] Moneeza K Siddiqui [1,6]

*These authors contributed equally to this work

1. Wolfson Institute of Population Health, Queen Mary University of London, United Kingdom

2. Novo Nordisk Research Centre Oxford, Oxford, United Kingdom

3. Precision Healthcare University Research Institute, Queen Mary University of London, United Kingdom

4. Berlin Institute of Health at Charite, Universitätsmedizin Berlin, Germany

5. Division of Endocrinology & Metabolism, University of San Diego School of Medicine

6. Madras Diabetes Research Foundation (ICMR- Collaborating Centre of Excellence), Chennai, India

7. Wellcome Sanger Institute, Hinxton, United Kingdom

8. Exeter Centre of Excellence for Diabetes Research (EXCEED), University of Exeter Medical School, Exeter, United Kingdom

9. Dr. Mohan's Diabetes Specialities Centre (IDF Centre of Excellence in Diabetes Care), Chennai, India

10 Blizard Institute, Queen Mary University of London, United Kingdom

11. Barts Health NHS Trust, London, UK

Address for correspondence:
MK Siddiqui
Wolfson Institute of Population Health,
Queen Mary, University of London
email: moneeza.siddiqui@qmul.ac.uk

Abstract

Type 2 diabetes (T2D) disproportionately affects individuals of South Asian ancestry, yet the underlying genetic aetiology remains poorly understood and in genetic studies these populations remain underrepresented. We conducted an exome-wide association study (ExWAS) in 13,674 T2D cases and 41,024 controls of British-Bangladeshi and -Pakistani ancestry from the Genes & Health study, focusing on damaging coding variants in single-variant and gene-level analyses. This represents the largest ExWAS of T2D to date. We tested the association of novel genes and variants with metabolic traits in Genes & Health, UK, Madras Diabetes Research Foundation Biobank, India (MDRF), and the UK Biobank.

Gene-level analyses revealed novel associations with *RNF19A* ($p = 1.8 \times 10^{-1}$□) and *HNF4A* ($p = 2.4 \times 10^{-13}$), a known monogenic diabetes gene. The *HNF4A* signal was driven by a South Asian-specific gain-of-function variant (rs150776703, Pro437Ser) protective against T2D (OR = 0.48, $p = 2.8 \times 10^{-1}$□), diabetic eye disease, and gestational diabetes but associated with elevated LDL-cholesterol, plasma PCSK9 upregulation, and increased myocardial infarction risk (HR = 1.62, p = 0.01). In Huh7 cells, Pro437Ser demonstrated similar transactivation of HNF4A target, *G6PC*, compared to wild-type, partially explaining the protective effect for T2D. Single-variant analyses highlighted a missense variant in *CDKAL1* and a T2D risk-increasing South Asian-specific variant, rs78193826 (*GP2*: Val429Met), linked to CKD risk and beta-cell dysfunction (via stimulated C-peptides). Meta-analysis with UK Biobank highlights a role for plasma GP2 levels as a biomarker for high genetic risk of beta-cell deficiency and low risk for obesity-mediated T2D. These novel findings provide a foundation for drug target development, precision medicine and deeper exploration of metabolic and cardiovascular disease mechanisms in global populations disproportionately affected by T2D.

Introduction

Type 2 diabetes (T2D) is a common complex metabolic condition with significant heterogeneity both within and across ancestries[1–4]. Compared to individuals of European ancestry (EUR), people of South Asian ancestry (SAS) are 2-4 times more likely to develop T2D, at much leaner BMI, have more severe diabetes and progress more rapidly toward complications [5–7]. Applying recently-developed genetic methods to characterise heterogeneity in T2D aetiology has shown that insulin deficiency and unfavourable fat distribution as key drivers of early-onset T2D in South Asians and are associated with worse outcomes and higher rates of complications[8–10]. However, these methods have been informed by data from EUR and East Asian cohorts.

Unfortunately, while SAS populations contribute disproportionately to the global burden of T2D[11], and while they constitute over 25% of the global population they represent less than 0.4% of populations in genetic studies of T2D[1] . The under-representation of genomic data from SAS is particularly true of whole-exome sequencing data which has the potential to identify rare variants of functional importance and is particularly valuable in understudied populations. Indeed, this is observed in a recent exome-wide association study of 646 traits in 44,028 British Pakistanis and Bangladeshis[12]. In this population, with high rates of autozygosity, 100 novel gene-phenotype associations and 546 genes with never previously reported homozygous loss-of-function variants were discovered.

In contrast to the most recent multi-ancestry T2D case-control exome-wide association study in 45,231 individuals[13], our study presents a 10-fold increase in SAS sample size, and a 4.6 fold increase in T2D cases from this ancestry group. We validate and characterise our findings through replication in other SAS populations, multi-ancestry meta-analyses, *in vitro* functional assays, and the use of proteomic data.

Results

Participants

We included 54,698 individuals (13,674 electronic health record-defined cases and 41,024 controls) from Genes & Health[14] in an exome-wide association study **(Figure 1)**. Descriptive details of included participants are given in **Table 1,** including event rates for diabetes-related complications, stratified by sex and genetically-inferred ancestry.

Figure 1. ***Participant flow diagram*** *to show numbers of individuals included in Exome-wide association study (ExWAS) and genome-wide association study (GWAS) of type 2 diabetes (T2D) in the Genes & Health study, and numbers of T2D cases and controls available for replication of effects and deep phenotyping in the Madras Diabetes Research Foundation biobank.*

Gene-level testing recapitulates findings in *MAP3K15*, identifies a novel association in *RNF19A*, and a novel ancestry-specific single variant-driven protective association in HNF4A for type 2 diabetes

We performed association testing of type 2 diabetes cases versus controls across 18,956 genes by aggregating rare variants with minor allele frequency (MAF) < 1% into functional categories (i.e. masks, as previously described[12] and in **Table S1**) and testing the association of these aggregates of variants with type 2 diabetes. Association testing results for three gene/mask combinations (*MAP3K15*: pLoF-HC|pDM, *RNF19A*:pLoF-HC and *HNF4A*: pLoF-HC|pDM; variants constituting each mask given in **Table S2**) crossed a Bonferroni-defined significance threshold of 9.35 x 10^{-8} after correcting for 534,135 tests (**Figs 2A,2B, S1)**. Rare *MAP3K15* variants have previously been reported to be negatively associated with T2D[15]; *HNF4A* variants have been reported to be risk-increasing for T2D[16,17]; and neither *RNF19A* rare variants, nor those in nearby genes (such as *SPAG1*), have previously been reported to be associated with T2D at a level crossing exome-wide significance[18].

Figure 2. ***Exome-wide association study of type 2 diabetes cases and controls in Genes & Health****. A: Symmetrical Manhattan plot showing association (-LOG10(p value)) between genetic variants and type 2 diabetes. Gene-burden results are shown in the top panel, and single variant results in the bottom panel. B: Chord diagram showing associations between gene-burden results and phenotypic traits in Genes & Health; these are shown agnostic of direction of effect as not all significance-passing mask-trait associations were associated with a beta, depending on the statistical test used (eg REGENIE GENE-P). Fasting glucose, non-fasting glucose and HbA1c were grouped under the heading "Glycaemic traits" . C: Meta-analysed associations between Genes & Health (GnH) variants passing suggestive threshold and multi-ancestry meta-analysed data from the type 2 diabetes knowledge portal. D: Heatmap showing association between single variants identified in exome-wide association study and diabetes-related traits. Betas and p values (represented by colour scale for beta, as per legend; and * (<0.05), ** (<0.01)*

*and *** (<0.001) for meta-analysed results between Genes & Health, MDRF, and T2D Knowledge portal are presented. These are presented separately as forest plots in fig S3.*

Deleterious, low frequency MAP3K15 variants are associated with lower risk of T2D and its complications

We recapitulate previously reported negative association[15] between *MAP3K15* variants and type 2 diabetes (**Fig 3A)**. The strongest associations were for the mask capturing deleterious missense variants with MAF < 0.01 (meta-analysed OR = 0.85, 95% CI = 0.80-0.90, p =1.7 x 10^{-8}), with no specific variant driving this effect in leave-one-variant-out (LOVO) analysis (**Fig 3B)**. In phenome-wide association testing within Genes & Health and meta-analysed with UK Biobank, this mask was significantly associated with lower lifetime median HbA1c (beta = -0.09, 95% CI = -0.12 - -0.07, p = 4.1e-14) lower non-fasting glucose (beta = -0.12, 95% CI = -0.14 - -0.08, p = 2.5 x 10^{-15}), and reduced risk of gestational diabetes (OR = 0.74, 95% CI = 0.63 – 0.86, p = 1.2 x 10^{-4}). A nominally significant association with cerebrovascular events in Genes & Health did not persist after meta-analysis (**Fig 3A).**

*Figure 3. **Summary of gene-based results for MAP3K15 and RNF19A**. A/C: Forest plots showing associations of MAP3K15 and RNF19A gene-based testing results with selected phenotypic traits, meta-analysed with UKBiobank data from all available ancestries published on the AZphewas portal for corresponding masks (ptv – protein-truncating variant ~ pLoF; dmg / raredmg ~ pDM);.SKAT results from AZPhewas portal were used; the most significant mask with a direction of effect was used for G&H. Because only results with p < 0.10 are available on this resource, not all ancestries are included in each meta-analysis. B/D: For MAP3K15 / RNF19A, lollipop plots showing genomic position relative to ENSEMBL canonical exonic regions (dark blue rectangles), overlaying single variant association testing LOG10P values for coding variants (top panel) and p values from leave-one-variant-out (LOVO) analysis using the most significant mask in gene-burden testing MAF < 0.01 in the bottom panel. In this bottom panel, the LOG10P value for the association of T2D with the mask without the named variant is plotted. Horizontal dotted lines represent exome-wide significance.*

High-confidence loss-of-function RNF19A rare variants drive higher risk of T2D and cardiovascular complications

We identified a novel, exome-wide significant association between HC-pLoF variants with MAF < 0.01 in *RNF19A:pLoF-HC* and type 2 diabetes (GENE_P p value = 1.8x10^{-14}; meta-analysed OR (incorporating G&H SKAT test results with direction of effect) = 1.65, 95% CI = 1.37 – 2.01, p = 2.8 x 10^{-7}). In Genes & Health, this mask was also nominally associated with HbA1c, BMI, HDL-C, and macrovascular diabetic complications including heart failure, peripheral arterial disease, myocardial infarction and transient ischaemic attack (**Fig2B,3C, Table S3)**. In LOVO analysis, the strongest individual signal arose from a risk-increasing stop-gain mutation (8:100288165:G:A, *RNF19A* Gln4Ter, MAF = 9.1x 10^{-6}, beta = 5.36, p = 0.001,) resulting in truncation of the protein at the fourth amino acid **(Fig 3D)**.

The effects of this SAS-specific variant could not be meta-analysed in UK-Biobank because no carriers could be identified. However, meta-analysis of *RNF19A* pLoF-HC variants from UKBiobank obtained via the AZPhewas[19] portal with Genes & Health identified significant associations with increased risk of ischaemic heart disease (OR = 2.38, 95% CI = 1.58 – 3.58, $p = 3.2 \times 10^{-5}$) and peripheral arterial disease (OR = 3.04, 95% CI = 1.70 – 5.43, $p = 1.8 \times 10^{-4}$), and nominal associations with cerebrovascular disease (OR = 1.89, 95% CI = 1.22 – 2.93, $p = 4.7 \times 10^{-3}$) and heart failure (OR = 1.09, 95% CI = 1.01 – 1.18, p = 0.03)(**Fig 3C)**. Significant continuous trait associations including HbA1c (beta = 0.27, 95% CI = 0.10 – 0.43 $p = 1.7 \times 10^{-3}$), systolic blood pressure (beta = 0.21, 95% CI = 0.10 – 0.32, $p = 2.4 \times 10^{-4}$), and HDL-cholesterol (beta = -0.59, 95% CI = -0.90 - -0.28, $p = 1.97 \times 10^{-4}$) were also observed after meta-analyses with UKBiobank **(Fig 3C).** A nominally significant association with increased BMI in Genes & Health (beta = 1.83, 95% CI = 0.52 – 3.15, $p = 6.4 \times 10^{-3}$), did not persist after meta-analysis. Results from the T2D knowledge portal[18] highlighted a nominally significant association with increased waist circumference adjusted for BMI (beta = 0.005, 95% CI = 0.0008 – 0.01, p = 0.04), largely driven by association results from European-ancestry individuals in UK Biobank[20].

HNF4A variant identified via gene-based testing protects against T2D, but increases risk of CVD

We observed a protective association between *HNF4A:*pLoF-HC|pDM and T2D risk ***(*Fig 2A/B, S1,S2A)***. LOF mutations in *HNF4A* are a known cause of monogenic diabetes (maturity onset diabetes of the young MODY)[16,21] and gene burden tests with masks containing low frequency, deleterious variants have been shown to increase T2D risk.[18] We undertook LOVO to understand why we observed a risk-decreasing effect from a mask containing predicted deleterious missense variants (MAF < 0.01) in *HNF4A* (OR = 0.56, 95% CI = 0.48 – 0.67, p = 1.2 x 10^{-11}) (**Fig 4A)**.

