

Genome-wide association and polygenic risk score analysis of prediabetes in South Indian women

Gayathri K. S.

Central University of Kerala

Tarun Sanjai G

Central University of Kerala

Elezebeth Mathews

Central University of Kerala

Manickavelu Alagu

amanicks@cukerala.ac.in

Central University of Kerala

Research Article

Keywords: prediabetes, genome-wide association study, polygenic risk score, MTNR1B, South Asian, India

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10628165/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose

India carries a large burden of prediabetes, but genomic studies of prediabetes in regionally sampled South Asian populations remain limited. We assessed common-variant association and type 2 diabetes polygenic risk score in women from Northern Kerala.

Methods

The cohort comprised 540 women aged 30–60 years genotyped using the Illumina Global Screening Array. After quality control, the genome-wide association study included 269 participants with prediabetes, 202 normoglycaemic controls, and 411,640 variants, with adjustment for four genetic principal components. A secondary analysis applied Polygenic Score Catalog score PGS000125 to all 540 participants; 70 of 80 loci were allele-compatible. Logistic regression, bootstrap confidence intervals, and repeated stratified tenfold cross-validation evaluated the score.

Results

The strongest regional signal occurred at *MTNR1B*, led by rs11020124 (odds ratio 1.83, 95% confidence interval 1.38–2.41; $P = 2.2 \times 10^{-5}$). The established glycaemic variant rs10830963 showed the same risk direction (odds ratio 1.77, 95% confidence interval 1.34–2.32; $P = 4.7 \times 10^{-5}$). Genomic inflation was limited ($\lambda_{GC} = 1.011$). Mean effect-allele counts were 77.53 in prediabetes and 75.94 in controls. Each standard-deviation increase was associated with 1.38-fold higher odds of prediabetes (95% confidence interval 1.16–1.65; $P = 0.00032$), but discrimination was weak (area under the receiver operating characteristic curve 0.583; bootstrap 95% confidence interval 0.534–0.630).

Conclusion

This study identified a suggestive *MTNR1B* regional signal consistent with prior Indian-Asian evidence. These findings support *MTNR1B* as an early dysglycaemia locus and highlight the need for ancestry-specific polygenic prediction models.

Introduction

Prediabetes is an intermediate stage of dysglycaemia characterized by glucose concentrations above the normal range but below the diagnostic threshold for type 2 diabetes mellitus (T2DM). It represents a critical window for disease prevention because individuals with prediabetes are at substantially increased risk of progressing to T2DM and developing cardiovascular and other metabolic complications [1]. According to the American Diabetes Association (ADA), prediabetes is diagnosed when any one of

the following criteria is met: impaired fasting glucose (IFG), impaired glucose tolerance (IGT), or elevated glycated haemoglobin (HbA1c) [2]. The International Diabetes Federation estimates that hundreds of millions of adults worldwide are affected by prediabetes, and this burden is expected to increase considerably over the coming decades, emphasizing the urgent need for effective risk stratification and early intervention strategies [3].

India bears one of the largest global burdens of prediabetes. The nationally representative ICMR-INDIAB study reported a weighted prevalence of 15.3% for prediabetes, exceeding the prevalence of diagnosed diabetes in several regions of the country [4, 5]. Furthermore, longitudinal studies from India, including Kerala, have demonstrated that nearly 60% of individuals with prediabetes progress to T2DM over time, highlighting the importance of identifying individuals at high risk before the onset of overt diabetes [6, 7]. Importantly, clinical trials such as the Indian Diabetes Prevention Programme (IDPP-1), the Diabetes Community Lifestyle Improvement Program (D-CLIP), and the Kerala Diabetes Prevention Program have demonstrated that progression from prediabetes to T2DM can be delayed through lifestyle modification and pharmacological intervention [8–10]. These findings underscore the importance of identifying genetic susceptibility during the earliest stages of dysglycaemia, when preventive strategies are most likely to be effective.

Despite the considerable clinical importance of prediabetes, its genetic architecture remains much less understood than that of established T2DM. Genome-wide association studies (GWAS) have successfully identified numerous susceptibility loci for T2DM and glycaemic traits across diverse populations, while multi-ancestry studies have demonstrated that greater population diversity improves locus discovery, fine-mapping, and the transferability of genetic risk prediction [11, 12]. However, South Asian populations remain underrepresented in genomic studies, despite developing T2DM at younger ages and lower body mass indices than many other ancestral groups. Women from Kerala represent an especially understudied population, even though they experience a substantial burden of cardiometabolic disease [13].

Polygenic risk scores (PRS) integrate the cumulative effects of multiple genetic variants and provide an estimate of inherited disease susceptibility. Although PRS have shown promise for predicting T2DM risk, their transferability across different ancestries and their applicability to prediabetes remain uncertain. Studies in Indian populations have demonstrated improved prediction using ancestry-specific, genome-wide PRS, but whether T2DM-derived polygenic scores can identify individuals at the prediabetic stage has received little attention [14–16]. Understanding the performance of PRS before the onset of overt diabetes may provide valuable insights into the genetic basis of early dysglycaemia and inform future precision prevention strategies.

The present study was designed to investigate the genetic basis of prediabetes as a distinct stage in the progression toward T2DM. Unlike conventional GWAS that compare individuals with established T2DM and normoglycaemic controls, we focused on prediabetes to identify genetic variants associated with early dysglycaemia before the onset of overt disease and its accompanying metabolic and therapeutic

influences. We therefore performed a genome-wide association study comparing women with prediabetes and normoglycaemic controls and evaluated whether a published T2DM polygenic risk score was associated with prediabetes in this South Indian population.

Methods

Study design and Phenotypic profile

This population-based study was conducted between 2022 and 2023 in Kasaragod, the northernmost district of Kerala, India. According to the 2011 Census of India, the district has a population of approximately 1.31 million, comprising 0.63 million males and 0.68 million females distributed across about 0.27 million households [17]. Participants were recruited using a multistage random sampling strategy as described previously [13]. Prediabetes was defined according to the American Diabetes Association (ADA) criteria. The overall study cohort consisted of 540 women, including 303 with prediabetes and 237 with normoglycaemia. The primary objective of this study was to investigate the genetic architecture underlying the transition from normoglycaemia to prediabetes, representing the earliest clinically detectable stage of dysglycaemia. Accordingly, the genome-wide association analysis compared women with prediabetes and normoglycaemic controls, while individuals with established T2DM were not included in the primary analysis. Excluding participants with overt T2DM enabled the investigation of genetic susceptibility to early dysglycaemia before the potential influence of disease progression, chronic hyperglycaemia, diabetes-related complications, and glucose-lowering therapy. For the GWAS, analyses were restricted to participants classified as either normoglycaemic or prediabetic who had complete phenotypic information for all 11 quantitative traits, including body mass index (BMI), body fat mass, waist circumference, fasting plasma glucose (FPG), 2-hour plasma glucose (2hPG), glycated haemoglobin (HbA1c), homeostatic model assessment of β -cell function (HOMA- β), homeostatic model assessment of insulin resistance (HOMA-IR), high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, and total cholesterol. The study was approved by the Institutional Human Ethics Committee of the Central University of Kerala (Ref. no. IHEC/CUK/2022/33 dated 28th February 2022 & CUK/IHEC/2019/061 dated 12th January 2020) written informed consent was obtained from all participants.

