https://doi.org/10.1038/s43856-025-00750-9

Excess of rare noncoding variants in several type 2 diabetes candidate genes among Asian Indian families

Check for updates

A list of authors and their affiliations appears at the end of the paper

Abstract

Background Type 2 diabetes (T2D) etiology is highly complex due to its multiple roots of origin. Polygenic risk scores (PRS) based on genome-wide association studies (GWAS) can partially explain T2D risk. Asian Indian people have up to six times higher risk of developing T2D than European people, and underlying causes of this disparity are unknown.
Methods We have performed targeted sequencing of ten T2D GWAS/candidate regions using endogamous Punjabi Sikh families and replication studies using unrelated Sikh people and families from three other Indian endogamous ethnic groups (EEGs).
Results We detect rare and ultra-rare variants (RVs) in *KCNJ11-ABCC8* and *HNF4A* (MODY genes) cosegregated with late-onset T2D. We also identify RV enrichment in two new genes, *SLC38A11* and *ANPEP*, associated with T2D. Gene-burden analysis reveals the highest RV burden contributed by *HNF4A* ($p = 0.0003$), followed by *KCNJ11/ABCC8* ($p = 0.0061$) and *SLC38A11* ($p = 0.03$). Some RVs detected in Sikh people are also found in Agarwals from Jaipur, both from Northern India, but were monomorphic in other two EEGs from South Indian people. Despite carrying a high burden of T2D and RVs, most families have a significantly lower burden of PRS. Functional studies show that an intronic regulatory variant (RV) in *ABCC8* affects the binding of Pax4 and NF-kB transcription factors, influencing downstream gene regulation.
Conclusions The high burden of T2D in these families may stem from the enrichment of noncoding RVs in a small number of major known genes (including MODY genes) with oligogenic inheritance alongside RVs from genes associated with polygenic susceptibility. These findings highlight the need to conduct deeper evaluations of families from non-European ancestries to identify potential novel therapeutics and implement preventative strategies.

Plain language summary

People with type 2 diabetes (T2D) have high levels of sugar in the blood, which can cause many health problems. T2D is a major global health issue, with Asian Indian people being up to six times more likely to develop it than European people. Although inherited factors contribute to this increased risk, they only partially explain the high T2D prevalence in Asian Indian families. In our study, we sequenced known T2D-related genes in Punjabi Sikh families. We found rare changes in inherited genes known to increase the incidence of T2D. Our findings emphasize the importance of investigating families from non-European backgrounds to identify new treatment and prevention strategies for T2D.

Type 2 diabetes mellitus (T2D) is a multifactorial metabolic disorder characterized by insulin secretory defects, insulin insufficiency, and resistance. It involves complex interactions of environmental, lifestyle, and genetic factors[1]. T2D has become one of the most severe health problems in the world as the prevalence of T2D continues to rise significantly in non-European ethnic populations. According to the International Diabetes Federation (IDF), the global prevalence of diabetes will increase from 463 million people in 2019 to 700 million by 2045, with 90 percent of these patients having T2D[2]. Due to rapid economic development and urbanization, people from developing countries bear the highest burden of this disease, accounting for more than 80% of T2D cases[3]. South Asian (SA) people from the Indian subcontinent comprise approximately one-quarter of the global population and are up to six times more likely to develop T2D than European people[4]. The IDF estimates that in 2019, there were 77 million people with diabetes in India, which is expected to rise to over 134 million by 2045[5]. Multiple studies have shown that the onset of T2D in SA people occurs about 10–12 years earlier and at lower body mass index (BMI) thresholds[4,6]. However, the underlying causes of this disparity are currently unknown and are not explained by conventional risk factors of T2D or by the currently available genetic studies[7].

The people of India have a complex population genetic history, which is further complicated by a deeply rooted caste system that has effectively

e-mail: Dharambir-sanghera@ouhsc.edu

prohibited inter-caste marriage for several generations and thus divided the population into numerous endogamous groups[8]. Recent findings of the Indian Genome Variation Consortium[9] revealed a high degree of genetic diversity between Indian ethnic groups, implying that pooling endogamous populations without accounting for ethnolinguistic factors could lead to false-positive disease associations. While genome-wide association studies (GWAS) have identified over 400 common genetic variants or regions linked to T2D, understanding their biological mechanisms and clinical implications remains limited[10–12]. Together, these loci explain modest odds ratios ranging between (1.05 to 1.30), per risk allele and, in the aggregate, account for <10% of the total genetic variance for T2D[13]. Cumulative genetic scores of common variants combined into a polygenic risk score (PRS) have successfully predicted the risk for certain cancers, T2D, and other diseases in people of European origin[14–16]. However, the lack of data on different ethnic groups and the limited transferability into other non-European ethnic groups still challenge the generalizability and clinical utility of PRS for identifying high-risk individuals[11,15,17].

Interestingly, the common variant signals we and others identified in Asian Indian T2D GWAS studies were not replicated in Sikh families. Despite having a high burden of T2D, some of these families had a very low PRS derived from extensive T2D GWAS studies[14,18]. These results prompted us to evaluate if the rare variants (RVs) underlying the GWAS regions could contribute to T2D in these families, as such data are unavailable in populations from India[19–22]. Here, we present the targeted sequencing results of ten T2D loci identified by GWAS studies in Asian Indian people[23]. The objectives were to determine (1) if RVs underlying the GWAS regions may influence T2D risk in Sikh families and validate the association in an additional dataset of 4602 (unrelated) individuals of Punjabi people and (2) to examine the extent to which they are transportable to families from other Endogamous Ethnic Groups (EEGs) of people from India.