Fig 4: ***An ancestry-specific variant 20:44429549:C:T in HNF4A (Pro437Ser) is protective against type 2 diabetes but associated with increased cholesterol and risk of myocardial infarction****. A: Lollipop plot showing Chr20 genomic position relative to ENSEMBL canonical HNF4A exonic regions (dark blue rectangles), overlaying single variant association testing LOG10P values for coding variants (top panel) and p values from leave-one-variant-out (LOVO) analysis using the most significant mask in gene-burden testing (MASK B (deleterious missense variants), MAF < 0.01 in the bottom panel. In this bottom panel, the LOG10P value for the mask without the named variant is plotted. Horizontal dotted lines represent exome-wide significance. B: Allele frequency of HNF4A Pro437Ser across ancestral groups; data are presented from 1000 Genomes. C: Box plots showing the distribution of lifetime median total cholesterol, HbA1c, non-fasting glucose, and age at type 2 diabetes diagnosis, stratified by number of HNF4A Pro437Ser T alleles. Overlaying these are summary fits from multivariable regression models depicting the direction of association with 95% CI between number of T alleles and each trait. D/E: Unadjusted Kaplan-Meier plots showing disease-free survival from birth, stratified by HNF4A Pro437Ser Genotype, for type 2 diabetes (D) and myocardial infarction (E). F: Association of Pro437Ser with incident myocardial infarction, presented from multivariable cox proportional hazards models. The base model is adjusted for age, sex, ancestry, and the first 20 genetic principal components. Subsequent LDL- and statin-adjusted models are adjusted for lifetime median LDL, and lifetime (binary) statin exposure, respectively. Only individuals with at least one LDL value and available prescribing data were included in analyses. E: Proteomic signatures associated with HNF4A Pro437Ser across two proteomic platforms (Olink, SomaLScan) in Genes & Health. Plotted result is the association testing output from REGENIE single variant testing reporting per-T-allele association with protein expression. H: Western blot of Huh7 cells overexpressing wild-type (WT), Pro437Ser (p.P437S), or Arg89Trp (p.R89W) HNF4A, or an empty vector (control). Two replicates shown. GAPDH included as loading control. I: Relative expression of HNF4A target gene G6PC in Huh7 cells from panel H, with GAPDH as housekeeping gene.*

LOVO sensitivity analysis revealed that the protective association observed for T2D was driven by a single missense variant (rs150776703, *HNF4A* Pro437Ser) (**Fig 4A).** In single variant association testing, this variant was strongly associated with lower risk of T2D (OR = 0.48, 95% CI = 0.40-0.59, p = 2.8 x 10^{-14}). rs150776703 had low frequency in South Asians (MAF =0.008) and West Asians (MAF = 0.0003) and was extremely rare (MAF < 0.000001) or not observed in all other ancestries (**Fig 4B).** It was slightly more common in British Bangladeshis than Pakistanis in Genes &

Health (MAF 0.0098 vs 0.0063)**.** Notably it was not included in any UKBiobank gene-burden sets on the AZ Phewas portal[19] and also not represented in publicly available gene-level testing[20] since they are from EUR cohorts. This variant is not linked to nearby variants, including those with significant T2D associations (maximum R^2 = 0.002, **Fig S2B-E).** Formal fine-mapping in both imputed genotype and whole exome sequencing data supported the observation that rs150776703 was a single, causal variant with posterior inclusion probability (PIP) = 0.997 in imputed genotype data and PIP = 1 in whole exome sequencing data (**Fig S1B-E, Table S5).**

Compared to a prevalence of 24.9% across the entire cohort, none of the 13 homozygous carriers of the effect had a diagnosis of T2D, **(Fig 4C)**. In Genes & Health, Pro437Ser was significantly associated with lower lifetime median HbA1c (beta = -2.23, 95% CI = -3.10 - -1.37, p = 4.02×10^{-7}), lower non-fasting glucose (beta = -0.44, 95% CI = -0.63 - -0.26, p = 2.4×10^{-6}), and reduced risk of diabetic eye disease (beta = -0.04, p = 3.0×10^{-4}). It was also nominally associated with lower 120-minute oral glucose tolerance test (beta = -2.35, 95% CI = -4.12 - -0.58, p = 0.009) and later onset of type 2 diabetes (ANOVA p = 0.01) (**Fig 4C/D, Table S4)**. However, Pro437Ser is associated with higher lifetime total cholesterol (beta = 0.21, 95% CI = 0.12 – 0.32, p = 1.3×10^{-5}) and nominally associated with non-HDL cholesterol (beta = 0.09,95% CI = 0.02 – 0.17, p = 0.009)(**Fig 4C**,**Table S4).** At the time of type 2 diabetes diagnosis in Genes & Health, this variant was associated with both increased BMI (beta = 1.96, 95% CI = 0.65 – 3.26, p = 0.003) and waist circumference (beta = 0.07, 95% CI = 0.02 – 0.11, p = 0.005)(**Table S4)**. These anthropometric findings replicated in MDRF (BMI beta = 0.88, p = 0.02; waist circumference beta = 1.99, p = 0.03), but none passed correction for multiple testing.

Using longitudinal healthcare data from Genes & Health in multivariable survival analyses. *HNF4A* Pro437Ser was nominally associated with increased hazards of myocardial infarction (HR = 1.62, 95% CI 1.09-2.44, p = 0.01) (**Fig 4E/F)**; this effect was not attenuated by adjustment for lifetime median LDL-C, , but was attenuated by adjustment for lifetime statin exposure (HR = 1.49, 95% CI 0.94 – 2.39, p = 0.08) in the same sample of individuals.

We explored the association of *HNF4A* Pro437Ser with a focused selection of known and putative downstream gene targets(*HNF4A, PCSK9 and ANGPTL3*) using affinity-based broad-capture proteomic signatures in Genes & Health; specifically, Olink Explore HT (targeting 5416 unique proteins in 1434 individuals) and SomaScan 11K v5 (targeting 9964 proteins in 1540 individuals) platforms **(Fig 4E).** For *HNF4A* Pro437Ser we observed increased *cis*-protein plasma abundance measured by Olink (beta = 0.54, 95% CI: 0.03–1.05, p = 0.035) but not SomaScan (beta = 0.19, 95% CI: –0.28–0.66, p = 0.41). Pro437Ser was also associated with increased abundance of PCSK9, an atherogenic protein which is upregulated by *HNF4A*[22] (SomaScan beta = 0.55, 95% CI: 0.06–1.04, p = 0.026), supported by a non-significant trend using the Olink platform (beta = 0.46, 95% CI: –0.07–0.99, p = 0.08). These findings suggest this variant may confer gain (as opposed to loss) of

function in terms of detectable plasma protein abundance. Supporting this, presumed loss-of-function protein-truncating variants in *HNF4A* are associated with higher risk of T2D but lower LDL-C in European-ancestry individuals in UKBiobank (**Fig 2B,S1)** (ie, the opposite direction to the effects we report for the putative gain of function *HNF4A* Pro437Ser).

However, *HNF4A Pro437Ser* was not associated with detectable expression levels of ANGPTL3, which *HNF4A* also upregulates, arguing against this hypothesis, although the direction of effect for both platforms weakly suggested a trend towards a positive signal(**Fig 4G)**. The lack of association for *ANGPTL3* could not be explained in terms of lower abundance compared to *HNF4A* or *PCSK9* (ANOVA p <2e-16 across all 3 proteins for raw abundance; *ANGPTL3* not least abundant on either platform **(Fig S2F);** and similar dilution factor required (a proxy of abundance) in somalogic assays (**Fig S2G)).** The lack of association could also not be explained by lack of detectability compared to *HNF4A* or *PCSK9* in olink assays (call rate for *ANGPTL3* greater than both *HNF4A* and *PCSK9*, **Fig S2G).**

Hepatocyte nuclear factor 4 alpha is a transcription factor that regulates expression of several genes and its main cites of expression are the liver, followed by colon, kidney, and pancreas. To determine the tissue of origin for detectable plasma protein levels of *HNF4A, PCSK9* and *ANGPTL3*, we used a recently constructed human-proteome distribution atlas from 29 tissue and cell types and developed a Global Label Score (GLS) [ranging from 1-4], where a higher score denotes greater confidence for inferring the tissue of origin for plasma proteins[23]. Scores ranged from 2.45, 2 and 2 respectively, suggesting a high degree of uncertainty, that observed effects are due to hepatic activity, and implying we must interpret these findings with caution.

To evaluate the transactivation potential of Pro437Ser *HNF4A* in hepatocytes, *HNF4A* wild-type (WT), Pro437Ser, or the known MODY associated variant Arg89Trp[24] were transfected into Huh7 cells. Western blotting revealed robust overexpression of *HNF4A* for each variant, compared to minimal detection of endogenous HNF4A in the control cells (**Fig 4H**). WT *HNF4A* induced expression of a known hepatic HNF4A transcriptional target, G6PC, by ~28-fold (**Fig 4I**). HNF4A p.P437S exhibited a similar degree of activity to WT while the MODY associated p.R89W variant, which inhibits nuclear localization[24], displayed the lowest transactivation (**Fig 4I**).

Overall, we find that *HNF4A* Pro437Ser is a South-Asian specific gain-of-function variant which maintains WT levels of activity on a known transcriptional target, *G6PC*. While the variant may be associated with increased risk of coronary artery disease via higher cholesterol levels, we also observed effects on improved glycaemic control and lower T2D risk. Upregulation of G6PC may contribute to these effects, although further analyses of the effects of Pro437Ser on HNF4A

transactivation and a more comprehensive analysis of downstream targets is warranted.

Single variant exome-wide association study (ExWAS) identifies coding variants in known genes and new ancestry-specific variants highlighting novel loci and genes implicated in type 2 diabetes

We performed single variant exome-wide association testing of type 2 diabetes status on 13,674 cases and 41,024 controls using REGENIE, including only 25,236 missense coding variants with MAF > 0.01 (aligning with masks included in gene-based testing, **Table S1)**.

We identified one variant passing exome-wide significance (chr8:117172544:C:T, (rs13266634), a missense variant in *SLC30A8* which has been previously identified in imputed genotype data and known to be strongly associated with T2D (previously reported OR = 0.91, p = 8.6x 10^{-25018}). We additionally identified suggestive signals (P < 5e-5) mapping to four genes: *GP2, CDKAL1, CCNL2, and WFS1,* which we assessed using meta-analyses (**Fig 2C)***.* After meta-analysis with existing associations obtained from the T2DKP, we observed stronger associations for *GP2 Val429Met (*OR = 1.10, 95% CI 1.08 – 1.11, p = 2.04 x 10^{-37}*), CDKAL1 Pro409Arg (*OR = 1.02, 95% CI 1.02 – 1.03, p = 1.88x10x^{-12}*) and WFS1 Val333Ile (*OR = 1.063, 95% CI = 1.059 – 1.067, p = 6.5X10^{-232}*),* and therefore focused on downstream analyses on these coding variants with external evidence for replication. Further summary information for these variants is shown in **Table 2.**

Ancestry-specific variant in pancreatic acinar cell-expressed *GP2* is associated with T2D via a beta-cell deficiency pathway

We identified a missense variant rs78193826 C-T (Val429Met), located in the *GP2* gene region, which increases T2D risk (meta-analysed OR = 1.10, 95% CI 1.08 – 1.11, p = 2.04 x 10^{-37}). In spite of other larger multi-ancestry studies, the variant has only been reported previously in individuals of East Asian ancestry from the Asian Genetic Epidemiology Network consortium[25] (OR = 1.09, p = 1.43 x 10^{-8}, n= 254,373). This is because *GP2* Val429, has the highest MAF in East Asians (EAS 0.07), followed by South Asians (0.03; MAF in Genes & Health Bangladeshis 0.047 vs Pakistanis 0.025) (**Fig 5A),** and is rare or ultra-rare in other populations.

Fig 5. ***Summary data for variant rs78193826, GP2 Val429Met****. A: Allele frequencies across ancestries for rs78193826, taken from 1000 Genomes estimates, showing specificity of the variant to East and South Asian populations. B: Multi-ancestry forest plots for selected trait associations of GP2:Val429Met, meta-analysing results from Genes & Health, the Madras Diabetes Research Foundation, and ancestry-stratified publicly available estimates from the Type 2 diabetes knowledge portal. C: Proteomic signatures associated with GP2 Val429Met across two proteomic platforms (Olink, SomaLScan) in Genes & Health. D: Meta-analysis between Genes & Health and the UK Biobank showing the association of subtype specifc genetic risk scores (partitioned polygenic risk scores, pPS) and plasma levels of GP2.*

We then meta-analysed the effect of the variant on glycaemic traits, using data from G&H, MDRF and EAS cohorts in T2DKP[25,26], and insulin response from the MAGIC consortium[27]. Val429Met was significantly associated with higher fasting glucose (beta = 0.03, 95% CI 0.02 – 0.04, p = 7.4e-11) and nominally with lower 2hr C-peptide levels (beta = -0.08, 95% CI = -0.13 - -0.03, p = 0.002, n=14,704)(**Fig 5B, Fig S3, Table S4).** In contrast, the variant was not associated with any measures of insulin resistance, e.g. Triglycerides: HDL-C ratio (p = 0.23), and was only nominally associated with lower HDL-C (beta = -0.009, 95% CI = -0.015 - -0.002, p = 0.008) and BMI (beta = -0.009, 95% CI = -0.016 - -0.002, p = 0.02). *GP2* variants have additionally previously been reported to be associated with use of diabetes-controlling medication in a study from Biobank Japan[28].

Data from GTEx[29], Translational Human Pancreatic Islet Genotype Tissue-Expression Resource (TIGER)[30], and Pankbase[31] shows *GP2* is almost exclusively expressed in the pancreas, minimally in islet cells, and more specifically, in acinar cells (**Figure S4A/B).**

Our proteomic analyses show an increase in plasma abundance of the GP2 *cis*-protein on the Olink platform (beta = 0.29, SE = 0.09, p = 0.001) but not SomaScan (beta = 0.05, SE = 0.10, p = 0.63, **Fig 5C**). Using the human-proteome distribution atlas[23], we determined that plasma protein GP2 has been demonstrated to be secreted from the pancreas, with a high GLS of 3 **(Fig S4C),** giving us confidence in interpreting our proteomic findings.

To further understand our findings in the context of the complex, heterogeneous aetiology of T2D, we investigated the association between genetic sub-type specific partitioned polygenic scores (pPS)[9] and plasma GP2 levels in people of European ancestry in UK Biobank and in the G&H study (Olink), irrespective of T2D status. Our meta-analysis revealed a strong association between a higher genetic burden of beta cell dysfunction and increased plasma levels of GP2 [beta cell 1 (insulin sensing): beta = 0.02, SE = 0.004, p = 1.74×10^{-7}; beta cell 2 (insulin production) beta = 0.02, SE = 0.004, p = 2.35×10^{-8}] (**Figure 5D / S4D).** Additionally, a higher genetic burden related to BMI- and cholesterol-mediated T2D was associated with lower GP2 levels. A strong association between a less-well understood insulin resistance pathway characterised by lower ALP and cholesterol levels was also associated with higher plasma GP2 (beta = 0.37, SE = 0.08, p = 7.2×10^{-7}), although this pPS is comprised of just 9 variants, and the strong association was driven by a single SNP in the *ABO* gene (rs529565) which is well known to associate with ALP (**Fig S4D)**.

We show strong direct and orthogonal evidence that GP2 Val4229Met is associated with beta-cell dysfunction and lean-onset T2D, with a possible role for higher plasma abundance of GP2 in this pathway.

A coding missense variant (*CDKAL1* Pro409Arg, rs77152992 C-T) may be associated with T2D independently of previously reported intronic variants in *CDKAL1*

To date, reported T2D-associated variants mapped to the *CDKAL1* gene, a well-established risk locus, have been intronic[32–34]. Here we highlight a coding missense variant (Pro409Arg (rs77152992) which, in addition to its association with T2D (meta-analysed OR = 1.02, 95% CI = 1.02 – 1.03, p = 1.88 $x10^{-12}$), is significantly associated with fasting glucose (beta = 0.012, 95% CI = 0.006 – 0.018, p = 7.7 $x10^{-5}$), HbA1c (beta = 0.002, 95% CI = 0.001 – 0.004, p = 0.001), progression to insulin therapy and microvascular complications (HR = 1.17 and HR = 1.10, p = 1.0 x 10^{-4} and p = 0.001, respectively), and earlier onset T2D (beta = -0.47, p = 0.001)(**Fig S3, Table S4)**. In MDRF, this variant was significantly associated with lower BMI (beta = -0.22, p = 8.1$x10^{-4}$), and nominally associated with higher 1h-oral glucose response (beta = 13.7, p = 0.004), lower HOMA-IR and HOMA-B (beta = -0.28 and -0.004, p = 0.006 and 0.04, respectively), lower fasting and stimulated C-peptide (beta = -0.02 and -0.04, p = 0.01 and 0.04, respectively). Burden testing in UK Biobank, demonstrates that rare-damaging variants (not including Pro409Arg) in EUR are associated with earlier onset diabetes (B=-0.57, p = 6.2$x10^{-4}$) .These associations support a role for Pro409Arg in beta-cell function, and perhaps more specifically in glucose sensing.