Genotyping and quality control

Peripheral venous blood (2 mL) was collected from each participant into EDTA tubes and stored at -80°C until processing. DNA was extracted with the Promega Wizard kit, quantified on a NanoDrop One, and genotyped using the Illumina Global Screening Array, with genome coordinates aligned to GRCh37. PLINK v1.9 was used for genotype data management and quality control. Sample quality control evaluated call rate, autosomal heterozygosity, duplicate or cryptic relatedness, and ancestry structure [18]. Autosomal variants were pruned for linkage disequilibrium using a 50-variant window, five-variant step, and r^2 threshold of 0.2 before heterozygosity and relatedness analyses. One participant from each

pair with $PI_HAT \geq 0.125$ was excluded. Variant quality control excluded unmapped and non-biallelic markers and applied call-rate, control-based Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$), and minor-allele-frequency ($\geq 1\%$) filters.

Population structure and genome-wide association analysis

Principal component analysis was conducted on approximately independent autosomal markers after linkage-disequilibrium pruning with a 200-variant window, 50-variant step, and r^2 threshold of 0.25. The first four principal components were included to control for population structure. The archived analysis identifies GAPIT v3 as the association-testing software and PLINK v1.9 as the quality-control software [18]. Prediabetes status was modelled under an additive genotype model using logistic regression. Genomic inflation was assessed using λGC . Genome-wide significance was defined as $P < 5 \times 10^{-8}$; weaker associations were considered suggestive and hypothesis-generating [20].

Haplotype analyses

The chromosomes were examined after the genome-wide scan. Haplotype blocks were defined using the Gabriel confidence-interval method. Expectation–maximisation estimates of haplotype frequency were compared between groups, with haplotypes below 1% combined and Bonferroni correction applied within blocks. Because regions were selected after inspection of the association results, these analyses are exploratory [21].

Polygenic score calculation and harmonisation

Candidate T2DM scores were identified in the Polygenic risk Score Catalog (<https://www.pgscatalog.org/>). PGS000125, an 80-locus T2DM score in which each effect-allele weight equals one, was selected for direct-genotype evaluation. A South Asian-derived score, PGS003731, was considered separately but required substantially greater variant recovery or imputation. Variants were matched by rsID and harmonised to the reported effect allele; direct and complementary-strand matches were permitted. A/T and C/G variants were flagged as palindromic and examined in sensitivity analyses. For each participant, the score was calculated as the sum of effect alleles and standardised to the cohort standard deviation. This is an unweighted effect-allele score rather than a genome-wide, log-odds-weighted PRS [14, 22].

Results

Characteristics of the cohort and phenotypic profile

The study involved 540 women, consisting of 237 with normal glucose levels and 303 classified as prediabetic. The process for selecting and screening participants is illustrated in Fig. 1. Supplementary

Table S1 summarizes the 11 quantitative metabolic traits (BMI, body fat mass, waist circumference, FPG, 2hPG, HbA1c, HOMA- β , HOMA-IR, HDL cholesterol, LDL cholesterol, and total cholesterol) of the women included in the GWAS.

The correlation analysis of the 11 cardiometabolic traits uncovered several significant biological relationships. The most robust positive correlations were found between total cholesterol and LDL cholesterol ($r = 0.894$) and between hip circumference and BMI ($r = 0.850$). A moderate correlation was noted between HbA1c and fasting glucose ($r = 0.408$), while triglycerides exhibited an inverse correlation with HDL cholesterol ($r = -0.316$), which aligns with dyslipidemia associated with insulin resistance. Principal component analysis (PCA) revealed four principal components that accounted for 66.3% of the total phenotypic variance. PC1 indicated an axis of obesity and insulin resistance, PC2 was related to lipid metabolism, PC3 focused on glycemic regulation, and PC4 highlighted atherogenic dyslipidemia. Detailed correlation coefficients and PCA loadings are provided in Supplementary Table S2.

Genotyping quality control and Principal component analysis

The genotyped dataset consisted of 540 women, which included 303 individuals diagnosed with prediabetes and 237 normoglycemic controls. After conducting sample-level quality control, 41 heterozygosity outliers and 28 related individuals were removed, leading to a final dataset of 471 unrelated participants (269 prediabetes cases and 202 controls) for the GWAS. Variant-level quality control preserved 411,640 variants for association testing. Among the 269 prediabetes cases that passed quality control, 201 (74.7%) exhibited IFG, 21 (7.8%) showed isolated IGT, and 47 (17.5%) had both IFG and IGT. The counts of samples and variants at each stage of quality control are detailed in Supplementary Table S3.

PCA was performed on the final quality-controlled dataset of 471 unrelated individuals genotyped at 653,589 variants. For PCA, chromosome 0 variants were excluded and linkage disequilibrium pruning was applied using a 200-SNP window, 50-SNP step size, and r^2 threshold of 0.25. This removed 400,538 variants and retained 252,842 approximately independent markers. The first ten principal components together accounted for the variance captured by the eigen, with PC1 explaining 13.99 percent, PC2 11.71 percent, PC3 11.37 percent, PC4 9.59 percent, and PC5 9.39 percent. Scatter plots of PC1 versus PC2, PC1 versus PC3, and PC2 versus PC3 showed overlapping distributions of normal and prediabetes samples illustrated in Fig. 2, indicating no major phenotype-driven population stratification. However, several minor clusters and a small number of outlier samples were visible, consistent with fine-scale structure within the cohort.

Genome Wide Association Analysis

After genotype quality control, 411,640 variants were available in 471 participants, comprising 269 prediabetes cases and 202 normoglycemic controls. Logistic regression implemented using the GAPIT general linear model produced valid tests for 411,592 variants (99.99%), with adjustment for four

principal components. The genomic inflation factor was close to unity, indicating little residual test-statistic inflation. The strongest signal was a cluster of SNPs in the *MTNR1B* region on chromosome 11q21.2. The lead variant was rs11020124 (chr11:92,957,495; OR = 1.83, 95% CI 1.38–2.41, $P = 2.2 \times 10^{-5}$). Multiple correlated SNPs in the same locus produced consistent signal, including rs1387153 (OR = 1.79, $P = 3.2 \times 10^{-5}$), rs10830956 (OR = 1.79, $P = 3.2 \times 10^{-5}$), the canonical glycaemic GWAS variant rs10830963 (OR = 1.77, 95% CI 1.34–2.32, $P = 4.7 \times 10^{-5}$) and rs7936247 (OR = 1.77, $P = 6.4 \times 10^{-5}$). Additional suggestive loci were identified outside chromosome 11. On chromosome 6, rs12111375 showed OR = 0.55 ($P = 2.5 \times 10^{-5}$) and rs2178373 OR = 0.50 ($P = 4.2 \times 10^{-5}$). On chromosome 17q21, rs11650969 (OR = 0.57, $P = 2.8 \times 10^{-5}$) and rs4796699 (OR = 1.79, $P = 6.0 \times 10^{-5}$) flanked a narrow region of moderate signal. Smaller suggestive peaks appeared on chromosomes 4, 12 and 13. Figure 3 displays the Manhattan plot that summarizes the results of the GWAS. The primary suggestive GWAS associations are presented in Table 1. Supplementary Table S4 presents the complete list of suggestive single-nucleotide polymorphism associations identified in the genome-wide analysis.