Our study reports the enrichment of noncoding RVs in a small number of major known genes with polygenic susceptibility and genes causing maturity onset of diabetes of the young (MODY) in some families with a high burden of T2D. These families also revealed having modestly low polygenic risk score (PRS) for T2D.

Materials and methods

Study cohorts

Our study investigated 6,437 individuals comprised of 4,890 Asian Indian people from India and the US[8,24,25]; the family cohort for targeted sequencing was part of the (Asian Indian Diabetic Heart Study/Sikh Diabetes Study [AIDHS/SDS]). All study participants of AIDHS/SDS were from the Northern part of India and were recruited from 2003 to 2009[8,24,26,27]. The Sikh people, a relatively young, inbred population of ~26 million (~2% of the Indian population), are from the northwestern province of India and follow a distinct and unique religion born ~500 years ago in Punjab[8]. From the AIDHS/SDS, we chose for targeted sequencing 32 families with large pedigrees comprising 300 individuals (with an average family size of 10.5 and a family size range of 53 to 6). Sikh people are a non-smoking community, with 50% of individuals being teetotalers and life-long vegetarians[8]. Diagnosis of T2D was confirmed as described previously[27] by reviewing medical records for symptoms, medication use, and fasting glucose levels following the guidelines of the American Diabetes Association (2004)[28]. The diagnosis for normoglycemic controls was based on fasting glycemia <110 mg/dL or 2-h glucose <140 mg/dL. BMI was calculated as weight (kg)/[height (m)^2], and waist-to-hip ratio (WHR) was calculated as the ratio of abdomen or waist circumference to hip circumference. Blood pressure (BP) was measured twice after a 5-min seated rest period with the participant's feet flat on the floor. Coronary artery disease (CAD) was assigned when there was a documented prior diagnosis of heart disease, electrocardiographic evidence of angina pain, coronary angiographic evidence of severe (>50%) stenosis, or echocardiographic evidence of myocardial infarction. All blood samples were obtained at the baseline visit. Education, socioeconomic status, dietary, and physical activity data were recorded. Smoking information was collected regarding past smoking, current smoking status, and length of time, the number of cigarettes smoked/day. The individuals on lipid-lowering medication, subjects with type 1 diabetes (T1D), or those with a family member with T1D or secondary diabetes (e.g., hemochromatosis or pancreatitis) were excluded from the study, as reported earlier[26].

Our validation dataset (as part of Phase II) of the AIDHS/SDS, a case/control cohort (n = 4602; (2574 T2D cases and 2028 controls)), is primarily from the same Punjabi ancestry and originated from the exact geographic location (Punjab) where the families were collected and explained previously[8,24,26,27]. An additional 1547 individuals from different EEGs from India include 517 people from Chettiars (Tamil Nadu family diabetes study, TNFDS), 530 people from Agarwals (Jaipur family diabetes study, JFDS), and 500 people from Reddys (Nellore family diabetes study, NFDS) and the pedigree, phenotypic (i.e., T2D and its related traits) and covariate data from these individuals were collected using standardized procedures as described in detail earlier[29]. All participants provided written informed consent for investigations. The study was reviewed and approved by the University of Oklahoma Health Sciences Center's Institutional Review Board (IRB) and the Human Subject Protection Committee at the participating hospitals and institutes in India. Details of the clinical and demographic characteristics of all cohorts are described in Table 1.

Target sequencing (discovery/sequencing cohort)

Genomic DNA samples of 300 Sikh individuals (AIDHS/SDS) were sequenced using Agilent's SureSelectXT2 kits designed for targeted resequencing of ten confirmed candidate gene regions for T2D. Targeted sequencing was performed at the Oklahoma Medical Research Foundation (OMRF), Oklahoma City, Oklahoma, USA. A custom gene panel (SureSelect QXT, Agilent Technologies Inc.) was designed to investigate genetic variation in ten T2D genes reported to be associated with diabetes in people from SA (Supplementary Table 1). Genomic positions of the coding sequence of all gene regions were obtained from the Consensus Coding Sequence database (www.ncbi.nlm.nih.gov/CCDS.CcdsBrowse.cgi, Release 20). The coding regions and all exon-intron boundaries were captured using a custom capture assay (Agilent Technologies, Santa Clara, CA, USA). Samples were fragmented using a Covaris ultrasonicator and prepared for sequencing on an Illumina 3000HiSeq (20–30X) using a custom DNA library preparation protocol described earlier[30,31]

Library production, targeted capture, sequencing

Genomic DNA was extracted from whole blood or buffy coats using Qiagen kits (Qiagen, Chatsworth, CA, USA) or salting-out procedures described previously[27,32,33]. 1 μg of genomic DNA was sent to the Oklahoma Medical Research Foundation (OMRF) for sequencing after target capture using Agilent XT2 capture technology. The quality and integrity of the DNA were assessed using Agilent's Analyzer and Tape Station reagents before target capture and library preparation. Library construction and custom capture were automated using the Perkin-Elmer Janus II system in a 96-well plate format. The purified DNA underwent a series of shotgun library construction steps, which included fragmentation via acoustic sonication (Covaris), end-polishing, A-tailing, ligation of sequencing adapters, and PCR amplification with eight-base pair barcodes for multiplexing. For the capture of libraries, Roche/Nimblegen SeqCap EZ custom-designed probes were employed. Before sequencing, the library concentration was quantified using triplicate qPCR, and molecular weight distributions were verified with the Agilent Bioanalyzer. Barcoded libraries were pooled using liquid handling robotics before clustering on the Illumina cBot and loading onto the sequencer. Massively parallel sequencing-by-synthesis was performed using fluorescently labeled, reversibly terminating nucleotides on the HiSeq sequencer allowing for high-throughput analysis of the DNA libraries as described earlier[30].