We explored the association and interaction of this variant with previously reported intronic variants through fine-mapping with SuSIE[35] using genome-wide imputed (GWAS) data on the same individuals **(Fig 6A-D)**. We observed only weak linkage disequilibrium (LD) between *CDKAL1* Pro409Arg and the intronic lead SNP (rs67131976, association with T2D in GWAS OR = 1.12, p = 3.3$x10^{-20}$, previously reported association on T2DKP OR = 1.14, p = 5$x10^{-324}$)[18]; R^2 = 0.10 in both Genes & Health and 1000 Genomes South Asian population. We also observed differential patterns of minor allele frequency between these two variants (e.g. rs67131976 vs *CDKAL1* Pro409Arg MAF in SAS 0.24 and 0.11 respectively, **Fig 6E)**. Pro409Arg has been reported previously in multi-ancestry analyses by Suzuki et al[1] (OR = 1.05, 95% CI 1.02 – 1.04, p = 3.2$x10^{-9}$), and displayed heterogeneity across ancestries (MR-MEGA P-heterogeneity = 0.0026), although this result was not fine-mapped.

Figure 6. ***Type 2 diabetes association with genetic variants mapping to CDKAL1, focusing on rs77152992 (CDKAL1 Pro409Arg.)****Locus zoom (panel A + C) and posterior inclusion probability finemap plots (panel B + D) for the CDKAL1 region including variant rs77152992 (chr6_21065218_C_T) on chromosome 6, performed separately in whole exome sequencing (plots A+B) and TOPMED-r3- imputed genotype data *(INFO>0.7)(plots C+D). Panel E: Allele frequencies across ancestries for rs77152992, taken from 1000 Genomes estimates. Panel F: association of rs77152992 missense variant with T2D in Genes & Health in 3673 individuals homozygous for previously reported intronic variant rs67131976(T/T), which is a well-established risk locus strongly associated with T2D (previously reported p value = 3.2e-324). The presented result is the beta from a multivariable regression model adjusted for age, age squared, sex, ancestry, and the first 20 principal components. The left panel of the graph shows results*

from a continuous per-effect-allele model; the second from a discrete model versus C/C individuals. Panel G: as for panel F, but showing association of rs77152992 missense variant with fasting C-peptide, the most strongly associated trait, in MDRF, in 1079 individuals homozygous for previously reported intronic variant rs67131976(T/T). The presented result is the beta from a multivariable regression model adjusted for age, age squared, sex, ancestry, and the first 5 principal components.

In GWAS, the sole 95% credible set contained 16 variants, all of which were intronic, with a maximum PIP of 0.16 (for rs67131976); all 16 of these variants were in strong LD (>0.94) with rs67131976, while *CDKAL1* Pro409Arg was not included in any credible set and had PIP = 1e-16 (**Fig 6C/D, Table S5**). In contrast, fine-mapping in whole exome sequence data identified a single variant (Pro409Arg) with a suggestive PIP (PIP = 0.94)**(Fig 6A/B)**.

In an orthogonal approach to assess for independent effects, we explored whether additional effect alleles (T) of the missense variant CDKAL1 Pro409Arg were associated with T2D, stratified by intronic lead SNP rs67131976 (C-T) genotype **(Fig *6F*),** reasoning that if effects were not independent, there would be no additional risk conferred by the missense coding variant in homozygote carriers of the intronic variant**.** In 3,673 homozygous carriers of the intronic lead variant rs77152992-T /T in Genes & Health, the association between T2D status and missense effect allele count (CDKAL1 Pro409Arg – T allele) was significant after adjustment for age, sex, and principal components (beta = 0.021, p = 0.04), an effect driven by heterozygotes likely due to sample size (**Fig *6F***). We observed a similar trend in MDRF when exploring CKDAL1 Pro409Arg associations with fasting c-peptide (the most strongly associated trait with the variant) in rs67131976 T/T homozygotes (continuous beta = -0.04 p = 0.05), although there was no significant association in a discrete model assessing heterozygotes and homozygotes separately (**Fig 6G).** Taken together, these findings weakly suggest independence of effects, in contrast to results from the genotype fine-mapping approach.

Replication of *WFS1* Val333Ile rs1801212 association with T2D

We identified a risk-increasing, previously described variant in the *WFS1* gene (rs1801212*,* Val333Ile*,* meta-analysed beta = 0.062, p = 6.5x 10^{-232}*)*[36]*(***Fig 2A,C,D, S3, S5, S6)**. As previously reported, this was significantly associated with higher HbA1c (beta = 0.55, p = 7.7x10^{-7}), higher non-fasting glucose (Beta = 0.084, p = 6.2x10^{-4})(**Table S4)**. For the first time, we demonstrate a significant association of this variant with higher serum alkaline phosphatase (Beta = 1.08, p = 4.1x10^{-4}); given recent multi-ancestry efforts highlighting associations with ALP and cholesterol as biologically plausible and independent mechanisms underlying T2D aetiology[9], this finding warrants further investigation.

Sex-specific effects do not underlie T2D associations in exome-wide association testing

For all analyses including discovery ExWAS and downstream associations, we ran sensitivity analyses in a sex-stratified manner to ascertain whether sex-specific effects drove any of the associations we identified. We observed several nominally significant sex-specific associations, but no associations where the pooled sex-combined effect was driven by a single sex. Sex-specific and all-sex associations are given in **Table S3 and S4.**

Discussion

Summary

This study in 54,698 South Asians is the largest exome-wide association study of T2D to date, including recent multi-ancestry efforts[13]. We identify a novel association between T2D and a protective variant in *HNF4A*, a known monogenic diabetes gene. We additionally highlight associations between T2D and *RNF19A*, and dissect the association of coding variants in *GP2* and *CDKAL1.* By integrating longitudinal health records with genomic and proteomic data, we move beyond identifying novel genetic associations to uncovering their downstream health impacts and gaining mechanistic insights into variant function.

Our focus on an under-represented South Asian population enables discovery of genetic effects not observable in cohorts of European ancestry. This is exemplified by the SAS-specific, likely gain-of-function variant in *HNF4A* (Pro437Ser) which is rare to monomorphic in other populations, and an EAS and SAS-specific variant in *GP2* (Val429Met).

The discovery of *HNF4A* Pro437Ser*,* a missense variant in a known MODY gene, is particularly striking as it highlights the importance of studying under-represented ancestries with a longitudinal view of health outcomes. While this variant demonstrates strong protective associations with T2D, diabetic retinopathy, and improved glycaemic control, it also confers increased risk of myocardial infarction and higher total cholesterol. This provides a more thorough context and suggests caution in considering HNF4A as a therapeutic target for T2D.

While we observe modest associations between *HNF4A* Pro437Ser and increased plasma protein abundance of HNF4A and PCSK9, in vitro assays modelling Pro437Ser in Huh7 cell lines do not show evidence of transactivation of *PCSK9* but do show transactivation of *G6PC* (glucose-6-phosphatase, catalytic subunit). The enzyme is a target gene for HNF4A and plays a key role in glucose homeostasis[37]. Western blotting confirms the variant exerts similar level of HNF4A expression as wild-type. Pro437Ser is located in the repressor domain of *HNF4A*, and is likely

exerting its gain-of-function effect by interrupting the repression of the transcription factor.

While increased plasma abundance of PCSK9 in Pro437Ser carriers is a compelling hypothesis for the observed increased risk of MI, there is uncertainty around the tissue secreting HNF4A or PCSK9 into whole blood and therefore pQTL findings for these proteins needs to be interpreted with caution.

In contrast our findings in *GP2* Val429Met suggest an organ-specific aetiology. *GP2* encodes glycoprotein 2, an integral membrane protein that serves as a cell-surface marker for pancreatic progenitor cells and is highly expressed in pancreatic tissue[38,39]. Furthermore, *GP2*-enriched pancreatic progenitor cells can differentiate into functional beta cells both *in-vivo* and *in-vitro*, supporting its potential therapeutic relevance for diabetes [40]. We validate the association between the *GP2* variant and increased risk of T2D, potentially mediated through beta-cell secretory dysfunction. Further we show that plasma GP2 levels are elevated in individuals with greater genetic burden of beta-cell dysfunction pathways while lower levels are observed in individuals with a genetic burden related to obesity and insulin resistance.

Further functional studies are necessary to explore the effect of the missense variant on GP2 activity, and to determine whether the variant itself or GP2 plasma levels are causally related to T2D risk. Additionally, it is crucial to determine whether the observed association reflects a compensatory increase in blood plasma levels of GP2 resulting from functional inactivity. These findings strongly suggest that plasma GP2 abundance may play an indicator of higher genetic burden of beta-cell dysfunction and lower obesity-mediated T2D, highlighting its role as a potential therapeutic target, but more immediately as a biomarker for this type of diabetes that is difficult to predict and that progresses very rapidly. Further validation studies looking at GP2 trajectories would be necessary to confirm its value as a biomarker.

Challenges in identifying causal variants

We observe an association between T2D and a coding variant in *CDKAL1 (Pro409Arg*). *CDKAL1* is a well-established T2D risk gene, however, all previously reported lead variants have been entirely intronic. While Pro409Arg has been reported as passing genome-wide significance in summary statistics as part of reported risk loci, albeit with an intronic lead variant within these loci[1]. Despite being in only weak LD with the most strongly suggestive non-coding variants (e.g. rs67131976; R2 = 0.10), Pro409Arg does not appear to be causal variant in GWAS fine-mapping in Genes & Health. However, in conditional analyses, the coding missense variant was associated with increased T2D risk among individuals who were homozygous for the most strongly associated intronic variant rs67131976, an effect driven by heterozygotes for Pro409Arg (**Fig 6F).** It is not clear from our analyses whether the variant represents a novel, independent signal, which would be the first independent coding locus identified in CDKAL1, or if it represents a weaker signal at the edge an established intronic locus, which is, by chance a coding variant,

as is suggested by our fine-mapping results. This unresolved finding highlights the importance of generating sequencing data in non-European ancestries.

Limitations

While, this represents the largest study of T2D using exome sequencing, so far, it is still a small sample size for a genetic discovery. We augment our findings by meta-analysing with external datasets wherever possible. Because we use routine electronic healthcare record data for phenotype ascertainment, there is risk of miscoding, misclassification, and misdiagnosis. We aim to mitigate these risks by using the best available phenotype data – the National Diabetes Audit – which requires which requires structured and standardised data entry and national guideline-based diagnosis and monitoring of T2D. Because few genomic datasets are available which include individuals of South Asian ancestry, the possibility of replication of key findings is limited; we have attempted to overcome this by meta-analysing with biologically concordant variants in other ancestral groups, but further replication is needed to confirm these findings.

Summary

In this exome-wide association study of T2D in South Asians, we show that leveraging multi-omic and longitudinal health data in under-represented populations is not only a springboard to identify novel ancestry-specific variants but also highlights plausible biological pathways and therapeutic targets with implications to diverse global populations.

Methods

Study design

We performed a case-control genome- and exome-wide association study using participant data from the Genes & Health study based in the UK, and investigated phenome-wide associations of variants of interest. We extended phenome association analyses to the Madras Diabetes Research Federation (MDRF)[41] biobank based in India, and replicated and meta-analysed findings using UK Biobank data via the AstraZeneca PheWAS portal[19], and the type 2 diabetes knowledge portal (T2DKP)[18], a live, open access resource collating and meta-analysing summary statistics across ancestries for type 2 diabetes and related traits.

Setting

Genes & Health is a long-term community-based study[14,42], funded to recruit 100,000 British Bangladeshi and British Pakistani individuals in the UK aged 16 and over, linking genetic testing to routine electronic health record (EHR) data. To date, Genes & Health has recruited 70k volunteers, among whom 55,273, have whole exome sequencing data, and 51,170 have imputed genotype data, with linkage to primary and secondary care clinical codes, laboratory test results, prescribing data, and multi-platform proteomic and metabolic data.

We replicated and meta-analysed findings in a separate setting, the MDRF biobank in India. This is a national-level, multicentre, clinic-based disease registry, combining phenotypic data with genetic testing in a nationally-representative sample of individuals with T2D (n = 19,898) and healthy controls (n = 4,711).

Participants

Inclusion criteria: We used no specific inclusion criteria beyond those of the Genes & Health study.

Exclusion criteria: We excluded from analyses individuals diagnosed with type 2 diabetes before 18 years of age, and those with a clinical code of type 1 diabetes, maturity onset diabetes of the young (MODY), neonatal diabetes, secondary diabetes (eg drug-induced diabetes), or clinical conditions strongly predisposing to secondary diabetes (eg cystic fibrosis, pancreatectomy). Clinical code lists used for exclusion criteria are provided in **Table S7**.

Case definition (Genes & Health): We defined type 2 diabetes as an entry of type 2 diabetes in the National Diabetes Audit (NDA)[43], an annual England-wide registry of type 2 diabetes patients. Because NDA data are not yet available from 2023 onwards, we also defined as type 2 diabetes individuals with a first clinical code of type 2 diabetes in the EHR from 2023 onwards. For cases, we defined “age at diagnosis” as the age at the earliest clinical code of type 2 diabetes in the EHR.

Control definition: We defined controls as individuals *with no record of type 2 diabetes in the* NDA, and with no records of *type 2 diabetes* in the EHR at any time.

MDRF: For this replication cohort, we included all participants with a diagnosis of type 2 diabetes made after 18 years of age. As individuals are enrolled in this cohort only after clinical assessment by an endocrinologist, we undertook no further quality control on included participants. Because of substantial case / control imbalance (n = 19,898 / n = 4711) we were unable to run an effective genome-wide association study for type 2 diabetes status. Instead, we explored and replicated associations between diabetes-related traits among individuals with and without diabetes, including fasting glucose, HbA1c, and HOMA estimates of insulin resistance, deficiency, and sensitivity[44].

Variables

Exposures: We performed genetic association testing using both whole exome sequencing (exome-wide association study, ExWAS) and imputed genotype data (genome-wide association study, GWAS).

Outcomes: We defined one primary outcome: type 2 diabetes state. This was a binary variable used in regression-based association testing.

We defined a larger number of secondary outcomes (20 quantitative and 10 binary traits, **Table S8).** We examined the association of these traits with genetic exposures of interest passing a pre-determined significance threshold in either single variant or gene-based testing. These included binary traits (ie disease codes in the EHR), and quantitative traits (ie measured variables in the EHR, such as HbA1c or weight); binary traits were examined through both logistic regression and survival models. Further detail is provided in the “statistical methods” section below.

Confounders: We adjusted analyses for confounders typically included in multivariable genetic association models: age at recruitment, age at recruitment squared, sex, and the first 20 genetic principal components.

Data sources and measurements

Sample Genotyping: Genotyping information from Illumina Infinium Global Screening Array v3 (with additional multi-disease variants) was performed as previously described[45] in 51k individuals, and imputed to the TOPMed-r3 panel. We excluded variants with low INFO scores (below 0.7), minor allele frequency below 1%, and limited analyses to single nucleotide polymorphisms only.

Sample Whole Exome Sequencing: Exome sequencing was performed on 55k individuals by the Broad Institute using Twist Reagents and Illumina 150bp PE Novaseq-6000 sequencing, mapped to hg38 to produce crams. Crams with less than 85% of based at >20x coverage were excluded. UMAP was performed on the first 7 genetically determined principal components to define Pakistani and Bangladeshi ancestry individuals, which correlated highly with 1000 genomes population

references; 198 outlier samples not corresponding to Pakistani or Bangladeshi ancestry were not included in analyses. Variant QC was performed using a random forest model trained on a set of genotype-corresponding variants identified as true positives.

Bias

We identify several sources of bias, including ascertainment bias, selection bias, immortal time bias, phenotype misclassification, and indirect parental genetic effects. Because the Genes & Health cohort is enriched for autozygosity – a measure of the relatedness of an individual's immediate ancestors – we may not have fully accounted for population stratification.