Table 1
Leading suggestive genome-wide association study associations

Chromosome/locus	Variant	OR	P value	Direction
11, <i>MTNR1B</i>	rs11020124	1.83	2.2×10^{-5}	Risk
6	rs12111375	0.55	2.5×10^{-5}	Protective
17	rs11650969	0.57	2.8×10^{-5}	Protective
11, <i>MTNR1B</i>	rs1387153	1.79	3.2×10^{-5}	Risk
11, <i>MTNR1B</i>	rs10830956	1.79	3.2×10^{-5}	Risk
6	rs2178373	0.50	4.2×10^{-5}	Protective
11, <i>MTNR1B</i>	rs10830963	1.77	4.7×10^{-5}	Risk
17	rs4796699	1.79	6.0×10^{-5}	Risk
11, <i>MTNR1B</i>	rs7936247	1.77	6.4×10^{-5}	Risk
OR odds ratio, <i>MTNR1B</i> melatonin receptor 1B				

Haplotype analysis

Haplotype block analysis was performed around the strongest associated region on chromosome 11 by extracting the *MTNR1B* locus (chr11:92.7–93.2 Mb, GRCh37). After filtering at MAF greater than or equal to 5 percent and genotype missingness less than or equal to 5 percent, 69 variants were retained. PLINK block estimation identified 17 haplotype blocks in this interval. Two adjacent blocks captured the main association peak. The downstream *MTNR1B* block spanned chr11:92,964,815 – 92,992,074 and contained rs10466351, rs10830962, rs10830963, GSA-rs11020131, and rs6483212. EM-based haplotype estimation showed a significant overall case-control difference (chi-square = 21.10, df = 6, $P = 0.0018$).

The major risk haplotype (designated 00100 in case-coded form) had estimated frequency 40.2 percent in prediabetes and 30.2 percent in normals, corresponding to haplotype OR 1.55 ($P = 0.0020$; Bonferroni-corrected $P = 0.012$). A protective haplotype (11000) had frequency 16.4 percent in prediabetes and 23.5 percent in normals, with OR 0.64 ($P = 0.0082$; Bonferroni-corrected $P = 0.049$).

The adjacent upstream block spanned chr11:92,940,662 – 92,962,848 and contained rs1387153, rs10830956, rs7936247, rs11020124, rs2166706, rs10830961, and GSA-rs4331050. In this block, the risk haplotype 1111100 was enriched in prediabetes (42.8 percent) relative to normals (31.3 percent), with OR 1.69 and $P = 2.0 \times 10^{-4}$. Conversely, the protective haplotype 0000011 was more common in normals (51.2 percent) than in prediabetes (39.7 percent), with OR 0.64 and $P = 0.0010$. These results indicate that the chromosome 11 signal is not driven by a single SNP alone, but by a correlated local haplotype structure spanning the *MTNR1B* region.

A second haplotype analysis was performed for the chromosome 17 region surrounding rs11650969. Four haplotype blocks were identified in chr17:41.6–41.9 Mb, with the lead pair rs11650969-rs74871137 forming a 5.35 kb block. In this block, haplotype 00 (C-A) had frequency 62.5 percent in prediabetes and 49.0 percent in normals, with OR 1.73 and $P = 4.8 \times 10^{-5}$, whereas haplotype 11 (A-G) had frequency 12.9 percent in prediabetes and 22.3 percent in normals, with OR 0.52 and $P = 2.0 \times 10^{-4}$. This supports a genuine local association signal at chromosome 17, although replication is needed. Figure 4 shows a regional association plot for the *MTNR1B* region located on chromosome 11 (chr11:92.5–93.3 Mb) displays a concentration of notable GWAS association signals surrounding the primary variants rs10830963 and rs11020124.

Polygenic score and prediabetes status

Seventy of 80 PGS000125 variants were scored in all 540 participants. Mean effect-allele count was 77.53 (standard deviation 4.91) in prediabetes and 75.94 (5.02) in controls, a mean difference of 1.58 alleles (bootstrap 95% confidence interval 0.73–2.42). Each standard-deviation increase was associated with higher odds of prediabetes (odds ratio 1.382, 95% confidence interval 1.159–1.649; $P = 0.00032$). The area under the curve was 0.583 (bootstrap 95% confidence interval 0.534–0.630), indicating weak separation (Fig. 5). Nagelkerke pseudo- R^2 was 0.033. The Brier score was 0.2403 compared with 0.2463 for a prevalence-only model, a 2.4% relative improvement.

Repeated ten-fold cross-validation produced mean area under the curve 0.5826 and mean Brier score 0.2412. The mean nested-threshold sensitivity was 44.9% and specificity 63.9%; these values are cohort-specific and do not define a clinical cutoff. Prediabetes prevalence increased from 45.2% in quartile 1 to 63.5% in quartile 4. The odds ratio was 1.287 per quartile (95% confidence interval 1.100–1.506; $P = 0.00165$), and quartile 4 had odds ratio 2.11 versus quartile 1 (95% confidence interval 1.28–3.47; $P = 0.0032$).

The complete-case analysis included 448 participants and gave odds ratio 1.35 per standard deviation with area 0.579. Excluding seven palindromic loci reduced the odds ratio to 1.26 and area to 0.558.

Across all 128 possible palindromic orientations, area ranged from 0.523 to 0.589 and the odds ratio from 1.11 to 1.42. The maximum absolute change in area after removing one variant was 0.010, for rs10830963. Only rs10830963 remained associated after false-discovery-rate adjustment across 70 exploratory locus tests ($q = 0.037$), but this single-cohort result does not establish causality (Table 2; Fig. 6).

Table 2
Robustness of the exploratory PGS000125 association

Analysis	Participants / loci	OR per SD (95% CI)	AUC
Primary	540 / 70	1.382 (1.159–1.649)	0.583
Complete cases	448 / 70	1.348 (1.111–1.635)	0.579
Exclude palindromic loci	540 / 63	1.264 (1.062–1.503)	0.558
Frequency-informed flips	540 / 70	1.315 (1.104–1.566)	0.567
All palindromic orientations	540 / 70	Range 1.111–1.419	0.523–0.589
AUC area under the receiver operating characteristic curve, CI confidence interval, OR odds ratio, SD standard deviation			