Read processing, quality control (QC) analysis

We used Genome Analysis ToolKit [GATK], Picard, Burrows-Wheeler Alignment (BWA), and SAMTools for base calling, sequence alignment, realignment, duplicate removal, quality recalibration, data merging, variant detection, and annotation as previously described[30]. The BAM files were aligned to a human reference using BWA Aligner (v0.6.2). Read data from a flow cell lane is treated independently for alignment and QC purposes in instances where the merging of data from multiple lanes is required (e.g., for sample multiplexing). The samples were sequenced using paired-end ~140 to 150 bp reads, and the insert sizes were at least 100 bp in length. Therefore, we expected to see ~240 to 250 bp on the Bioanalyzer. Read pairs not mapping within ±2 standard deviations of the average library size (~150 ± 15 bp for the targeted region) were removed. Lanes/samples that failed QC were flagged in the system and could be re-queued for library prep (<5% failure) or further sequencing. Duplicate reads were removed using Picard Mark Duplicates; v1.70, indel realignment was performed using GATK Indel Realigner; v1.6-11-g3b2fab9, and base qualities were recalibrated using GATK Table Recalibration; v1.6-11-g3b2fab9 as described[30].

Table 1 | Clinical characteristics of the study groups

Trait	Discovery		Sikh replication		EEGs replication					
	Discovery (*N* = 288)		AIDHS/SDS (*N* = 4602)		TNFDS (*N* = 517)		NFDS (*N* = 500)		JFDS (*N* = 530)	
	T2D	Controls	T2D	Controls	T2D	Controls	T2D	Controls	T2D	Controls
Age (years)	54.2 ± 12.6	45 ± 14.7	54.9 ± 11.4	48.6 ± 14.5	57.2 ± 13.2	42.3 ± 15.6	54.2 ± 14.3	41.3 ± 15.2	55.9 ± 12.5	36.1 ± 13.8
SYSBP (mmHg)	151.9 ± 24.3	135.4 ± 23.8	144.4 ± 24.1	132.7 ± 31.3	139.7 ± 20.2	129.2 ± 16.8	132.6 ± 18.7	122.2 ± 17.3	133.6 ± 19.6	120.8 ± 15.9
DBP (mmHg)	87 ± 11.3	82 ± 11.7	84.7 ± 12.2	81.1 ± 12.3	84.4 ± 12.4	82 ± 10.7	82.7 ± 10.7	79.6 ± 12.7	85.2 ± 11.4	80.8 ± 10
Waist circumference (cm)	96.1 ± 10.9	91 ± 10.9	94.4 ± 12.1	90.7 ± 12.3	100.4 ± 12.8	96.8 ± 14	98.3 ± 13.1	93 ± 13.5	97.9 ± 12.4	90.7 ± 11.5
Hip circumference (cm)	98.4 ± 9.3	96.1 ± 8.8	98.2 ± 10.6	98.0 ± 10.3	98 ± 12.5	93.3 ± 13.4	101.8 ± 12.9	99.7 ± 13.1	100.6 ± 9.2	99.8 ± 9.9
Height (cm)	163.2 ± 9.6	163.9 ± 9.4	161.8 ± 10.3	163.3 ± 10.3	159.8 ± 8.5	160 ± 9.4	161.9 ± 8.8	159.7 ± 9	161.8 ± 9.1	165.1 ± 8.8
Weight (kg)	75.5 ± 14.9	72.9 ± 14.1	71.4 ± 14.1	69.9 ± 14.3	69.9 ± 12.4	67.3 ± 14.7	71.1 ± 16.2	66.6 ± 14.6	74.5 ± 12.9	71.9 ± 14.6
BMI (kg/m^2)	28.3 ± 4.8	27.2 ± 4.6	27.3 ± 4.9	26.2 ± 4.7	27.8 ± 4.9	26.3 ± 5.4	27.1 ± 5.9	26.1 ± 5.3	27.8 ± 5.2	26.2 ± 4.9
Fasting blood glucose (mg/dL)	174.9 ± 61.5	99 ± 10.4	175.9 ± 71.9	96.5 ± 13.8	173.8 ± 65.9	102.2 ± 11.8	149.8 ± 56	93.8 ± 11.2	147.3 ± 31.2	101.7 ± 8.2

Values are in mean ± SD.
SYSBP systolic blood pressure, *DBP* diastolic blood pressure, *BMI* body mass index, *T2D* type 2 diabetes, *AIDHS/SDS* Asian Indian Diabetes Heart Study/Sikh Diabetes Study, *EEGs* Endogamous Ethnic Groups, *TNFDS* Tamil Nadu family diabetes study, *NFDS* Nellore family diabetes study, *JFDS* Jaipur family diabetes study.