Study size

We included 54,698 individuals (13,674 cases and 41,024 controls, 575 excluded) in the exome-wide association study in Genes & Health, and 45,643 in the genome-wide association study (12,011 cases and 33,623 controls, 5,536 excluded); further details are given in **Figure 1.** Because we fine-mapped variants identified in the whole-exome study using imputed genotype data, we restricted genotype analyses including GWAS to only those individuals who were included in the ExWAS; there is not complete overlap between individuals included in exome sequencing and imputed genotype samples, and therefore there is a high level of exclusion (~10%) from the imputed genotype sample.

Quantitative variables

We included quantitative trait variables defined by the Genes & Health QUANT-PY pipeline, which has been described in more detail elsewhere[46]. In brief, this identifies all primary and secondary care quantitative treat readings in the EHR; harmonises units; removes duplicates (including serial readings within narrow windows); excludes outlier values above clinically curated reasonable upper and lower bounds; and presents individual and study-level mean, median, min and max values.

Statistical methods

Genome-wide association study (GWAS):

We examined the association of single nucleotide polymorphisms (SNP) with minor allele frequency (MAF) > 1% and imputation quality metric (INFO) score > 0.7 in 45,643 participants using REGENIE v4.0, principally to enable fine-mapping of single variants identified in WES analysis. Association models were adjusted for age at recruitment, age at recruitment squared, sex, and the first 20 genetic principal components. Models were run as REGENIE "binary trait" models with approximate Firth correction below p = 0.05. We defined genome-wide significant loci as loci with a lead SNP p value < 5e-8. A summary Manhattan plot for this data is given in **Figure S7**. We did not identify any novel lead SNPs at a genome-wide significance threshold of p<5e-8, but include the summary statistics in our data deposit.

Because we intended to perform simultaneous fine-mapping in both whole exome sequence and imputed genotype data, we restricted our GWAS analyses to individuals who had both imputed genotype and whole exome sequence data. Because these samples do not overlap, we saw high numbers of participants excluded (n = 5,536, ~10%) from GWAS, which may have limited power to a small extent.

Exome-wide association study (ExWAS)– single variant:

We examined the association of 25,326 genetic variants with MAF > 1% in 54,698 participants using REGENIE v4.1[47]. Because of high inbreeding coefficients among variants with MAF ~ 0.50, we cautiously excluded all variants in the window 0.46<MAF<0.54, and restricted analysis to single nucleotide polymorphisms. The methods for this analysis were otherwise identical to the GWAS methods above. We defined exome-wide significance as a Bonferroni-corrected p value of $p < 1.97 \times 10^{-6}$ (ie, 0.05 / 25,326 tests).

Exome-wide association study (ExWAS) – gene-based testing:

We used REGENIE v4.0 to examine the association of 534,315 gene-based "masks" previously defined[12] as sets of variants below three MAF thresholds (0.01, 0.001, 0.0001) or singleton, with specific functional annotations (LOFTEE_High_confidence_loss_of_function (LOFTEE-HC-LoF), deleterious missense variants (CADD > 20 & PolyPhen > 0.466 and SIFT<0.05), other missense variants (not meeting the above criteria), and synonymous variants); full definitions are given in **Table S1**. We defined gene testing exome-wide significance as a Bonferroni-corrected p value of $p < 9.35 \times 10^{-8}$. (ie, 0.05 / 534,315 tests).

Sex-stratified analysis:

For each analysis, we additionally ran sensitivity sex-stratified models (defined as the presence or absence of a Y chromosome) to explore the possibility of sex-specific effects (**Table S4)**.

Fine mapping: To identify likely causal variants, we performed fine mapping using SuSIE v0.12.35 [35]. We defined linkage disequilibrium (LD) windows of 2Mb (1Mb either side) of the lead SNP, and present posterior inclusion probabilities (PIPs) for variants within these windows. We performed fine mapping in both imputed genotype and whole exome sequencing data, using the same individuals for both analyses. We additionally presented locuszoom plots using the R package locuszoomr[48].

Association testing with metabolic traits:

For selected variants of interest which passed significance thresholds and appeared novel, we aimed to better characterise associations with metabolic traits through an association-testing pipeline in Genes & Health. This aimed to explore the per-minor-allele association of each variant with relevant binary and quantitative traits. Traits selected were based on availability and relevance: glycaemic, anthropometric, metabolically related, and lipid traits at the time of type 2 diabetes diagnosis, and

disease-free survival from insulin resistance phenotypes and diabetic complications (**Table S8)**. For variants within a gene for which data was available, we explored association of the variant with protein expression levels from two platforms (Olink and SomaLogic), and meta-analysed results between these platforms. Further detail on proteomic quality control is provided elsewhere.

We explored the association of variants of interest with binary and quantitative traits in the Genes & Health study defined using robust, open-source pipelines[49] using multivariable regression models adjusted for age at study recruitment, sex, ancestry, and the first 20 principal components. These include binary trait phenotypes defined as ICD10 codes, including codes mapped from SNOMED (used in Genes & Health primary care data) to ICD10 (which are used for Genes & Health secondary care data), as described previously. For quantitative traits, we explored the association of variants of interest with lifetime median values.

We explored the association of variants of interest with traits related to insulin resistance at the time of diagnosis in 13,468 type 2 diabetes cases. We defined traits at the time of diagnosis as a reading within 1 year (either before or after) the time of earliest date of type 2 diabetes diagnosis in the EHR, as described previously. Where multiple measures were available within this window, we selected the value closest in time to diagnosis data . Associations were reported from regression models adjusted for age at study recruitment, sex, ancestry, and the first 20 principal components.

We additionally examined the association of variants of interest with survival from birth to insulin resistance associated phenotypes including microvascular complications of type 2 diabetes (neuropathy, retinopathy, nephropathy, and a pooled composite of all these), macrovascular complications of type 2 diabetes (coronary artery disease, peripheral vascular disease, cerebrovascular disease, and a pooled composite of these), and initiation of insulin **(**case counts given in **Table 1)**. We performed this analysis using Cox proportional hazards models adjusted for age at study recruitment, sex, ancestry, and the first 20 principal components. Individuals were censored at the earliest occurrence of: death, de-enrolment from the study, date of last data extraction, or time of event.

Meta-analysis

We meta-analysed single variant summary statistics for type 2 diabetes case-control status and phenotypes of interest identified at nominal significance ($p<0.05$) in Genes & Health trait association testing with publicly available summary statistics aggregated via the T2DKP.

For gene-based results, we meta-analysed with gene-burden summary statistics from UK Biobank available through the AstraZeneca PheWAS portal. This resource provides gene-burden summary statistics across multiple masks, stratified by ancestry; we selected the most closely aligned mask to the Genes & Health discovery mask for each gene (eg, protein-truncating variants in UKB for high-

confidence predicted loss of function variants in GnH). Because not all REGENIE gene burden masks produce a direction of effect with associated beta and standard error (eg GENE-P), for phenotypes with nominal significance ($p < 0.05$) in internal association testing but no associated beta in the most significant mask*test combination, we selected the most significant gene-burden mask*test combination within Genes & Health with a direction of effect, measured as the smallest p value. Meta-analysis was performed using random-effects meta-analysis with the R package metafor v4.6-0[50].

Where possible, we meta-analysed with phenotypes from MDRF, including meta-analysis between MDRF and the T2DKP for traits not included in the "discovery" sample Genes & Health, and particularly for traits related to HOMA estimates of insulin resistance and production.

Functional characterization of *HNF4A* variant

pTwist-CMV-puro expression vectors were generated by Twist Bioscience (San Francisco, CA, USA) for the following *HNF4A* variants: wild-type (WT, Uniprot isoform P41235-1, NM_000457.6), Pro437Ser (NM_000457.6:c.1309C>T), and Arg89Trp (NM_000457.6:c.265C>T), a confirmed loss of function variant[24]. Each construct included a C-terminal Myc tag. The *HNF4A* constructs were transfected into WT HUH7 hepatocytes using TransIT-LT1 transfection reagent (#MIR2304, Mirus Bio, Madison, WI, USA) following the manufacturer's protocol. Variant overexpression was confirmed 24 hours after transfection using Western blotting with primary antibodies against c-Myc (#3-2500, Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA) or HNF4A (#3113SCell Signaling Technology, Danvers, MA, USA) in three independent replicates. Expression of the HNF4A transcriptional target *G6PC* was assayed in HUH7 cells transiently transfected with an empty control vector (pLenti SpBsmBI sgRNA Puro (#62207, Addgene, Watertown, MA, USA)) or HNF4A WT, Pro437Ser, or Arg89Trp Myc-tagged constructs after 24 hours by qPCR as we have published previously[51]. The following primers were used for *G6PC:* FWD 5'- GCTGTGATTGGAGACTGGCTCA-3', REV 5'- GTCCAGTCTCACAGGTTACAGG-3';[52] *PCSK9*: FWD 5'- AGGGGAGGACATCATTGGTG-3', REV 5'- CAGGTTGGGGGTCAGTACC-3';[53] and the housekeeping control gene *GAPDH*: FWD 5'- CATCTTCTTTTGCGTCGCCA-3', REV 5'- TTAAAAGCAGCCCTGGTGACC-3'.

Acknowledgements

SH is funded by Wellcome Health Advances in Underrepresented Populations (HARP) Doctoral Fellowship 227532/Z/23/Z. VB is funded by the Wellcome Trust PhD programme - health data in practice: human-centred science (218584/Z/19/Z). AM, AB and S Hu are supported by I01BX006293 from VA Office of Research and Development. DS and SF are funded by the Tackling Multimorbidity at Scale

Strategic Priorities Fund programme (MR/W014416/1) delivered by the Medical Research Council and the National Institute for Health Research in partnership with the Economic and Social Research Council and in collaboration with the Engineering and Physical Sciences Research Council. IB and SF are funded by a Wellcome Discovery Award (227897/Z/23/Z) and IB acknowledges support from the National Institute for Health and Care Research Exeter Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. GK & HCM's involvement in this research was funded in part by Wellcome Trust (grant no. 220540/Z/20/A, "Wellcome Sanger Institute Quinquennial Review 2021–2026"). MKS is supported by Barts Charity grants (MGU0504) and G-002995

We acknowledgement the support of Indian Council of Medical Research (ICMR) for setting up the MDRF – Biobank.

Analyses in UKBiobank by AYD were performed under study accession number 53639.

Genes & Health is/has recently been core-funded by Wellcome (WT102627, WT210561), the Medical Research Council (UK) (M009017, MR/X009777/1, MR/X009920/1), Higher Education Funding Council for England Catalyst, Barts Charity (845/1796), Health Data Research UK (for London substantive site), and research delivery support from the NHS National Institute for Health Research Clinical Research Network (North Thames). We acknowledge the support of the National Institute for Health and Care Research Barts Biomedical Research Centre (NIHR203330); a delivery partnership of Barts Health NHS Trust, Queen Mary University of London, St George's University Hospitals NHS Foundation Trust and St George's University of London

Genes & Health is/has recently been funded by Alnylam Pharmaceuticals, Genomics PLC; and a Life Sciences Industry Consortium of AstraZeneca PLC, Bristol-Myers Squibb Company, GlaxoSmithKline Research and Development Limited, Maze Therapeutics Inc, Merck Sharp & Dohme LLC, Novo Nordisk A/S, Pfizer Inc, Takeda Development Centre Americas Inc.

We thank Social Action for Health, Centre of The Cell, members of our Community Advisory Group, and staff who have recruited and collected data from volunteers. We thank the NIHR National Biosample Centre (UK Biocentre), the Social Genetic & Developmental Psychiatry Centre (King's College London), Wellcome Sanger Institute, and Broad Institute for sample processing, genotyping, sequencing and variant annotation. This work uses data provided by patients and collected by the NHS as part of their care and support. This research utilised Queen Mary University of London's Apocrita HPC facility, supported by QMUL Research-IT, http://doi.org/10.5281/zenodo.438045

We thank: Barts Health NHS Trust, NHS Clinical Commissioning Groups (City and Hackney, Waltham Forest, Tower Hamlets, Newham, Redbridge, Havering, Barking

and Dagenham), East London NHS Foundation Trust, Bradford Teaching Hospitals NHS Foundation Trust, Public Health England (especially David Wyllie), Discovery Data Service/Endeavour Health Charitable Trust (especially David Stables), Voror Health Technologies Ltd (especially Sophie Don), NHS England (for what was NHS Digital) - for GDPR-compliant data sharing backed by individual written informed consent.

Most of all we thank all of the volunteers participating in Genes & Health.

A favourable ethical opinion for the main Genes & Health research study was granted by NRES Committee London - South East (reference 14/LO/1240) on 16 Sept 2014. Queen Mary University of London is the Sponsor, and Data Controller.

Author contribution

SH, VB, MKS: conceptualisation, formal analyses, visualisation, writing original draft and manuscript review. MB: replication data analysis, visualisation, manuscript review. RMA, VM, GP, RV, SF, DvH, CL: supervision, data generation, project administration, resources, funding acquisition, manuscript review. CM, YJ, AYD: methodology, analysis, visualisation, and reviewing manuscript. DS, JS, VB: data curation, methodology, manuscript revision.

REFERENCES

1. Suzuki, K. *et al.* Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* (2024) doi:10.1038/s41586-024-07019-6.

2. Mahajan, A. *et al.* Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet.* **54**, 560–572 (2022).

3. Loh, M. *et al.* Identification of genetic effects underlying type 2 diabetes in South Asian and European populations. *Commun Biol* **5**, 329 (2022).

4. Srinivasan, S. *et al.* Common and Distinct Genetic Architecture of Age at Diagnosis of Diabetes in South Indian and European Populations. *Diabetes Care* **46**, 1515–1523 (2023).

5. Gujral, U. P., Pradeepa, R., Weber, M. B., Narayan, K. M. V. & Mohan, V. Type 2 diabetes in South Asians: similarities and differences with white Caucasian and other populations. *Ann. N. Y. Acad. Sci.* **1281**, 51–63 (2013).

6. Unnikrishnan, R., Gupta, P. K. & Mohan, V. Diabetes in south Asians: Phenotype, clinical presentation, and natural history. *Curr. Diab. Rep.* **18**, 30 (2018).

7. Narayan, K. M. V. & Kanaya, A. M. Why are South Asians prone to type 2 diabetes? A hypothesis based on underexplored pathways. *Diabetologia* **63**, 1103–1109 (2020).

8. Hodgson, S. *et al.* Genetic basis of early onset and progression of type 2 diabetes in South Asians. *Nat. Med.* **31**, 323–331 (2025).

9. Smith, K. *et al.* Multi-ancestry polygenic mechanisms of type 2 diabetes. *Nat. Med.* (2024) doi:10.1038/s41591-024-02865-3.

10. Siddiqui, M. K. *et al.* Young-onset diabetes in Asian Indians is associated with lower measured and genetically determined beta cell function. *Diabetologia* **65**, 973–983 (2022).

11. Williams, R. & Diabetes Atlas Committee, I. D. F. … and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Res.* (2019).

12. Kim, H. I. *et al.* Exome sequencing and analysis of 44,028 British South Asians enriched for high autozygosity. *medRxiv* 2025.06.05.25329068 (2025) doi:10.1101/2025.06.05.25329068.

13. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).

14. Finer, S. *et al.* Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *Int. J. Epidemiol.* **49**, 20–21i (2020).

15. Nag, A. *et al.* Human genetics uncovers MAP3K15 as an obesity-independent therapeutic target for diabetes. *Sci. Adv.* **8**, eadd5430 (2022).