Discussion

Prediabetes is a heterogeneous intermediate stage of dysglycaemia in which impaired fasting glucose (IFG), impaired glucose tolerance (IGT), and their combination represent partially distinct pathophysiological mechanisms of glucose dysregulation. Focusing specifically on prediabetes rather than established type 2 diabetes mellitus (T2DM) is a key strength of the present study because this stage reflects the earliest clinically detectable phase of dysglycaemia, before the potential confounding effects of prolonged disease duration, glucose-lowering therapy, and chronic diabetic complications. Consequently, genetic associations identified at this stage are more likely to reflect inherited susceptibility to early glucose dysregulation than secondary consequences of established diabetes. Despite the high burden of dysglycaemia in South Asian populations, genomic studies focusing specifically on prediabetes remain scarce. To our knowledge, this is the first genome-wide genetic study of prediabetes conducted in women from Kerala. Two principal findings emerged from this study. First, multiple correlated variants within the *MTNR1B* locus demonstrated consistent suggestive associations with prediabetes, with the strongest regional signal centred on the established glycaemic variant rs10830963 and its neighbouring variants. Second, an exploratory T2DM polygenic risk score constructed from 70 available loci was significantly associated with prediabetes, although its discriminatory performance remained modest, indicating that existing T2DM-derived polygenic scores have limited predictive utility for identifying prediabetes in this South Indian population. The phenotype distribution was predominantly characterised by IFG, with 201 of 269 prediabetes cases exhibiting isolated IFG, compared with 21 cases of isolated IGT and 47 cases with combined IFG and IGT. This distribution may have favoured the detection of genetic variants influencing fasting glucose regulation,

hepatic glucose metabolism, early β -cell dysfunction, and circadian regulation, consistent with the observed association at the *MTNR1B* locus. In contrast, isolated IGT is more closely related to postprandial glucose regulation and peripheral insulin resistance, whereas combined IFG and IGT represents a more advanced metabolic disturbance. Because the numbers within each subtype were insufficient for adequately powered subgroup analyses, these mechanistic interpretations remain speculative. Future studies with larger cohorts should investigate IFG, isolated IGT, combined IFG + IGT, and continuous glycaemic traits separately to better define the genetic architecture underlying the progression from normoglycaemia to overt diabetes.

Indian diabetes is also aetiologically heterogeneous. WellGen analyses identified different polygenic profiles among insulin-deficient, obesity-related and other T2DM subgroups. Prediabetes is likely at least as heterogeneous. A single score may therefore average across beta-cell, adiposity, hepatic-lipid and insulin-resistance pathways. Larger studies should compare pathway-partitioned scores and test whether genetic components interact with fasting glucose, 2-hour glucose, insulin, waist circumference or reproductive history ¹⁶.

The *MTNR1B* result is biologically coherent. Prior work in Indian Asians identified genome-wide-significant fasting-glucose associations near *MTNR1B* and showed association with T2DM. Experimental and human studies indicate that increased melatonin signalling can impair insulin secretion, providing a plausible connection to fasting dysglycaemia. The risk direction at rs10830963-G in Kerala agrees with this evidence. Concordance of direction is more informative here than the point estimate, which is likely inflated by small sample size and selection of the strongest observed regional signal ^{23,24}. The local haplotype results provide additional support for correlated variation across the *MTNR1B* region. Risk haplotypes were enriched among prediabetes cases, whereas alternative haplotypes were more frequent among controls. However, these analyses were selected after inspection of the GWAS results. Independent replication and conditional analyses are needed to determine whether the haplotype associations contain information beyond the lead variants.

The polygenic risk score result should be interpreted against better-resourced Indian studies. Pemmasani and colleagues used imputed genotypes, South Asian DIAMANTE weights and LDpred2, reporting AUC 0.795 in development and 0.778 in an Indian test dataset. Our area of 0.583 is lower for several expected reasons: the outcome was prediabetes rather than T2DM ; PGS000125 contains only 80 unweighted loci; 10 loci were unavailable; clinical covariates were absent; and the score was not developed for this population. The contrast shows that score architecture, imputation, ancestry matching and phenotype alignment matter more than the label “PRS” ¹⁵. From a clinical perspective, genetic risk assessment is likely to be most valuable during the prediabetic stage, when progression to overt diabetes remains preventable through lifestyle modification and targeted interventions. Although the predictive performance of the present exploratory PRS was modest and is not suitable for clinical implementation, the observed association suggests that inherited genetic susceptibility is already detectable before the onset of overt T2DM. Future studies integrating ancestry-specific polygenic scores with established clinical risk factors, including body mass index, family history, fasting plasma glucose,

and lifestyle characteristics, may substantially improve risk stratification and help identify individuals who would benefit most from early preventive interventions.

Li and colleagues used five pathway-partitioned polygenic scores to cluster participants with prediabetes and related those clusters and lifestyle scores to incident T2DM in two prospective cohorts. That framework is useful for organizing future Kerala studies, but it cannot be reproduced from the present cross-sectional dataset. A comparable analysis would require substantially larger subtype samples, multiple validated pathway scores, standardized lifestyle measurements, and prospective diabetes follow-up²⁵.

The comparative analysis of the *MTNR1B* locus on chromosome 11 confirms that G is the risk-associated minor allele. Evaluating the minor allele frequency (MAF) of rs10830963 across global population reference panels places Northern Kerala females (MAF = 0.30) squarely within the broader South Asian range. This frequency is comparable to American populations (0.32) and substantially higher than African populations (0.20). These population-level variations highlight the distinct genetic architecture underlying glycemic susceptibility in South Asian lineages, consistent with findings from large-scale South Asian and Indian-Asian genome-wide association studies^{26,27}.

The validity of these genetic findings is grounded in a rigorous analytical framework. By applying strict sample and variant-level quality control, pruning for linkage disequilibrium, and adjusting for four genetic principal components, we achieved an ideal genomic inflation factor ($\lambda_{GC} = 1.011$), confirming that the observed signals are free from test-statistic inflation or major population stratification. The study further emphasizes transparency through repeated ten-fold cross-validation, bootstrap confidence intervals, complete-case evaluations, and sensitivity analyses regarding palindromic variant orientations. This study has strengths. It examines an underrepresented regional population, applies sample and variant quality control, controls the GWAS for genetic principal components, uses a transparent catalogued score and reports uncertainty, internal validation and sensitivity analyses. It also connects a regional signal with established Indian-Asian *MTNR1B* evidence rather than claiming a novel locus.

Conclusion

This study represents the first genome-wide genetic investigation of prediabetes in women from Kerala, South India. The genome-wide association analysis identified a suggestive cluster of variants within the *MTNR1B* locus, consistent with the established role of this gene in glycaemic regulation and early dysglycaemia. In addition, an exploratory polygenic risk score based on 70 established T2DM-associated loci was significantly associated with prediabetes, although its discriminatory performance remained modest, indicating that currently available T2DM-derived polygenic scores have limited predictive utility for identifying prediabetes in this population. Collectively, these findings provide new insights into the genetic architecture of prediabetes in an understudied South Indian population and highlight the potential of integrating genome-wide association analysis with polygenic risk assessment to investigate genetic susceptibility during the earliest stages of dysglycaemia. Future studies incorporating larger

cohorts, refined prediabetes subphenotypes, imputed genome-wide data, ancestry-specific polygenic scores, relevant clinical risk factors, and external prospective validation will be essential to determine whether the observed *MTNR1B* association and polygenic risk can be translated into improved early risk stratification and prevention of progression to T2DM.

Declarations

Acknowledgment

The authors thank the DBT India Alliance and Wellcome Trust–Department of Biotechnology India Alliance for fellowship support during the study and the study participants for their cooperation during sample collection.