Variant detection and annotation

All sequence data underwent further QC protocol before performing annotation and further processing. We assessed total reads, the ratio of unique reads to total reads mapped to the target, capture efficiency ratio of reads mapped to humans versus reads mapped to the target, the reads with selected coverage distribution <80% at 20X were not included. We also assessed capture uniformity and raw error rates. The Transition/Transversion ratio (Ti/Tv) was chosen to be 3 for known sites and ~2.5 for novel sites as detailed previously[30,31,34].

Variant detection and genotyping were performed using the UnifiedGenotyper (UG) tool from GATK (v1.6-11-g3b2fab9). Variant data for each sample were formatted (variant call format [VCF]) and poor calls were flagged using the filtration walker (GATK) to mark sites that were of lower quality/false positives including low-quality scores (Q50), allelic imbalance (ABHet 0.75), long homopolymer runs (HRun > 3) and low quality by read depth (QD < 5).

We used an automated pipeline to annotate variants derived from targeted sequencing data, the SeattleSeq Annotation Server (http://gvs.gs.washington.edu/). These publically accessible servers return annotations, including dbSNP and GnomAD rsID (or whether the coding variant is novel), gene names and accession numbers, predicted functional effect (e.g., splice-site, nonsynonymous, missense, etc.), protein positions and amino acid changes, PolyPhen predictions, conservation scores (e.g., PhastCons, GERP), ancestral allele, dbSNP allele frequencies, and known clinical associations. Based on the minor allele frequency (MAF) reported in gnomAD for the SA population, we defined RVs as ultra-rare, MAF < 0.1%; rare, 0.1% to <1%; low-frequency, 1% to <5%; and common variants, MAF ≥ 5% (Supplementary Table 3).

Replication studies

AIDHS/SDS replication cohort (unrelated). AIDHS/SDS cohorts were genotyped on Illumina using a Human 660 W Quad BeadChip panel; Illumina's Global Screening Arrays with multi-disease content (GSA +); and GSA (Illumina, Inc., San Diego, CA) as explained in Saxena et al.[8,24]. We performed pairwise identity-by-state clustering in PLINK across all individuals to assess population stratification, and outliers were removed. Related individuals with pi-hat 0.3 and samples with <93% call rate were excluded, as were SNPs with call rate <95%. Also excluded were SNPs with HWE $p < 10^{-6}$ or MAF < 1%, as described previously[8,25]. To increase genome coverage, imputation was performed using Minimac4[35] (https://imputationserver.sph.umich.edu/) with 1000G Phase 3v5 multi-ethnic reference panel in NCBI Build 37 (hg19) coordinates as described[24,25]. Quality control for the imputed SNPs included the removal of variants with an imputation certainty info score <0.8, and SNPs significantly deviated from HWE in controls ($p < 1 \times 10^{-6}$). Replication studies for RV were performed in all available variants with MAF < 1% and excluding variants with poor info score (<0.8).

Endogamous ethnic groups (EEGs)

The other replication cohort comprised 1547 individuals from families from three different EEGs of people from India. These families were recruited for genetic studies for T2D as part of the joint partnership of the US National Institutes of Health and the Indian Council of Medical Research for performing genetics studies of T2D. These EEG people include the TNFDS; $n = 517$, Chennai, Tamil Nadu (South India); the NFDS; $n = 500$, Nellore, Andhra Pradesh (South India); and the JFDS; $n = 530$, Jaipur, Rajasthan (Northwest India). The recruitment and genetic research activities for these cohorts were initiated after obtaining IRB approvals from the US and Indian institutes for the protection of human subjects as well as institution-specific collaborative research policies of the US and India project number 55/6/2/Indo-US/ 2014-NCD-II jointly funded by National Institutes of Health and the Indian Council of Medical Research. All participants provided informed consent for participating in research studies as described previously[29].

Probands were recruited randomly from the T2D registries, and at least a 3–4 generation pedigree was drawn for each family. Anthropometric data were collected, including weight, height, waist, and hip circumferences (i.e., an average of three values collected for a given trait). BMI was measured as weight (kg) divided by height squared (m^2). Systolic and diastolic BP and heart rate (i.e., average of three values collected for a given trait) were measured. Fasting (at least an 8-h overnight fast) and post-prandial capillary blood glucose levels were measured at the study sites, as described earlier[29]. All the individuals were recorded for physical activity, disease-related history, and food habits. Biochemical phenotyping (HbA1c, lipid profile, hs-CRP, blood/urine creatinine, and fasting insulin) was also performed, followed by the genetic analysis.

For replication studies, EEG families were not available with genome-wide genotypes but only selected variants could be genotyped due to funding restrictions. The genotyping was performed at MedGenome Inc. using the Fluidigm technology in India. Of these 132 SNPs, 18 markers were found to be monomorphic in all three EEGs of people, and six markers with poor call rate or quality were omitted. After excluding 12 of the 118 SNPs that could not be designed for Fluidigm. The genotypic data were cleaned for Mendelian inconsistencies by blanking the genotypes in error using the PLINK program, and 106 variants were available for replication analysis. The markers that required less than 5% blanking were retained for the analysis. In addition, the program PEDSYS was used to infer missing genotypes from the genotypes of relatives, when possible, without any ambiguity. Using the SOLAR program, the allele frequencies and minor allele counts were calculated for each of the three EEGs, and SNPs were tested for HWE.