16. Warncke, K. *et al.* Frequency and characteristics of MODY 1 (HNF4A mutation) and MODY 5 (HNF1B mutation): Analysis from the DPV database. *J. Clin. Endocrinol. Metab.* **104**, 845–855 (2019).

17. Kooner, J. S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* **43**, 984–989 (2011).

18. Costanzo, M. C. *et al.* The Type 2 Diabetes Knowledge Portal: An open access genetic resource dedicated to type 2 diabetes and related traits. *Cell Metab.* **35**, 695-710.e6 (2023).

19. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).

20. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom.* **2**, 100168 (2022).

21. Kind, L. *et al.* Molecular mechanism of HNF-1A-mediated HNF4A gene regulation and promoter-driven HNF4A-MODY diabetes. *JCI Insight* **9**, (2024).

22. Lau, P., Soubeyrand, S., Hegele, R. A., Lagace, T. A. & McPherson, R. Molecular mechanism linking a novel PCSK9 copy number variant to severe hypercholesterolemia. *Atherosclerosis* **304**, 39–43 (2020).

23. Malmström, E. *et al.* Human proteome distribution atlas for tissue-specific plasma proteome dynamics. *Cell* **188**, 2810-2822.e16 (2025).

24. Kaci, A. *et al.* Functional characterization of HNF4A gene variants identify promoter and cell line specific transactivation effects. *Hum. Mol. Genet.* **33**, 894–904 (2024).

25. Spracklen, C. N. *et al.* Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**, 240–245 (2020).

26. Kim, Y. J. *et al.* The contribution of common and rare genetic variants to variation in metabolic traits in 288,137 East Asians. *Nat. Commun.* **13**, 6642 (2022).

27. Chen, J. *et al.* The trans-ancestral genomic architecture of glycemic traits. *Nat. Genet.* **53**, 840–860 (2021).

28. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).

29. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

30. Alonso, L. *et al.* TIGER: The gene expression regulatory variation landscape of human pancreatic islets. *Cell Rep.* **37**, 109807 (2021).

31. Vu, H. A. T. H. *et al.* 2121-LB: PanKbase integrated single-cell map—A comprehensive atlas of human pancreatic islets unlocking insights into Type 1 and Type 2 diabetes. *Diabetes* **74**, 2121-LB (2025).

32. Chidambaram, M., Radha, V. & Mohan, V. Replication of recently described type 2 diabetes gene variants in a South Indian population. *Metabolism* **59**, 1760–1766 (2010).

33. Steinthorsdottir, V. *et al.* A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39**, 770–775 (2007).

34. Palmer, C. J. *et al.* Cdkal1, a type 2 diabetes susceptibility gene, regulates mitochondrial function in adipose tissue. *Mol. Metab.* **6**, 1212–1225 (2017).

35. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the “Sum of Single Effects” model. *PLoS Genet.* **18**, e1010299 (2022).

36. Sandhu, M. S. *et al.* Common variants in WFS1 confer risk of type 2 diabetes. *Nat. Genet.* **39**, 951–953 (2007).

37. Hutton, J. C. & O'Brien, R. M. Glucose-6-phosphatase catalytic subunit gene family. *J. Biol. Chem.* **284**, 29241–29245 (2009).

38. Lin, Y. *et al.* Genome-wide association meta-analysis identifies GP2 gene risk variants for pancreatic cancer. *Nat. Commun.* **11**, 3175 (2020).

39. Hase, K. *et al.* Uptake through glycoprotein 2 of FimH(+) bacteria by M cells initiates mucosal immune response. *Nature* **462**, 226–230 (2009).

40. Aghazadeh, Y. *et al.* GP2-enriched pancreatic progenitors give rise to functional beta cells in vivo and eliminate the risk of teratoma formation. *Stem Cell Reports* **17**, 964–978 (2022).

41. Anjana, R. M. *et al.* ICMR-MDRF diabetes biosamples: Cohort profile. *Indian J. Med. Res.* **160**, 514–526 (2024).

42. Spreckley, M. *et al.* Advancing health and fostering community involvement in medical research through the Genes & Health study. *Trends Endocrinol. Metab.* **36**, 196–198 (2025).

43. Holman, N. *et al.* Completion of annual diabetes care processes and mortality: A cohort study using the National Diabetes Audit for England and Wales. *Diabetes Obes. Metab.* **23**, 2728–2740 (2021).

44. Staten, M. A., Miller, W. G., Bowsher, R. R., Steffes, M. W. & for the Insulin Standardization Workgroup. Insulin assay standardization: Leading to measures of insulin sensitivity and secretion for practical clinical care. *Diabetes Care* **33**, e84–e84 (2010).

45. Hodgson, S. *et al.* Integrating polygenic risk scores in the prediction of type 2 diabetes risk and subtypes in British Pakistanis and Bangladeshis: A population-based cohort study. *PLoS Med.* **19**, e1003981 (2022).

46. srison-qmul & gamma-two. *Genes-and-Health/QUANT_PY: QUANT_PY: A Python Pipeline for Quantitative Trait Attribution in Genes & Health*. (Zenodo, 2025). doi:10.5281/ZENODO.17121461.

47. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

48. Lewis, M. J. & Wang, S. locuszoomr: an R package for visualizing publication-ready regional gene locus plots. *Bioinform. Adv.* **5**, vbaf006 (2025).

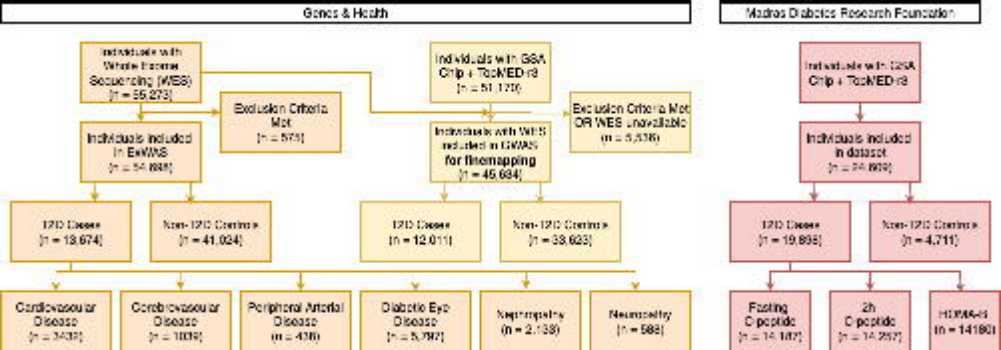
49. srison-qmul & gamma-two. *Genes-and-Health/BI_PY: BI_PY: A Python Pipeline for Binary Trait Attribution in Genes & Health*. (Zenodo, 2025). doi:10.5281/ZENODO.17121440.

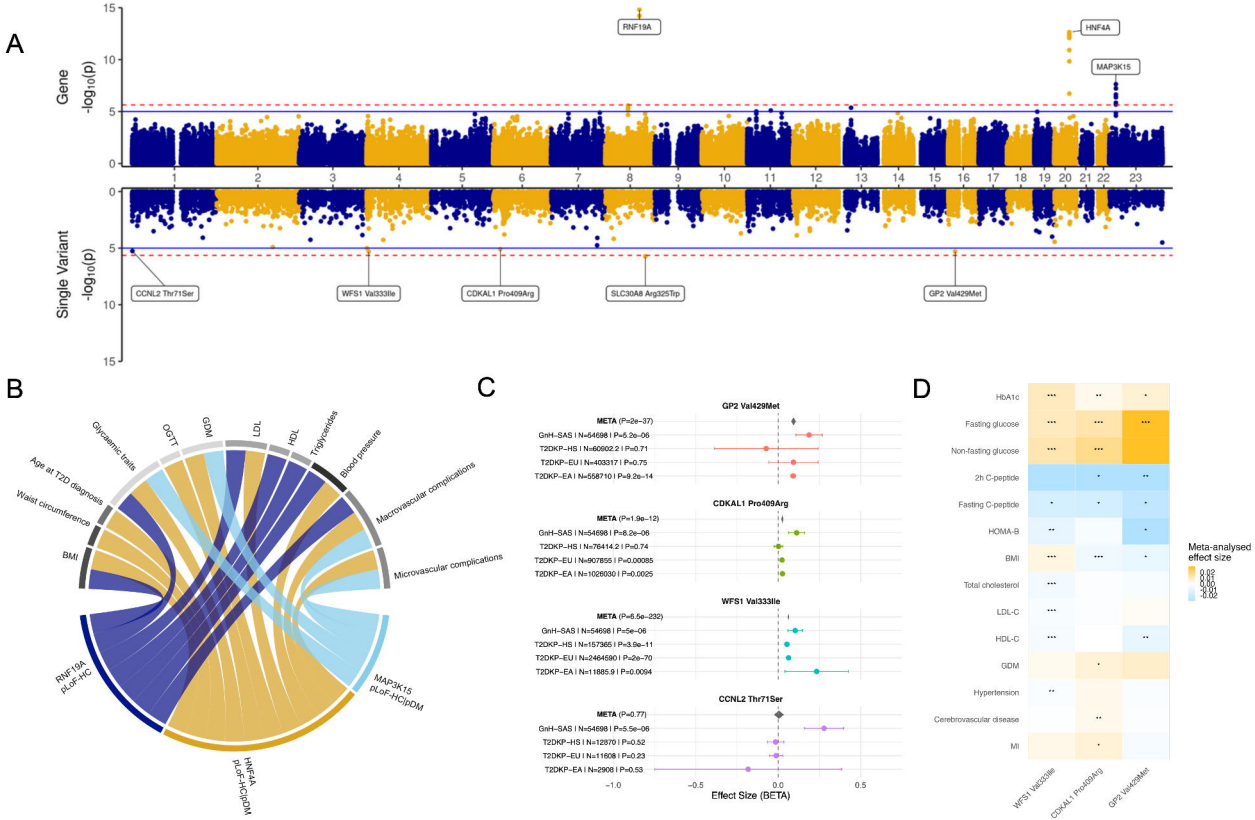
50. Lortie, C. J. & Filazzola, A. A contrast of meta and metafor packages for meta-analyses in R. *Ecol. Evol.* **10**, 10916–10921 (2020).

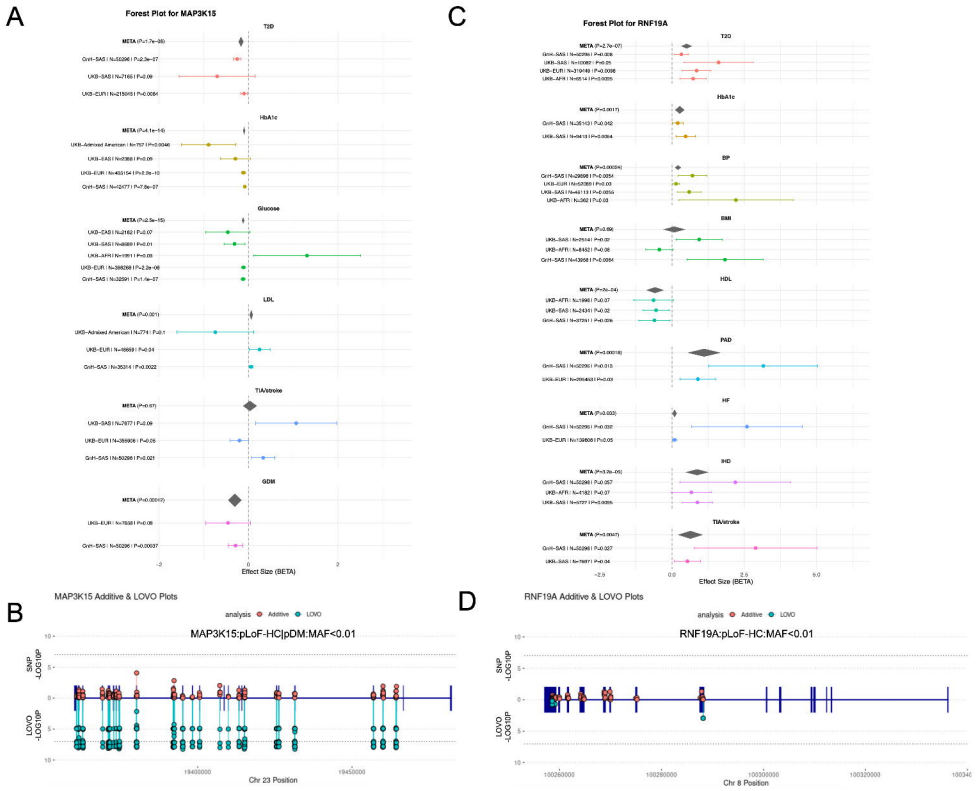
51. Du, X. *et al.* An alternatively translated isoform of PPARG suggests AF-1 domain inhibition as an insulin sensitization target. *Diabetes* **74**, 651–663 (2025).

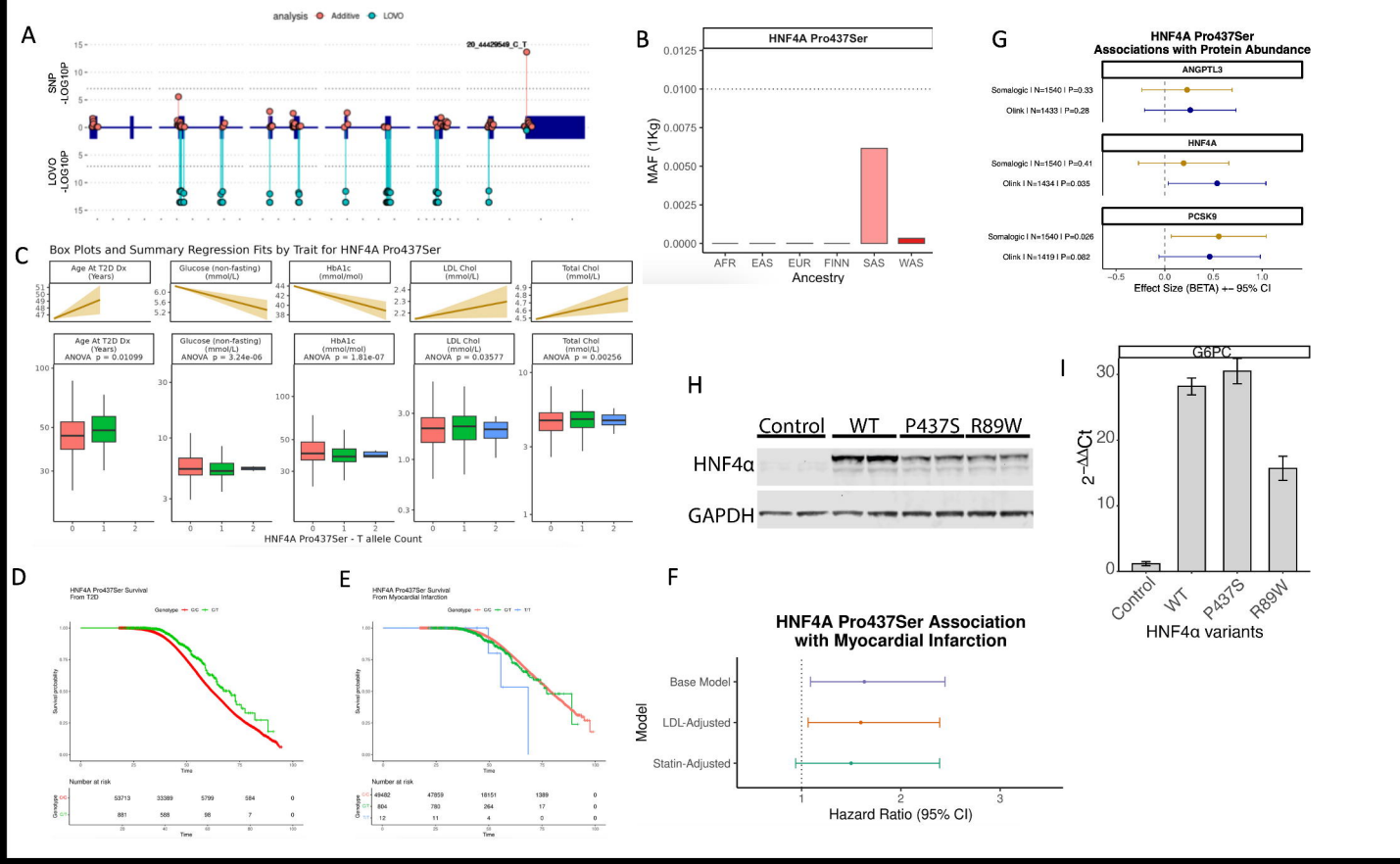
52. Morikawa, S. *et al.* Functional analysis of a novel HNF4A variant identified in a patient with MODY1. *J. Endocr. Soc.* **8**, bvae090 (2024).