Author contributions

Gayathri K S: Experiments, Data analysis and interpretation, Writing

Tarun Sanjai : Data analysis and interpretation, Writing

Elezebeth Mathews: Funding acquisition, Supervision and Conceptualization, Writing- Reviewing and Editing,

Manickavelu Alagu: Supervision and Conceptualization, Data Analysis, Writing-Reviewing and Editing.

Funding: This study was supported by a Clinical and Public Health Early Career Fellowship (grant no. IA/CPHE/17/1/503345) awarded to E.M. by the DBT India Alliance/Wellcome Trust–Department of Biotechnology India Alliance (2018–2024).

Competing interests: The authors declare that they have no competing interests.

Ethics approval: The study was approved by the Institutional Human Ethics Committee of the Central University of Kerala (Ref. no. IHEC/CUK/2022/33 dated 28th February 2022 & CUK/IHEC/2019/061 dated 12th January 2020). Written informed consent was obtained from all participants.

Consent to participate: Written informed consent was obtained from all participants.

Data availability: The dataset analysed in this study is available from the corresponding author on reasonable request, subject to applicable legal and ethical restrictions.

References

1. Hostalek, U. Global epidemiology of prediabetes - present and future perspectives. *Clin Diabetes Endocrinol* 5, 5 (2019). <https://doi.org/10.1186/s40842-019-0080-0>.

2. American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical Care in Diabetes-2019. *Diabetes Care*. 2019, 42:S13–28.
3. International diabetes federation: IDF diabetes atlas - 8th edition, 2017.
4. Ranjit Mohan Anjana, Mohan Deepa, Rajendra Pradeepa, Jagadish Mahanta, Kanwar Narain, Hiranya Kumar Das et al. Prevalence of diabetes and prediabetes in 15 states of India: results from the ICMR–INDIAB population-based cross-sectional study. *The Lancet Diabetes & Endocrinology*. 2017;5(8):585-596. 10.1016/s2213-8587(17)30174-2
5. Ranjit Mohan Anjana, Coimbatore Subramanian Shanthi Rani, Mohan Deepa, Rajendra Pradeepa, Vasudevan Sudha, Haridas Divya Nair et al. Incidence of Diabetes and Prediabetes and Predictors of Progression Among Asian Indians: 10-Year Follow-up of the Chennai Urban Rural Epidemiology Study (CURES). *Diabetes Care*. 2015;38(8):1441-1448. 10.2337/dc14-2814.
6. Gadadharan Vijayakumar, Sreeja Manghat, Revathi Vijayakumar, Leena Simon, Liss Maria Scaria, Aswathi Vijayakumar et al. Incidence of type 2 diabetes mellitus and prediabetes in Kerala, India: results from a 10-year prospective cohort. *BMC Public Health*. 2019;19(1). 10.1186/s12889-019-6445-6
7. Ranil Jayawardena, Priyanga Ranasinghe, Nuala M Byrne, Mario J Soares, Prasad Katulanda, Andrew P Hills. Prevalence and trends of the diabetes epidemic in South Asia: a systematic review and meta-analysis. *BMC Public Health*. 2012;12(1). 10.1186/1471-2458-12-380
8. A. Ramachandran, C. Snehalatha, S. Mary, B. Mukesh, A. D. Bhaskar, V. Vijay. The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). *Diabetologia*. 2006;49(2):289-297. 10.1007/s00125-005-0097-z
9. Mary Beth Weber, Harish Ranjani, Lisa R. Staimez, Ranjit M. Anjana, Mohammed K. Ali, K.M. Venkat Narayan et al. The Stepwise Approach to Diabetes Prevention: Results From the D-CLIP Randomized Controlled Trial. *Diabetes Care*. 2016;39(10):1760-1767. 10.2337/dc16-1241
10. Kavumpurathu R. Thankappan, Thirunavukkarasu Sathish, Robyn J. Tapp, Jonathan E. Shaw, Mojtaba Lotfaliany, Rory Wolfe et al. A peer-support lifestyle intervention for preventing type 2 diabetes in India: A cluster-randomized controlled trial of the Kerala Diabetes Prevention Program. *PLOS Medicine*. 2018;15(6):e1002575. 10.1371/journal.pmed.1002575
11. Jaspal S Kooner, Danish Saleheen, Xueling Sim, Joban Sehmi, Weihua Zhang, Philippe Frossard et al. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature Genetics*. 2011;43(10):984-989. 10.1038/ng.921
12. Anubha Mahajan, Cassandra N. Spracklen, Weihua Zhang, Maggie C. Y. Ng, Lauren E. Petty, Hidetoshi Kitajima et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nature Genetics*. 2022;54(5):560-572. 10.1038/s41588-022-01058-3
13. Mathews E, Sathish T, Joseph A, Kodapally B, Thulaseedharan JV, Narayan K MV, et al. Effectiveness and implementation of a lifestyle modification intervention for women with isolated impaired fasting

- glucose: study protocol for a hybrid type 2 study in Kerala, India. *Wellcome Open Res* 2022; 7:62.
14. Samuel A. Lambert, Laurent Gil, Simon Jupp, Scott C. Ritchie, Yu Xu, Annalisa Buniello et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*. 2021;53(4):420-425. 10.1038/s41588-021-00783-5
 15. Sandhya Kiran Pemmasani, Shravya Atmakuri, Anuradha Acharya. Genome-wide polygenic risk score for type 2 diabetes in Indian population. *Scientific Reports*. 2023;13(1). 10.1038/s41598-023-38768-5
 16. Chittaranjan S. Yajnik, Rucha Wagh, Pooja Kunte, Olof Asplund, Emma Ahlqvist, Dattatrey Bhat et al. Polygenic scores of diabetes-related traits in subgroups of type 2 diabetes in India: a cohort study. *The Lancet Regional Health - Southeast Asia*. 2023;14:100182. 10.1016/j.lansea.2023.100182
 17. Office of the Registrar General & Census Commissioner. Census 2011, India. 2011. Accessed on April 20, 2026.
 18. Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. 2007;81(3):559-575. 10.1086/519795
 19. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81:559-75.
 20. Alexander E. Lipka, Feng Tian, Qishan Wang, Jason Peiffer, Meng Li, Peter J. Bradbury et al. GAPIT: genome association and prediction integrated tool. *Bioinformatics*. 2012;28(18):2397-2399. 10.1093/bioinformatics/bts444.
 21. Stacey B. Gabriel, Stephen F. Schaffner, Huy Nguyen, Jamie M. Moore, Jessica Roy, Brendan Blumenstiel et al. The Structure of Haplotype Blocks in the Human Genome. *Science*. 2002;296(5576):2225-2229. 10.1126/science.1069424
 22. Qibin Qi, Adrienne M. Stilp, Tamar Sofer, Jee-Young Moon, Bertha Hidalgo, Adam A. Szpiro et al. Genetics of Type 2 Diabetes in U.S. Hispanic/Latino Individuals: Results From the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Diabetes*. 2017;66(5):1419-1425. 10.2337/db16-1150
 23. John C. Chambers, Weihua Zhang, Delilah Zabaneh, Joban Sehmi, Piyush Jain, Mark I. McCarthy et al. Common Genetic Variation Near Melatonin Receptor MTNR1B Contributes to Raised Plasma Glucose and Increased Risk of Type 2 Diabetes Among Indian Asians and European Caucasians. *Diabetes*. 2009;58(11):2703-2708. 10.2337/db08-1805
 24. Tiinamaija Tuomi, Cecilia L.F. Nagorny, Pratibha Singh, Hedvig Bennet, Qian Yu, Ida Alenkvist et al. Increased Melatonin Signaling Is a Risk Factor for Type 2 Diabetes. *Cell Metabolism*. 2016;23(6):1067-1077. 10.1016/j.cmet.2016.04.009
 25. Yang Li, Guo-Chong Chen, Jee-Young Moon, Rhonda Arthur, Daniela Sotres-Alvarez, Martha L. Daviglus et al. Genetic Subtypes of Prediabetes, Healthy Lifestyle, and Risk of Type 2 Diabetes. *Diabetes*. 2024;73(6):893–906. doi:10.2337/db23-0699