Functional studies using DNA in a pull-down assay

We amplified the region surrounding the (rs117727754 C>G) variant found in the *ABCC8* gene (located at Chr 11:17415057 position) by PCR. Using 5′-biotin-labeled forward primer (AGTAACGCCAGCCTAACATATAA), the region of *ABCC8* containing the RV was immobilized to streptavidin beads. Nuclear fraction (100 μg) was extracted from HCT-116 cells using NE-PER™ Nuclear and Cytoplasmic Extraction Reagents Kit. Proteins in the nuclear fraction were incubated with 5′-biotinylated DNA probes (1 μg) and the streptavidin beads on a rotating shaker at 4 °C for 2–4 h with gentle rocking. Following this incubation, tubes were placed on a magnet stand, and after 3 min of incubation at room temperature, the supernatant was removed. Lastly, the beads were washed three times with cold PBS. After the last wash, the pull-down mixture was resuspended in a 4X Laemmli buffer. The eluted transcription factors (TFs) from the beads were detected by Western blot analysis using monoclonal antibodies of Pax4 (1:500) (Invitrogen, MA, USA; catalog # PA1-108) and Nf-kβ (1:1000) (Cell Signaling Technology, MA, USA; catalog # D14E12) as described previously[36]. The band intensity was quantified using ImageJ software[37].

Statistical analysis and reproducibility

RVs were screened from the ten targeted gene regions. They were looked up for detection in the Punjabi Sikh people replication cohort ($N = 4602$), which is available with genome-wide genotypes and imputations, as discussed. Logistic regression was used to test the association of T2D with each RV in the replication cohort with adjustment for covariates age, gender, and BMI using SPSS software version 29. In families, we performed likelihood-based probability statistics and analyzed RV association with T2D in each family using a 2 × 2 chi-square in R. To further assess the cumulative burden of RVs on T2D, we performed gene/locus-centric analysis using combined multivariant and collapsing (CMC) methods with the Hotelling T2 test and CMC with regression tests[38]. We used the variance-component-based sequence kernel association test (SKAT) that tests for association by evaluating the distribution of genetic effects for a group of variants instead of aggregating variants[39]. We then tested the association of this RV in our unrelated case-control as part of replication studies adjusting for age, gender, and BMI. The power of RV association with T2D for the replication cohort was calculated using the Genetic Power Calculator[40]. These analyses were performed using SVS version 8.9.1 (Golden Helix, Bozeman, MT, USA) as described previously[30,41].

Multiple bioinformatics tools were used to assess the pathogenicity of the identified missense mutations in the present study, including HOPE[42] and varCards[43], to predict the functional consequences of variants through 23 in silico predictive algorithms. The variants with a damaging score of four of 23 defined by these algorithms were considered potentially damaging. The 3D models for wildtype and deleterious nonsynonymous missense RVs were generated using the Swissmodel[44]. Multiple sequence alignment was performed to understand the evolutionary conservation of the amino acid residue across species using Clustal omega[45]. The allele frequencies of RVs among other populations were taken from GnomAD v2.1.1[46].

To explore the role of RVs in the intronic or splice regions in affecting gene regulation, we used various bioinformatics tools. The effect of these RVs on TF binding sites (TFBS) was analyzed using the atSNP search[47]. The atSNP Search is a comprehensive web database that evaluates motif matches to the human genome with both reference and variant alleles and assesses the overall significance of the variant alterations on the motif matches. It also enables convenient interpretation of regulatory roles of noncoding genetic variants by statistical significance testing and composite logo plots, which are graphical representations of motif matches with the reference and variant alleles. The pathogenic impact of intronic or splice-site RVs was predicted using RegSNP-intron software. The algorithm is based on a supervised learning random forest classifier that integrates RNA splicing, protein structure, and evolutionary conservation features[48]. Chromatin loops (or chromatin interactions) are important elements of chromatin structures, and their disruption can lead to many diseases. ChromLoops is a comprehensive, multispecies, and specific protein-mediated chromatin loop database that integrated ChIA-PET, HiChIP, and PLAC-Seq datasets from 13 species and documented high-quality chromatin loops. This database identifies genes with high-frequency chromatin interactions in the collected species[49].

Genome-wide PRS

We constructed an Asian Indian ancestry-specific PRS (PRS_{AI}) for T2D using candidate variants derived from genome-wide genotypes derived from our AIDHS/SDS people and variants from people from other SA studies[8,24,25,50,51]. We also built a European PRS (PRS_{EU}) using summary statistics from GWAS meta-analyses from people of seven European cohorts[52].

The selection criteria of SNPs from 46,985,978 (common and rare) for constructing Sikh-specific PRS_{AI} were used as described previously[41,53]. Briefly, we selected all significant SNPs ($p < 10^{-2}$) following the regression analysis for T2D. After excluding INDELs, duplicate and multiallelic SNPs and SNPs with info score <0.80 and including SNPs with MAF > 0.01 and MAF < 0.45, SNPs with $p < 10^{-4}$ were chosen. After linkage disequilibrium (LD) pruning using R^2 (LD) = 0.50, a total of 2921 significant SNPs were used for the construction of the PRS_{AI}[54]. The individual-level regression coefficients were multiplied by the number of risk alleles to compute the PRS in training and test sets as described previously[41,53,55]. The weighted PRS was