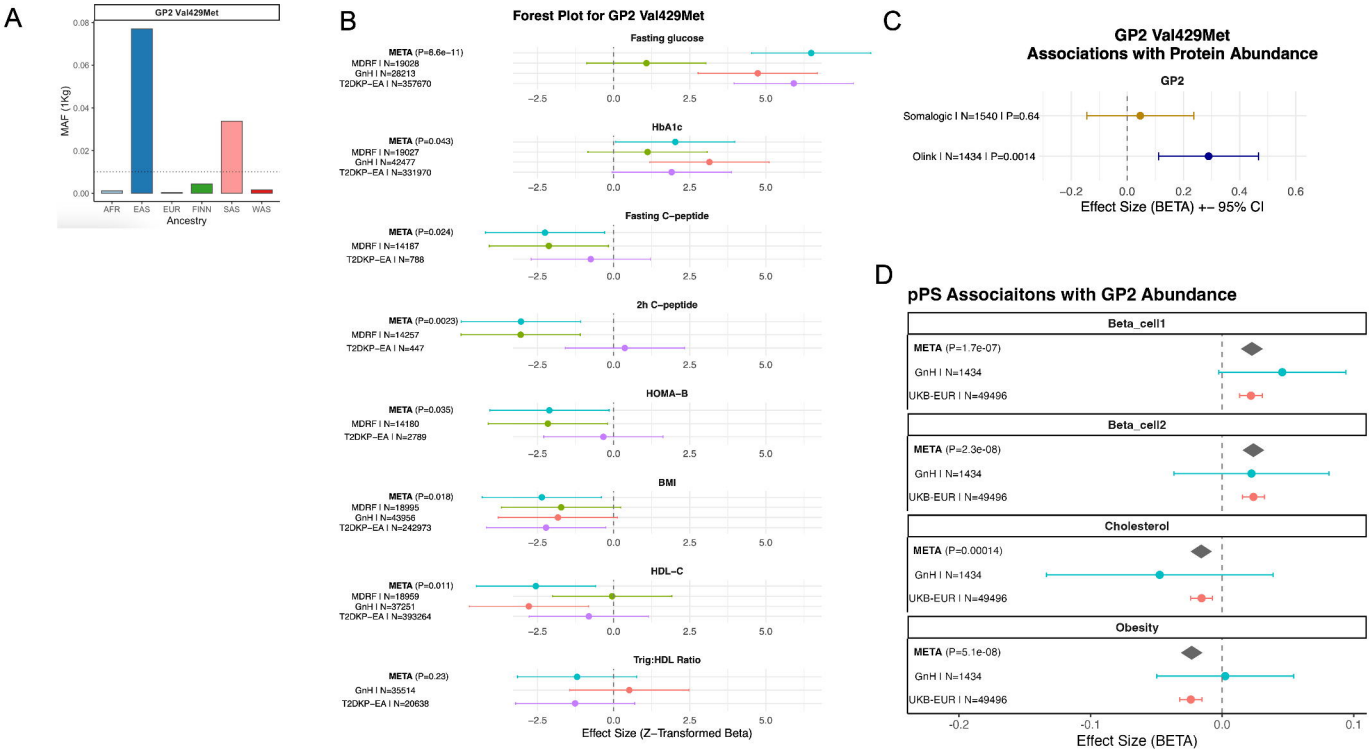
53. Dong, B. *et al.* Strong induction of PCSK9 gene expression through HNF1alpha and SREBP2: mechanism for the resistance to LDL-cholesterol lowering effect of statins in dyslipidemic hamsters. *J. Lipid Res.* **51**, 1486–1495 (2010).

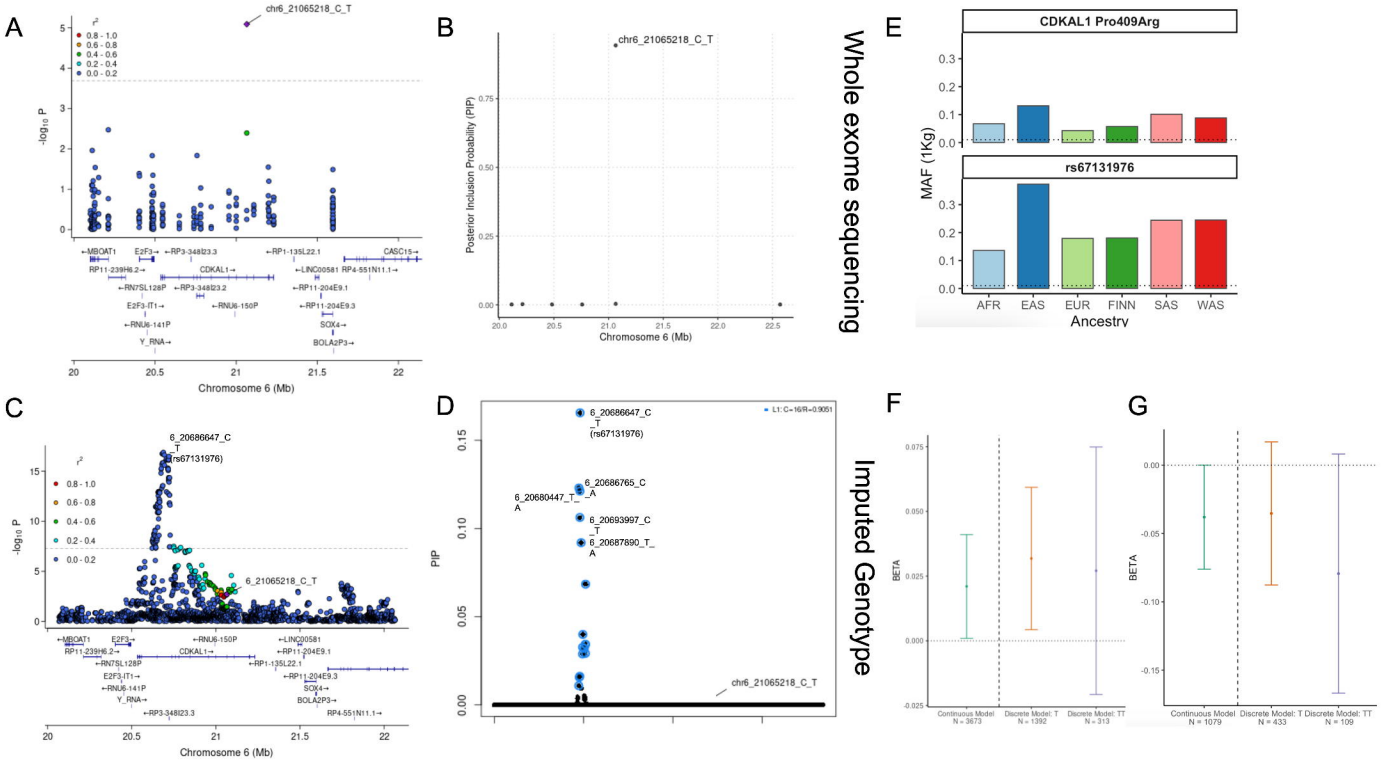

Genes & Health
Madras Diabetes Research Foundation
Individuals with Whole Exome Sequencing (WES) (n = 55,273)
Exclusion Criteria Met (n = 573)
Individuals included in ExWAS (n = 54,688)
Individuals with GSA Chip + TopMED-r3 (n = 51,170)
Exclusion Criteria Met OR WES unavailable (n = 5,536)
Individuals with WES included in GWAS for finemapping (n = 45,634)
Individuals with GSA Chip + TopMED-r3
Individuals included in dataset (n = 24,609)
T2D Cases (n = 13,674)
Non-T2D Controls (n = 41,024)
T2D Cases (n = 12,011)
Non-T2D Controls (n = 33,623)
T2D Cases (n = 19,898)
Non-T2D Controls (n = 4,711)
Cardiovascular Disease (n = 3432)
Cerebrovascular Disease (n = 1039)
Peripheral Arterial Disease (n = 436)
Diabetic Eye Disease (n = 5,797)
Nephropathy (n = 2,133)
Neuropathy (n = 588)
Fasting C-peptide (n = 14,187)
2h C-peptide (n = 14,257)
HOMA-B (n = 14180)


A
Gene -log10(p)
Single Variant -log10(p)
RNF19A
HNF4A
MAP3K15
CCNL2 Thr71Ser
WFS1 Val333Ile
CDKAL1 Pro409Arg
SLC30A8 Arg325Trp
GP2 Val429Met
B
Glycemic traits
OGTT
GDM
LDL
HDL
Triglycerides
Blood pressure
Macrovascular complications
Microvascular complications
Age at T2D diagnosis
Waist circumference
BMI
RNF19A pLoF-HC
HNF4A pLoF-HC/pDM
MAP3K15 pLoF-HC/pDM
C
GP2 Val429Met
META (P=2e−37)
GnH−SAS | N=54698 | P=5.2e−06
T2DKP−HS | N=60902.2 | P=0.71
T2DKP−EU | N=403317 | P=0.75
T2DKP−EA | N=558710 | P=9.2e−14
CDKAL1 Pro409Arg
META (P=1.9e−12)
GnH−SAS | N=54698 | P=8.2e−06
T2DKP−HS | N=76414.2 | P=0.74
T2DKP−EU | N=907855 | P=0.00085
T2DKP−EA | N=1026030 | P=0.0025
WFS1 Val333Ile
META (P=5.5e−232)
GnH−SAS | N=54698 | P=5e−06
T2DKP−HS | N=157365 | P=3.9e−11
T2DKP−EU | N=2464590 | P=2e−70
T2DKP−EA | N=11885.9 | P=0.0094
CCNL2 Thr71Ser
META (P=0.77)
GnH−SAS | N=54698 | P=5.5e−06
T2DKP−HS | N=12870 | P=0.52
T2DKP−EU | N=11608 | P=0.23
T2DKP−EA | N=2908 | P=0.53
−1.0
−0.5
0.0
0.5
Effect Size (BETA)
D
HbA1c
Fasting glucose
Non-fasting glucose
2h C-peptide
Fasting C-peptide
HOMA-B
BMI
Total cholesterol
LDL-C
HDL-C
GDM
Hypertension
Cerebrovascular disease
MI
WFS1 Val333Ile
CDKAL1 Pro409Arg
GP2 Val429Met
Meta-analysed effect size
0.02
0.01
0.00
−0.01
−0.02

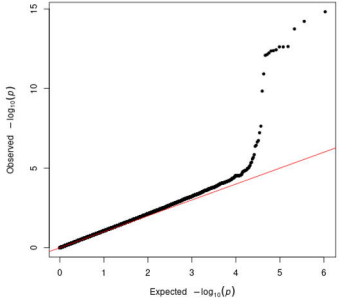

A
Forest Plot for MAP3K15
T2D
HbA1c
Glucose
LDL
TIA/stroke
GDM
Effect Size (BETA)
C
Forest Plot for RNF19A
T2D
HbA1c
BP
BMI
HDL
PAD
HF
IHD
TIA/stroke
Effect Size (BETA)
B
MAP3K15 Additive & LOVO Plots
analysis
Additive
LOVO
MAP3K15:pLoF-HC|pDM:MAF<0.01
SNP -LOG10P
LOVO -LOG10P
Chr 23 Position
D
RNF19A Additive & LOVO Plots
analysis
Additive
LOVO
RNF19A:pLoF-HC:MAF<0.01
SNP -LOG10P
LOVO -LOG10P
Chr 8 Position


A
analysis Additive LOVO
20_44429549_C_T
SNP -LOG10P
LOVO -LOG10P
B
HNF4A Pro437Ser
MAF (1Kg)
AFR EAS EUR FINN SAS WAS
Ancestry
C
Box Plots and Summary Regression Fits by Trait for HNF4A Pro437Ser
Age At T2D Dx (Years)
Glucose (non-fasting) (mmol/L)
HbA1c (mmol/mol)
LDL Chol (mmol/L)
Total Chol (mmol/L)
Age At T2D Dx (Years) ANOVA p = 0.01099
Glucose (non-fasting) (mmol/L) ANOVA p = 3.24e-06
HbA1c (mmol/mol) ANOVA p = 1.81e-07
LDL Chol (mmol/L) ANOVA p = 0.03577
Total Chol (mmol/L) ANOVA p = 0.00256
HNF4A Pro437Ser - T allele Count
D
HNF4A Pro437Ser Survival From T2D
Genotype GG GT
Survival probability
Time
Number at risk
53713 33389 5799 584 0
881 588 98 7 0
E
HNF4A Pro437Ser Survival From Myocardial Infarction
Genotype GG GT TT
Number at risk
49482 47859 18151 1389 0
804 780 264 17 0
12 11 4 0 0
F
HNF4A Pro437Ser Association with Myocardial Infarction
Model
Base Model
LDL-Adjusted
Statin-Adjusted
Hazard Ratio (95% CI)
G
HNF4A Pro437Ser Associations with Protein Abundance
ANGPTL3
Somalogic | N=1540 | P=0.33
Olink | N=1433 | P=0.28
HNF4A
Somalogic | N=1540 | P=0.41
Olink | N=1434 | P=0.035
PCSK9
Somalogic | N=1540 | P=0.026
Olink | N=1419 | P=0.082
Effect Size (BETA) +– 95% CI
H
Control WT P437S R89W
HNF4α
GAPDH
I
G6PC
2−ΔΔCt
Control WT P437S R89W
HNF4α variants