26. Thomas Sparsø, Amélie Bonnefond, Ehm Andersson, Nabila Bouatia-Naji, Johan Holmkvist, Lise Wegner, Niels Grarup, Anette P. Gjesing, Karina Banasik, Christine Cavalcanti-Proença, Marion Marchand, Martine Vaxillaire, Guillaume Charpentier, Marjo-Riitta Jarvelin, Jean Tichet, Beverley Balkau, Michel Marre, Claire Lévy-Marchal, Kristine Færch, Knut Borch-Johnsen, Torben Jørgensen, Sten Madsbad, Pernille Poulsen, Allan Vaag, Christian Dina, Torben Hansen, Oluf Pedersen, Philippe Froguel; G-allele of Intronic rs10830963 in *MTNR1B* Confers Increased Risk of Impaired Fasting Glycemia and Type 2 Diabetes Through an Impaired Glucose-Stimulated Insulin Release: Studies Involving 19,605 Europeans. *Diabetes* 1 June 2009; 58 (6): 1450–1456.
<https://doi.org/10.2337/db08-1660>
27. Lyssenko V, Nagorny CL, Erdos MR, Wierup N, Jonsson A, Spégel P, Bugliani M, Saxena R, Fex M, Pulizzi N, Isomaa B, Tuomi T, Nilsson P, Kuusisto J, Tuomilehto J, Boehnke M, Altshuler D, Sundler F, Eriksson JG, Jackson AU, Laakso M, Marchetti P, Watanabe RM, Mulder H, Groop L. Common variant in *MTNR1B* associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat Genet.* 2009 Jan;41(1):82-8. doi: 10.1038/ng.288. Epub 2008 Dec 7. PMID: 19060908; PMCID: PMC3725650.

Figures

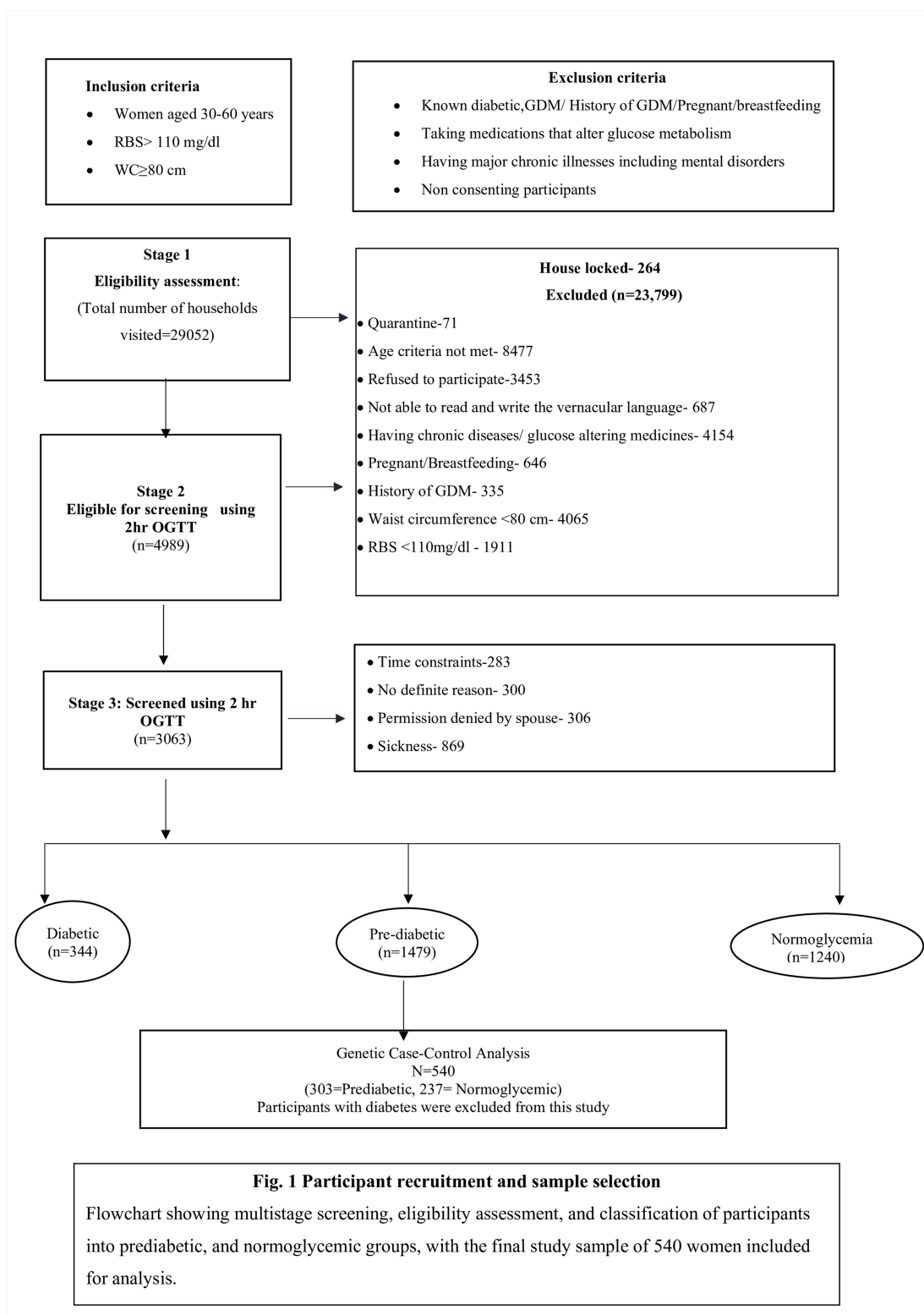


Figure 1

Participant recruitment and sample selection. Flow chart showing multistage screening, eligibility assessment, and classification of participants into prediabetic and normoglycaemic groups, with the final study sample of 540 women included for analysis.