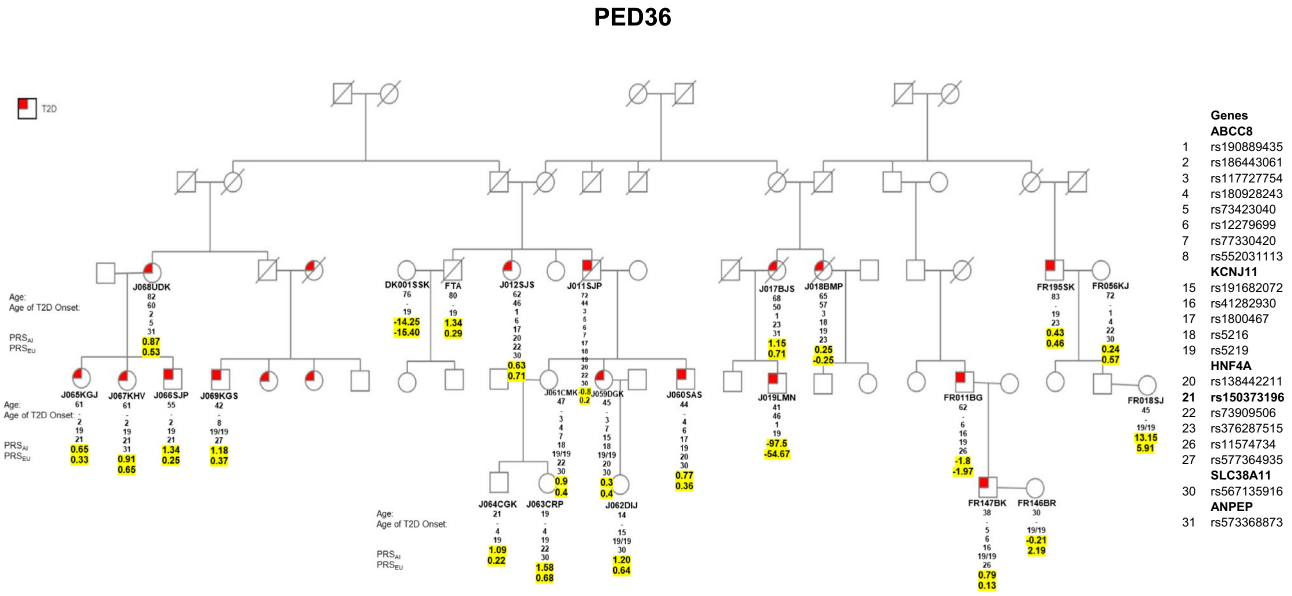

PED36
T2D
Age:
Age of T2D Onset:
PRS$_{AI}$
PRS$_{EU}$
J068UDK 82 60 2 5 31 0.87 0.53
J065KGJ 61 - 2 19 21 0.65 0.33
J067KHV 61 - 2 19 21 31 0.91 0.65
J066SJP 55 - 2 19 21 1.34 0.25
J069KGS 42 - 8 19/19 27 1.18 0.37
DK001SSK 76 - 19 -14.25 -15.40
FTA 80 - 19 1.34 0.29
J012SJS 62 46 1 6 17 20 22 30 0.63 0.71
J011SJP 72 44 3 5 6 7 17 18 19 20 22 30 -0.8 0.2
J061CMK 47 - 3 4 7 18 19/19 22 30 0.9 0.4
J059DGK 45 - 3 7 15 18 19/19 20 30 0.3 0.4
J060SAS 44 - 4 6 17 19 20 30 0.77 0.36
J064CGK 21 - 4 19 1.09 0.22
J063CRP 19 - 4 19 22 30 1.58 0.68
J062DIJ 14 - 15 19/19 30 1.20 0.64
J017BJS 68 50 1 23 31 1.15 0.71
J018BMP 65 57 3 18 19 23 0.25 -0.25
J019LMN 41 46 1 19 -97.5 -54.67
FR011BG 62 - 6 16 19 26 -1.8 -1.97
FR147BK 38 - 5 6 16 19/19 26 0.79 0.13
FR146BR 30 - 19/19 -0.21 2.19
FR195SK 83 - 19 23 0.43 0.46
FR056KJ 72 - 1 4 22 30 0.24 0.57
FR018SJ 45 - 19/19 13.15 5.91
Genes
ABCC8
1 rs190889435
2 rs186443061
3 rs117727754
4 rs180928243
5 rs73423040
6 rs12279699
7 rs77330420
8 rs552031113
KCNJ11
15 rs191682072
16 rs41282930
17 rs1800467
18 rs5216
19 rs5219
HNF4A
20 rs138442211
21 rs150373196
22 rs73909506
23 rs376287515
26 rs11574734
27 rs577364935
SLC38A11
30 rs567135916
ANPEP
31 rs573368873

Fig. 1 | The burden of rare variants of multiple MODY genes (*ABCC8*, *KCNJ11*, and *HNF4A*) in Ped 36 of Sikh families. Highlighted fonts depict polygenic risk scores using Asian Indian (PRS$_{AI}$) and European (PRS$_{EU}$) models. Details of the rare variants from *ABCC8, KCNJ11, HNF4A, SLC38A11*, and *ANPEP* detected in this family are listed on the right side. Detection of heterozygous variant alleles in each family member is indicated by the numerical code for the listed SNPs under the individual's ID.