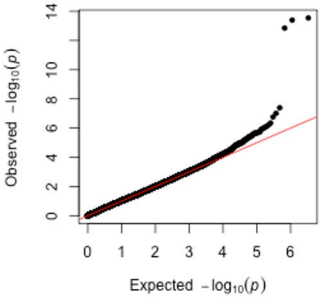

A
GP2 Val429Met
MAF (1Kg)
0.08
0.06
0.04
0.02
0.00
AFR
EAS
EUR
FINN
SAS
WAS
Ancestry
B
Forest Plot for GP2 Val429Met
Fasting glucose
META (P=6.6e−11)
MDRF I N=19028
GnH I N=28213
T2DKP−EA I N=357670
HbA1c
META (P=0.043)
MDRF I N=19027
GnH I N=42477
T2DKP−EA I N=331970
Fasting C−peptide
META (P=0.024)
MDRF I N=14187
T2DKP−EA I N=788
2h C−peptide
META (P=0.0023)
MDRF I N=14257
T2DKP−EA I N=447
HOMA−B
META (P=0.035)
MDRF I N=14180
T2DKP−EA I N=2789
BMI
META (P=0.018)
MDRF I N=18995
GnH I N=43956
T2DKP−EA I N=242973
HDL−C
META (P=0.011)
MDRF I N=18959
GnH I N=37251
T2DKP−EA I N=393264
Trig:HDL Ratio
META (P=0.23)
GnH I N=35514
T2DKP−EA I N=20638
−2.5
0.0
2.5
5.0
Effect Size (Z−Transformed Beta)
C
GP2 Val429Met
Associations with Protein Abundance
GP2
Somalogic I N=1540 I P=0.64
Olink I N=1434 I P=0.0014
−0.2
0.0
0.2
0.4
0.6
Effect Size (BETA) +− 95% CI
D
pPS Associaitons with GP2 Abundance
Beta_cell1
META (P=1.7e-07)
GnH I N=1434
UKB-EUR I N=49496
Beta_cell2
META (P=2.3e-08)
GnH I N=1434
UKB-EUR I N=49496
Cholesterol
META (P=0.00014)
GnH I N=1434
UKB-EUR I N=49496
Obesity
META (P=5.1e-08)
GnH I N=1434
UKB-EUR I N=49496
-0.2
-0.1
0.0
0.1
Effect Size (BETA)


A
chr6_21065218_C_T
r²
0.8 - 1.0
0.6 - 0.8
0.4 - 0.6
0.2 - 0.4
0.0 - 0.2
-log10 P
Chromosome 6 (Mb)
B
chr6_21065218_C_T
Posterior Inclusion Probability (PIP)
Chromosome 6 (Mb)
Whole exome sequencing
E
CDKAL1 Pro409Arg
rs67131976
MAF (1Kg)
AFR
EAS
EUR
FINN
SAS
WAS
Ancestry
C
6_20686647_C_T (rs67131976)
6_21065218_C_T
-log10 P
Chromosome 6 (Mb)
D
6_20686647_C_T (rs67131976)
6_20686765_C_A
6_20680447_T_A
6_20693997_C_T
6_20687890_T_A
chr6_21065218_C_T
PIP
Imputed Genotype
F
BETA
Continuous Model N = 3673
Discrete Model: T N = 1392
Discrete Model: TT N = 313
G
BETA
Continuous Model N = 1079
Discrete Model: T N = 433
Discrete Model: TT N = 109

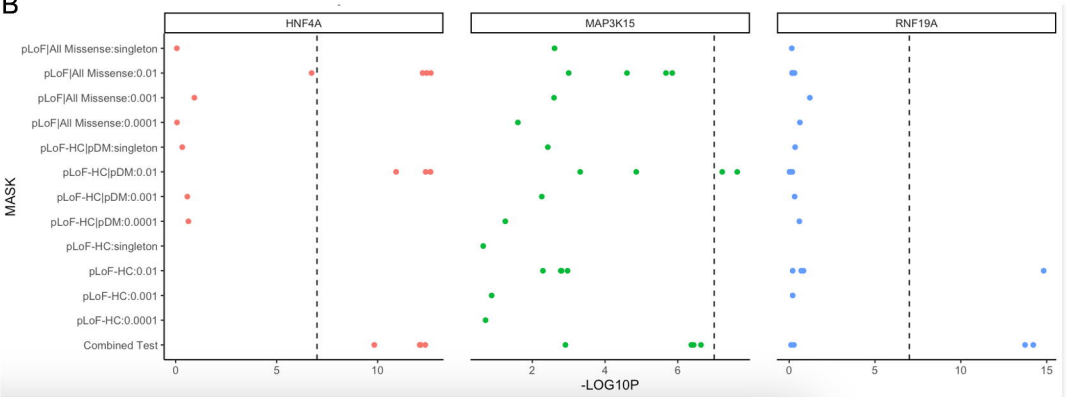
A

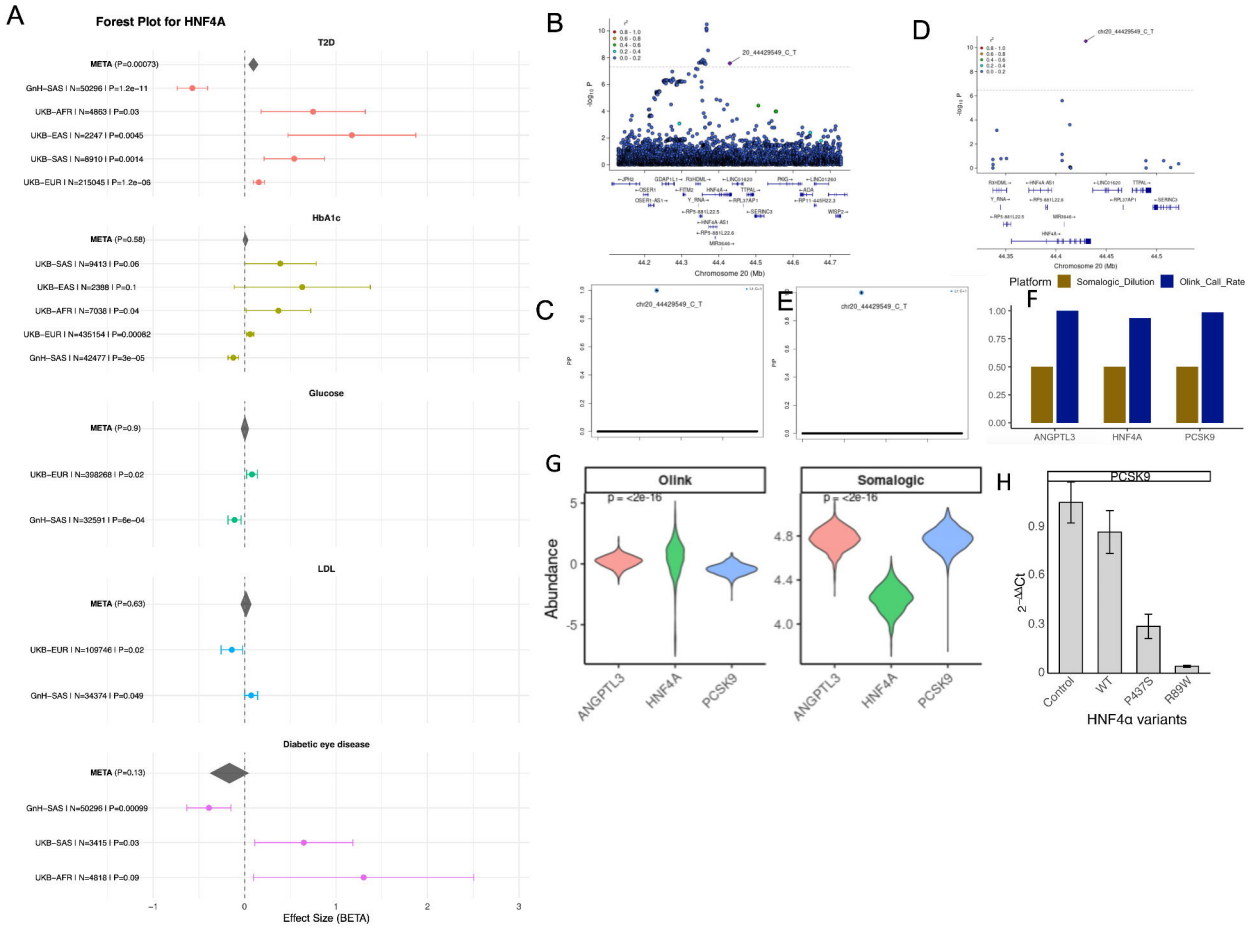

Gene-Based QQ: LambdaGC = 1.11
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
5
10
15
0
1
2
3
4
5
6

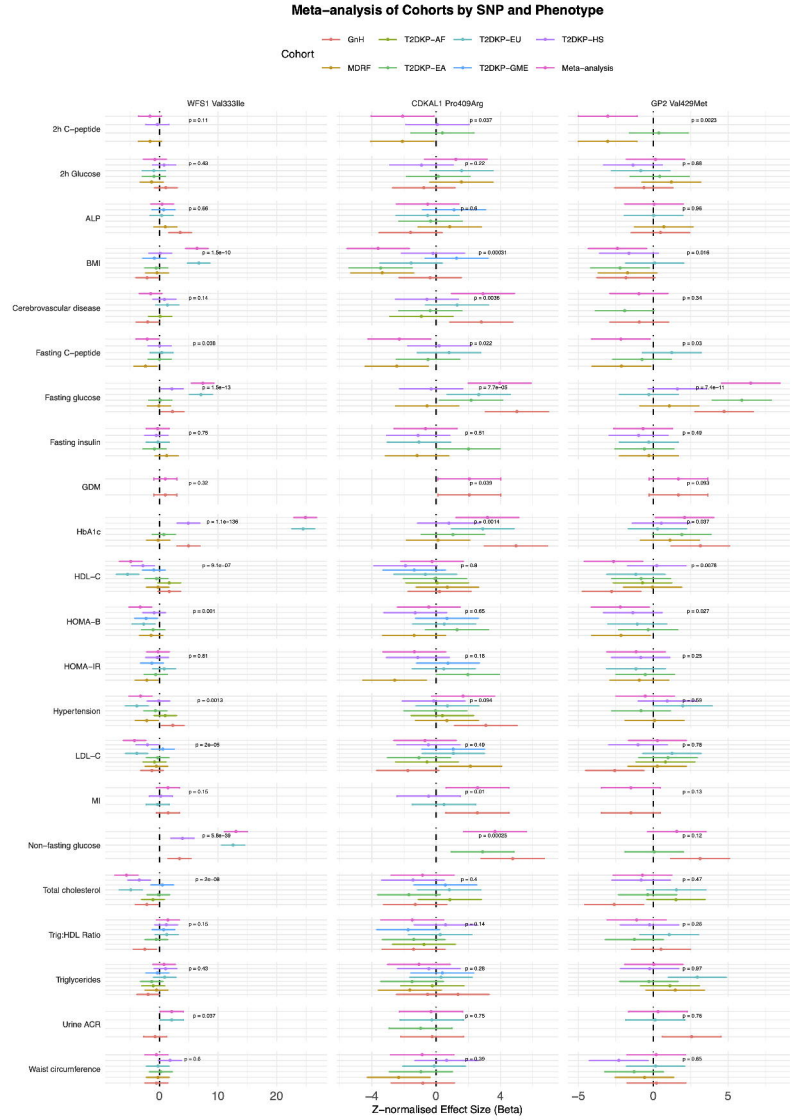
C

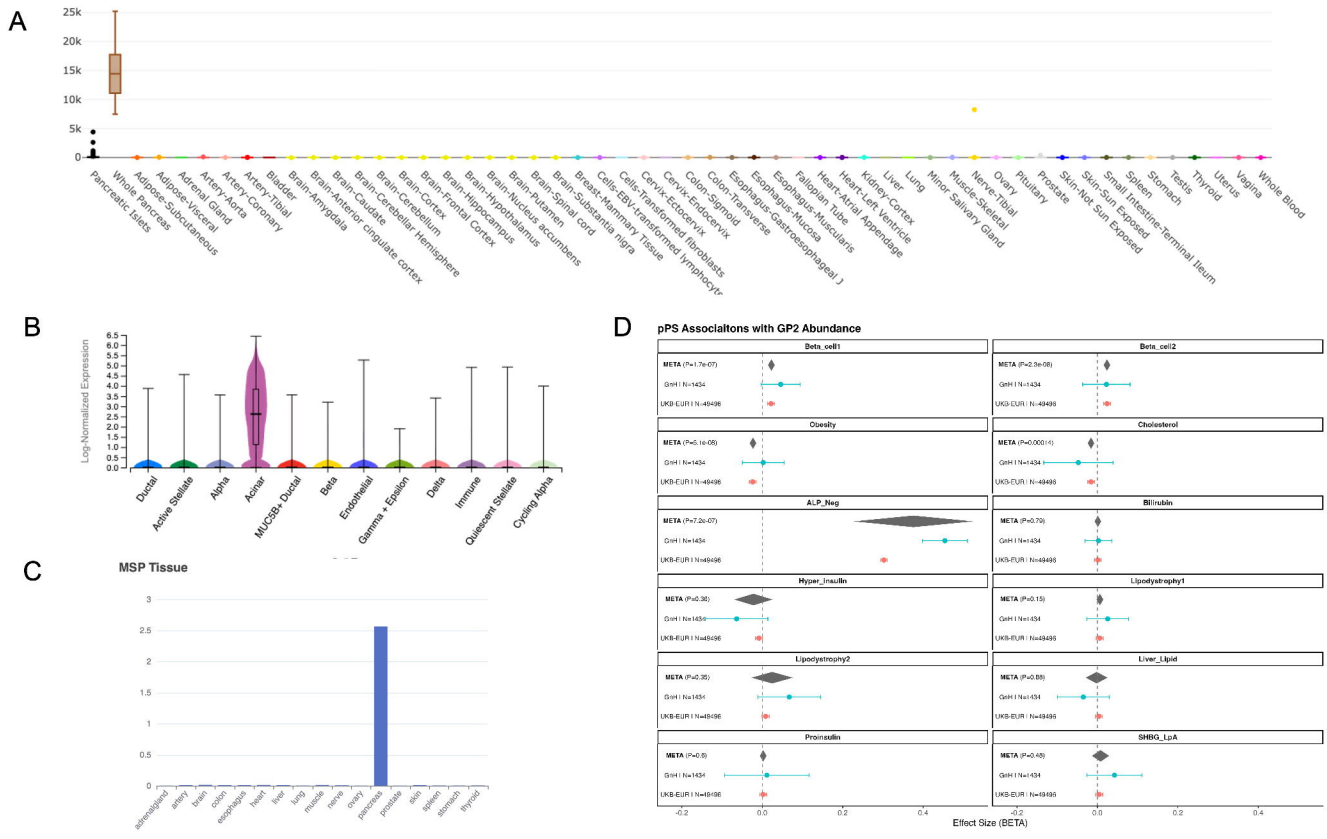

Single Variant QQ: LambdaGC = 1.10
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
2
4
6
8
10
14
0
1
2
3
4
5
6

B


HNF4A
MAP3K15
RNF19A
MASK
pLoF|All Missense:singleton
pLoF|All Missense:0.01
pLoF|All Missense:0.001
pLoF|All Missense:0.0001
pLoF-HC|pDM:singleton
pLoF-HC|pDM:0.01
pLoF-HC|pDM:0.001
pLoF-HC|pDM:0.0001
pLoF-HC:singleton
pLoF-HC:0.01
pLoF-HC:0.001
pLoF-HC:0.0001
Combined Test
0
5
10
2
4
6
0
5
10
15
-LOG10P