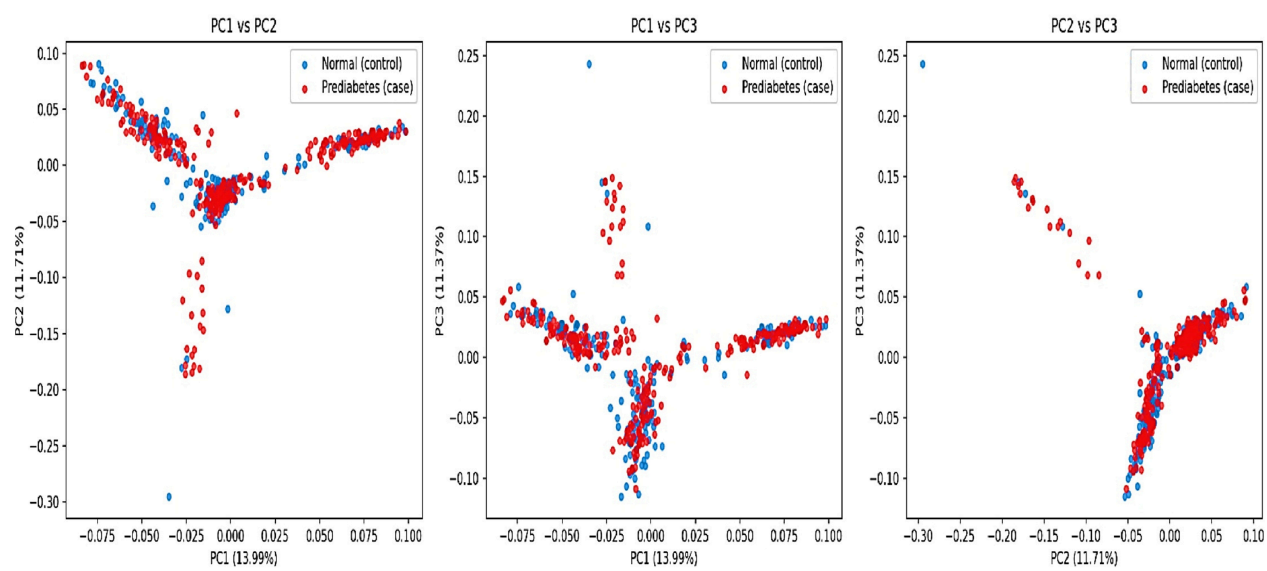


Figure 2

PCA scatter plots of the first three principal components for 471 individuals coloured by phenotype (normal vs prediabetes).

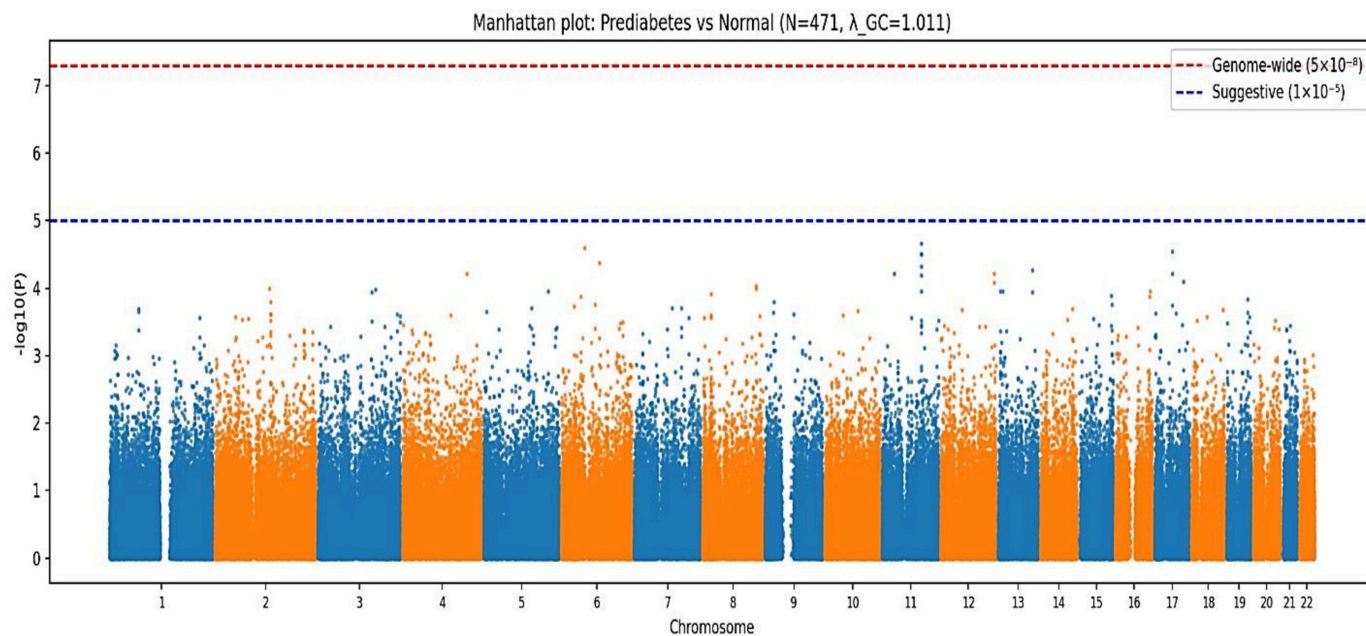


Figure 3

Manhattan plot of GWAS (prediabetes vs normal).

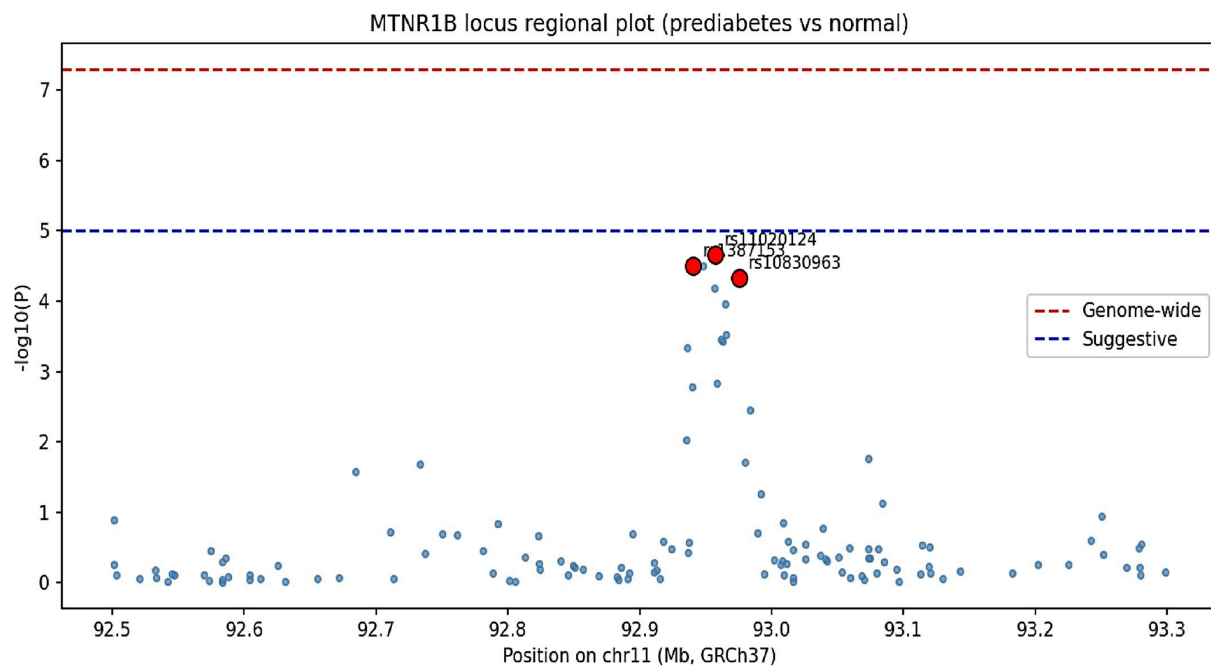


Figure 4

Regional association plot of the *MTNR1B* locus on chromosome 11 (chr11:92.5–93.3 Mb) showing the cluster of suggestive SNPs around rs10830963 / rs11020124.