calculated using the following equation 1:

$$PRS_j = \sum_{i}^{N} \beta_i * dosage_{ij}$$

where N is the number of SNPs in the score, β_i is the effect size (or beta) of variant i, and dosage$_{ij}$ is the number of copies of SNP in the genotype of individual j[56]. To construct PRS$_{EU}$, we used the summary statistics data from O'Connor et al.[57], which comprised datasets from people of seven European cohorts ($n = 312,646$) containing 33,122,978 variants and available for both additive and recessive models. The SNP information of these seven cohorts was obtained from Supplementary information from O'Connor et al.[57], available online. We used similar SNP selection criteria and QC for constructing PRS$_{EU}$ as described above for PRS$_{AI}$ and selected a total of 1847 significant SNPs for the construction of the PRS$_{EU}$. Next, we compared the distribution of PRS$_{AI}$ and PRS$_{EU}$ in T2D and controls between the entire pedigree cohort of 32 families (discovery) in selected five pedigrees carrying a high burden of rare and ultra RVs and in unrelated cases and controls ($n = 4602$). Significant differences were analyzed using a two-tailed t-test.

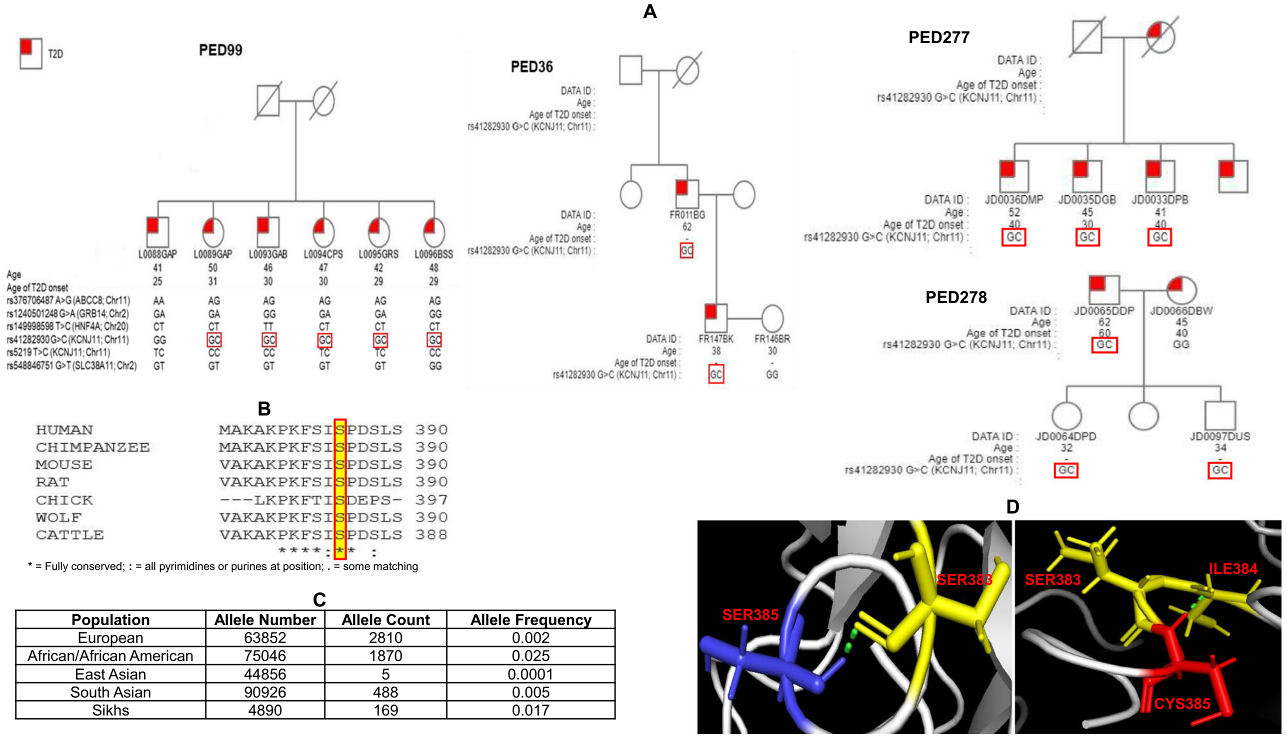
Results

Punjabi Sikh discovery and replication analysis

Of 6437 individuals studied, targeted sequencing of candidate genes for ten T2D loci (harboring 48 genes and intergenic regions) was performed on 300 subjects, and validation studies included 6137 subjects. As reported previously, these participants were part of the AIDHS/SDS and INDIGENIUS consortium[29,31,32]. Details of candidate gene regions selected for target sequencing are provided in Supplementary Table 1. Of the 300 Sikh individuals who underwent targeted sequencing, 288 remained after sequencing QC. To identify RVs associations spanning ethnicities, we extended the replication of 118 SNPs in the other three EEG people (543 cases and 1004 controls) by denovo genotyping. Our study detected several RVs cosegregating with the T2D phenotype in some families. Many RVs were found in genes associated with T2D and known MODY genes, such as the ATP-sensitive inward rectifier potassium channel 11 gene (*KCNJ11*, MODY-13) on chromosome 11, the sulfonylurea receptor gene (*ABCC8*, MODY-12) also on chromosome 11, and the hepatocyte nuclear factor 4 alpha gene (*HNF4A*, MODY-1) on chromosome 20.