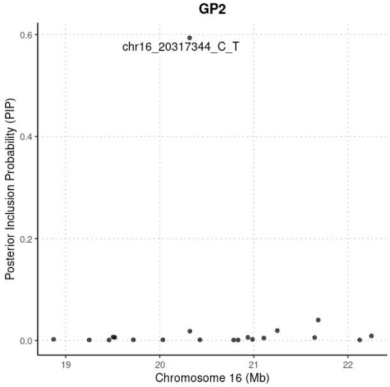

A
Forest Plot for HNF4A
T2D
META (P=0.00073)
GnH−SAS | N=50296 | P=1.2e−11
UKB−AFR | N=4863 | P=0.03
UKB−EAS | N=2247 | P=0.0045
UKB−SAS | N=8910 | P=0.0014
UKB−EUR | N=215045 | P=1.2e−06
HbA1c
META (P=0.58)
UKB−SAS | N=9413 | P=0.06
UKB−EAS | N=2398 | P=0.1
UKB−AFR | N=7038 | P=0.04
UKB−EUR | N=435154 | P=0.00082
GnH−SAS | N=42477 | P=3e−05
Glucose
META (P=0.9)
UKB−EUR | N=398268 | P=0.02
GnH−SAS | N=32591 | P=6e−04
LDL
META (P=0.63)
UKB−EUR | N=109746 | P=0.02
GnH−SAS | N=34374 | P=0.049
Diabetic eye disease
META (P=0.13)
GnH−SAS | N=50296 | P=0.00099
UKB−SAS | N=3415 | P=0.03
UKB−AFR | N=4818 | P=0.09
−1
0
1
2
3
Effect Size (BETA)
B
20_44429549_C_T
Chromosome 20 (Mb)
C
chr20_44429549_C_T
PP
D
chr20_44429549_C_T
Chromosome 20 (Mb)
E
chr20_44429549_C_T
PP
Platform
Somalogic_Dilution
Olink_Call_Rate
F
ANGPTL3
HNF4A
PCSK9
G
Olink
p = <2e-16
Somalogic
p = <2e-16
Abundance
ANGPTL3
HNF4A
PCSK9
H
PCSK9
2−ΔΔCt
Control
WT
P437S
R89W
HNF4α variants

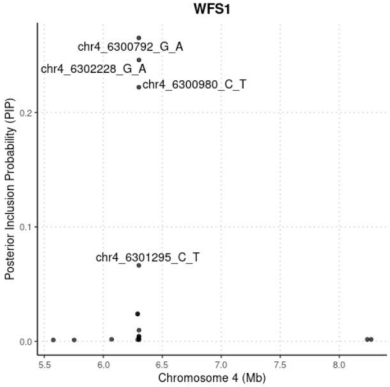

Meta−analysis of Cohorts by SNP and Phenotype
Cohort
GnH
T2DKP−AF
T2DKP−EU
T2DKP−HS
MDRF
T2DKP−EA
T2DKP−GME
Meta−analysis
WFS1 Val333Ile
CDKAL1 Pro409Arg
GP2 Val429Met
2h C−peptide
2h Glucose
ALP
BMI
Cerebrovascular disease
Fasting C−peptide
Fasting glucose
Fasting insulin
GDM
HbA1c
HDL−C
HOMA−B
HOMA−IR
Hypertension
LDL−C
MI
Non−fasting glucose
Total cholesterol
Trig:HDL Ratio
Triglycerides
Urine ACR
Waist circumference
Z−normalised Effect Size (Beta)


A
25k
20k
15k
10k
5k
0
Pancreatic-Islets
Whole Pancreas
Adipose-Subcutaneous
Adipose-Visceral
Adrenal Gland
Artery-Aorta
Artery-Coronary
Artery-Tibial
Bladder
Brain-Amygdala
Brain-Anterior cingulate cortex
Brain-Caudate
Brain-Cerebellar Hemisphere
Brain-Cerebellum
Brain-Cortex
Brain-Frontal Cortex
Brain-Hippocampus
Brain-Hypothalamus
Brain-Nucleus accumbens
Brain-Putamen
Brain-Spinal cord
Brain-Substantia nigra
Breast-Mammary Tissue
Cells-EBV-transformed lymphocyte
Cells-Transformed fibroblasts
Cervix-Ectocervix
Cervix-Endocervix
Colon-Sigmoid
Colon-Transverse
Esophagus-Gastroesophageal J
Esophagus-Mucosa
Esophagus-Muscularis
Fallopian Tube
Heart-Atrial Appendage
Heart-Left Ventricle
Kidney-Cortex
Liver
Lung
Minor Salivary Gland
Muscle-Skeletal
Nerve-Tibial
Ovary
Pituitary
Prostate
Skin-Not Sun Exposed
Skin-Sun Exposed
Small Intestine-Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
B
Log-Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Ductal
Active Stellate
Alpha
Acinar
MUC5B+ Ductal
Beta
Endothelial
Gamma + Epsilon
Delta
Immune
Quiescent Stellate
Cycling Alpha
C
MSP Tissue
3
2.5
2
1.5
1
0.5
0
adrenalgland
artery
brain
colon
esophagus
heart
liver
lung
muscle
nerve
ovary
pancreas
prostate
skin
spleen
stomach
thyroid
D
pPS Associaltons with GP2 Abundance
Beta_cell1
META (P=1.7e-07)
GrrH I N=1434
UKB-EUR I N=49496
Beta_cell2
META (P=2.3e-08)
GrrH I N=1434
UKB-EUR I N=49496
Obesity
META (P=5.1e-08)
GrrH I N=1434
UKB-EUR I N=49496
Cholesterol
META (P=0.00014)
GrrH I N=1434
UKB-EUR I N=49496
ALP_Neg
META (P=7.2e-07)
GrrH I N=1434
UKB-EUR I N=49496
Bilirubin
META (P=0.79)
GrrH I N=1434
UKB-EUR I N=49496
Hyper_insulin
META (P=0.38)
GrrH I N=1434
UKB-EUR I N=49496
Lipodystrophy1
META (P=0.15)
GrrH I N=1434
UKB-EUR I N=49496
Lipodystrophy2
META (P=0.35)
GrrH I N=1434
UKB-EUR I N=49496
Liver_Lipid
META (P=0.89)
GrrH I N=1434
UKB-EUR I N=49496
Proinsulin
META (P=0.5)
GrrH I N=1434
UKB-EUR I N=49496
SHBG_LpA
META (P=0.45)
GrrH I N=1434
UKB-EUR I N=49496
-0.2
0.0
0.2
0.4
Effect Size (BETA)

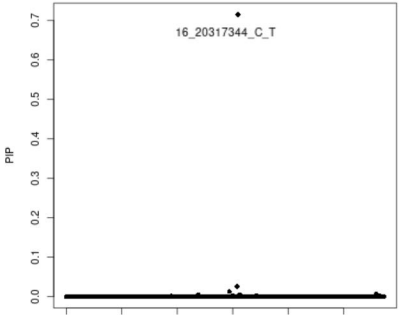
A

Whole exome sequencing

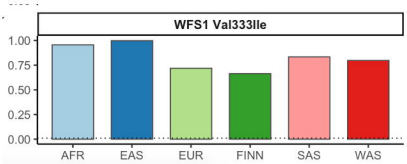


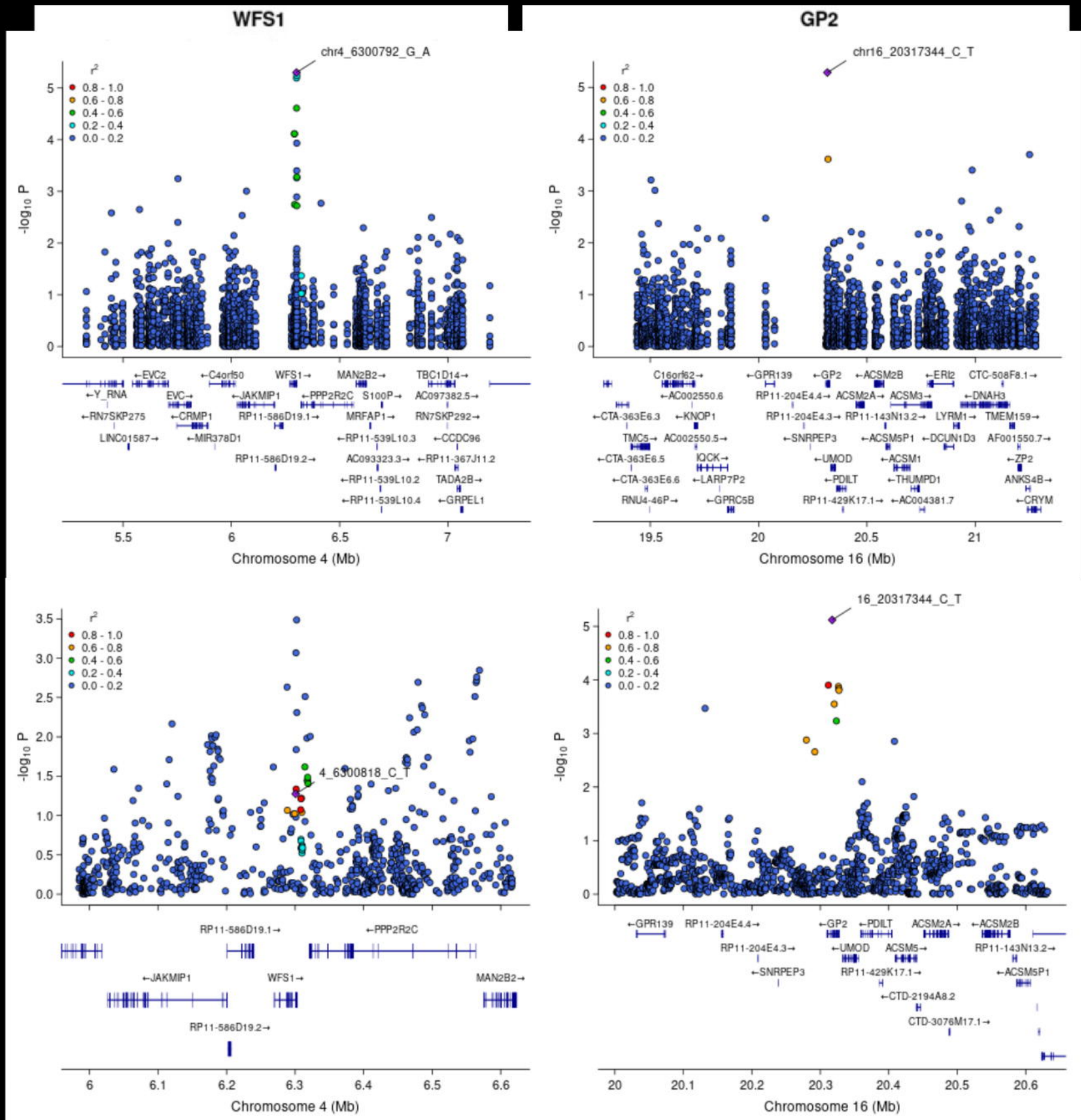


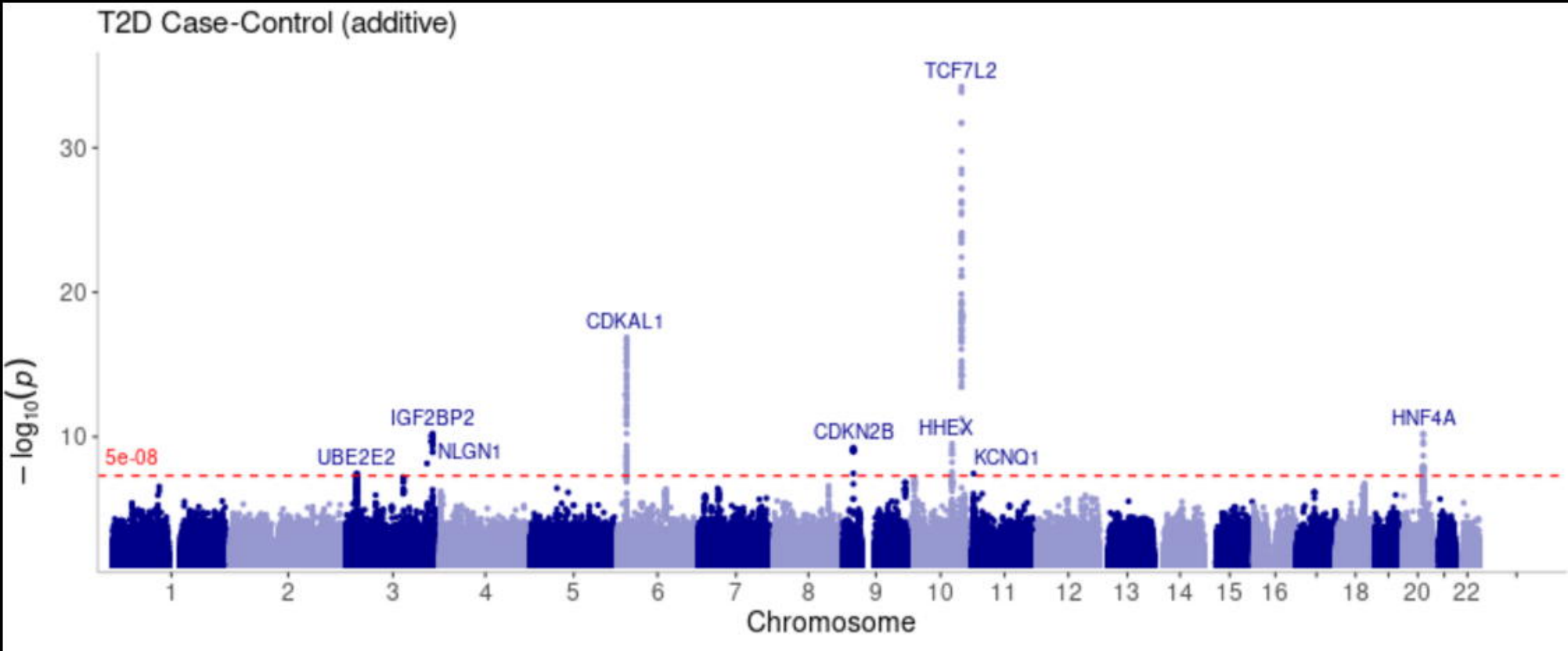
Imputed Genotype



B




WFS1
GP2
chr4_6300792_G_A
chr16_20317344_C_T
4_6300818_C_T
16_20317344_C_T
r²
0.8 - 1.0
0.6 - 0.8
0.4 - 0.6
0.2 - 0.4
0.0 - 0.2
-log10 P
Chromosome 4 (Mb)
Chromosome 16 (Mb)
EVC2
C4orf50
WFS1
MAN2B2
TBC1D14
Y_RNA
EVC
JAKMIP1
PPP2R2C
S100P
AC097382.5
RN7SKP275
CRMP1
RP11-586D19.1
MRFAP1
RN7SKP292
LINC01587
MIR378D1
RP11-539L10.3
CCDC96
RP11-586D19.2
AC093323.3
RP11-367J11.2
RP11-539L10.2
TADA2B
RP11-539L10.4
GRPEL1
C16orf62
GPR139
GP2
ACSM2B
ERI2
CTC-508F8.1
AC002550.6
RP11-204E4.4
ACSM2A
ACSM3
DNAH3
CTA-363E6.3
KNOP1
RP11-204E4.3
RP11-143N13.2
LYRM1
TMEM159
TMC5
AC002550.5
SNRPEP3
ACSM5P1
DCUN1D3
AF001550.7
CTA-363E6.5
IQCK
UMOD
ACSM1
ZP2
CTA-363E6.6
LARP7P2
PDILT
THUMPD1
ANKS4B
RNU4-46P
GPRC5B
RP11-429K17.1
AC004381.7
CRYM
ACSM5
CTD-2194A8.2
CTD-3076M17.1


T2D Case-Control (additive)
TCF7L2
CDKAL1
IGF2BP2
NLGN1
UBE2E2
CDKN2B
HHEX
KCNQ1
HNF4A
5e-08
30
20
10
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
22
Chromosome