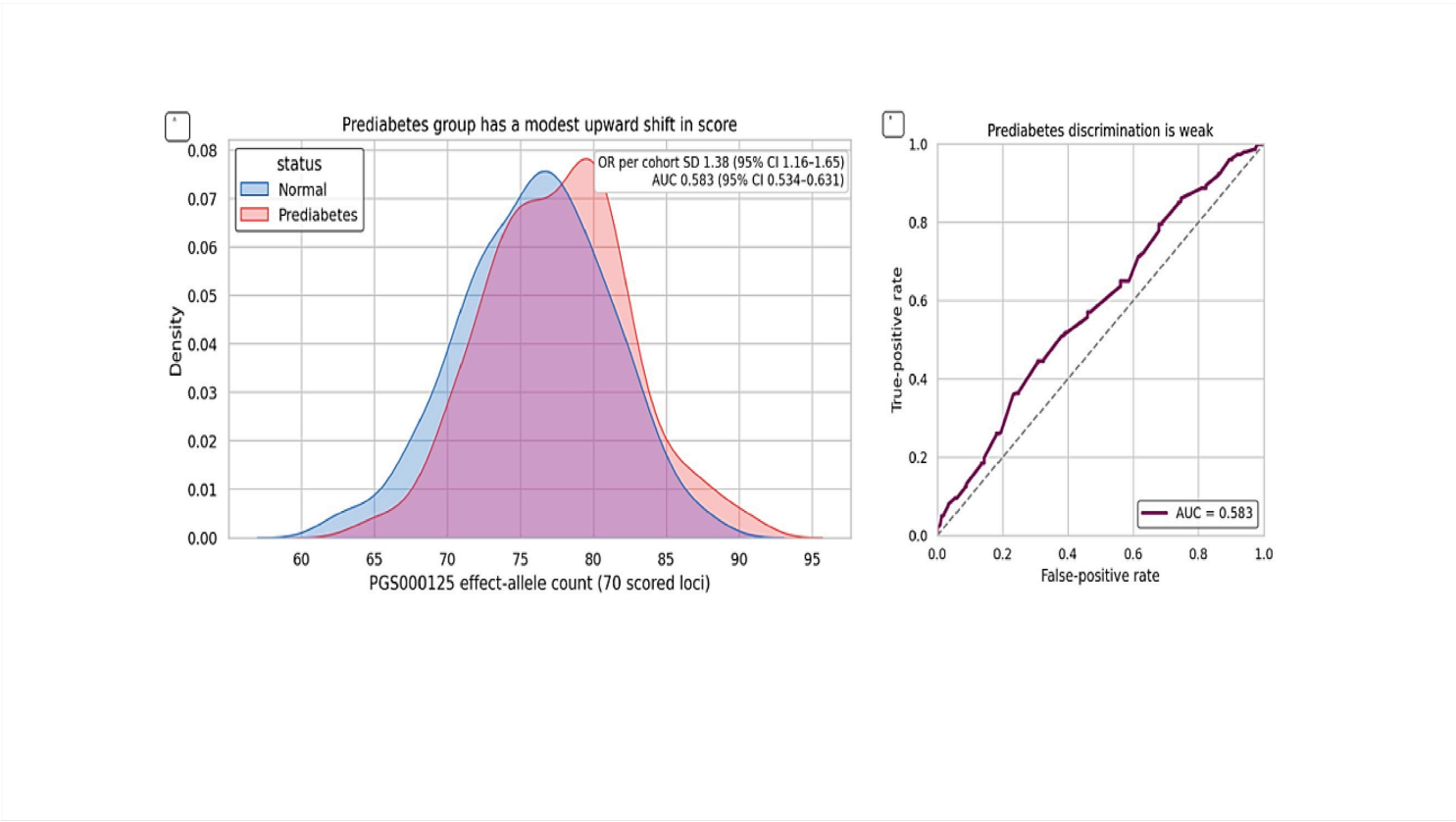


Figure 5

Exploratory PGS000125 performance in 540 archived participants. (A) Score distributions by phenotype. (B) Receiver operating characteristic curve.

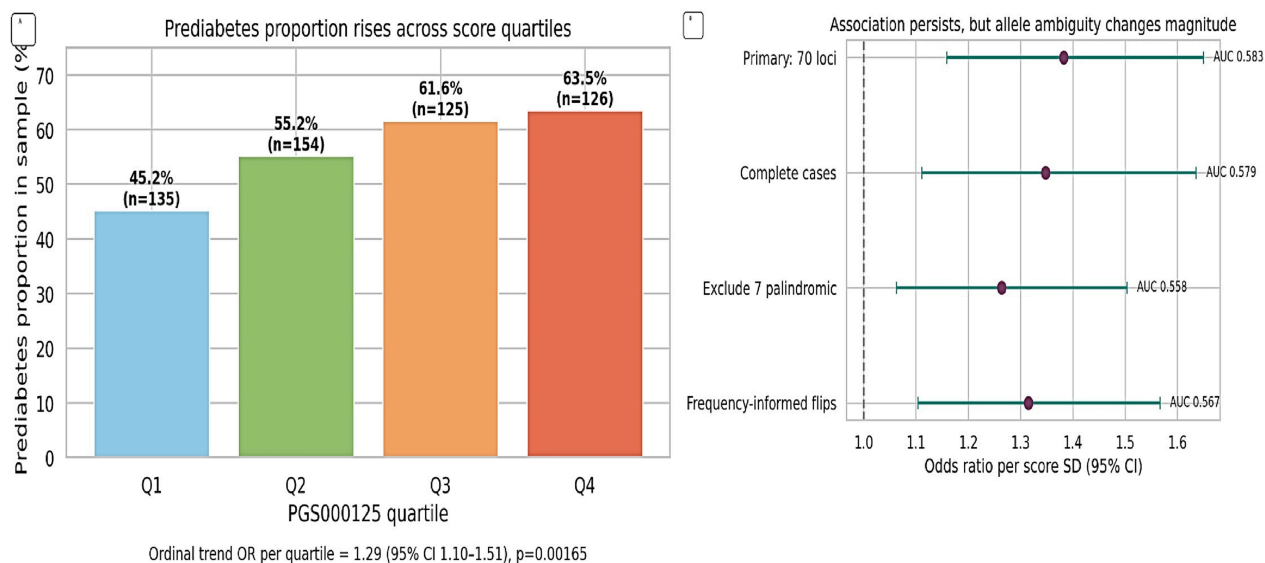


Figure 6

Dose-response and robustness of the exploratory polygenic score. (A) Observed prediabetes proportion across score quartiles. These are sample proportions, not prospective risks. (B) Odds ratios and 95% confidence intervals across sensitivity analyses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.docx](#)