Our study identified both coding and noncoding RVs matching the patient's phenotype, and some of the RVs were in the highly conserved coding regions. The MAF of some of these RVs differed among Sikhs, SA, and European people (Supplementary Table 3). Strong evidence for multiple ultra RVs association with T2D was found in the segregation of *HNF4A* variants in Ped 36 of AIDHS/SDS showing different RVs in different sibships (Fig. 1). All carriers of HNF4A (rs150373196; G>A) had T2D, and this variant was only found in Ped 36 (Fig. 1). Despite being an intronic variant, this variant was predicted to be damaging. The additional 5 of 9 (56%) carriers of this variant in our unrelated people of Punjabi cohort were diabetic (Supplementary Table 4). A rare missense variant in the *KCNJ11* (rs41282930; Ser385Cys) was found in four families: Ped 99, 36, 277, and 278. The two non-T2D carriers of this variant, both in Ped 278, had impaired glucose tolerance (Fig. 2A); these individuals were ages 32 and 34 in a pedigree with the age of onset ranging from 40 to 60 years, suggesting that these two may yet be diagnosed with T2D. An ultra-RV in the *ABCC8* gene (rs376706487; A>G) appeared to be in complete linkage disequilibrium (LD) with the *KCNJ11* (rs41282930 (Ser385Cys)) (D' = 1, $R^2 = 0.99$) in this and two other families (Supplementary Fig. 1). Additionally, 41 unrelated individuals carried the *ABCC8* rs376706487 variant, and 63% of carriers had T2D. While 78.5% of compound heterozygous carriers of rs376706487 (*ABCC8* A>G) and rs41282930 (*KCNJ11* C>T) were diabetic. In Ped 99, all individuals who were carriers of *ABCC8* rs376706487 also carried the mutant allele of the *KCNJ11* (rs5219; Lys23Glu), and the majority of members had an RV from the 3'UTR (rs149998598; C>T) encoding *HNF4A* (Supplementary Fig. 2A); indicating the RV enrichment from multiple MODY genes. We also detected a high prevalence of *ABCC8* rs376706487 in people of Agarwal families from Jaipur. Of the 24 carriers from three families, 7 of the 10 (70%) parent carriers and one young offspring (24 years) had T2D; the remaining rare allele carriers were below 40 or younger offspring without T2D phenotype (Supplementary Fig. 3).

Unlike the monogenic inheritance of MODY diabetes, some Sikh families showed a load of variants from multiple MODY genes, the most extreme being Ped 36 (AIDHS/SDS), with 31 RVs detected in multiple genes, and the majority of these were noncoding RVs detected in untranslated regions or introns (Fig. 1). Different families showed different RV loads. Highest correlation of RV with T2D was observed in *HNF4A* in Ped 36 ($r = 0.59$; $p = 0.002$) followed by *ABCC8* ($r = 0.46$; $p = 0.01$) and


A
T2D
PED99
L0088GAP L0089GAP L0093GAB L0094CPS L0095GRS L0096BSS
Age 41 50 46 47 42 48
Age of T2D onset 25 31 30 30 29 29
rs376706487 A>G (ABCC8; Chr11) AA AG AG AG AG AG
rs1240501248 G>A (GRB14; Chr2) GA GA GG GA GA GG
rs149998598 T>C (HNF4A; Chr20) CT CT TT CT CT CT
rs41282930 G>C (KCNJ11; Chr11) GG GC GC GC GC GC
rs5219 T>C (KCNJ11; Chr11) TC CC CC TC TC CC
rs548846751 G>T (SLC38A11; Chr2) GT GT GT GT GT GG
PED36
DATA ID :
Age :
Age of T2D onset :
rs41282930 G>C (KCNJ11; Chr11) :
FR011BG 62 - GC
FR147BK 38 - GC
FR146BR 30 - GG
PED277
JD0036DMP 52 40 GC
JD0035DGB 45 30 GC
JD0033DPB 41 40 GC
PED278
JD0065DDP 62 60 GC
JD0066DBW 45 40 GG
JD0064DPD 32 - GC
JD0097DUS 34 - GC
B
HUMAN MAKAKPKFSISPDSLS 390
CHIMPANZEE MAKAKPKFSISPDSLS 390
MOUSE VAKAKPKFSISPDSLS 390
RAT VAKAKPKFSISPDSLS 390
CHICK ---LKPKFTISDEPS- 397
WOLF VAKAKPKFSISPDSLS 390
CATTLE VAKAKPKFSISPDSLS 388
****:.** :
* = Fully conserved; : = all pyrimidines or purines at position; . = some matching
C
D
SER385
SER383
SER383
ILE384
CYS385

Population	Allele Number	Allele Count	Allele Frequency
European	63852	2810	0.002
African/African American	75046	1870	0.025
East Asian	44856	5	0.0001
South Asian	90926	488	0.005
Sikhs	4890	169	0.017

Fig. 2 | Segregation and in silico functional characterization of a missense RV in a MODY gene (*KCNJ11*) in four Sikh families. A Pedigrees of Sikh families show the representation of a new rare damaging variant rs41282930 (Ser385Cys) in *KCNJ11* gene segregating with T2D; mutant carriers of this variant are shown in red squares; **B** Sequence alignment reveals absolute conservation of rs41282930 at position Ser385Cys of the *KCNJ11* gene across species; **C** Allele frequency of Ser385Cys among different populations; **D** The wildtype residue (blue) forms hydrogen bonds with Serine at position 383. However, the mutant residue (red) Cysteine forms hydrogen bonds not only with Serine 383 but also with Isoleucine at position 384.

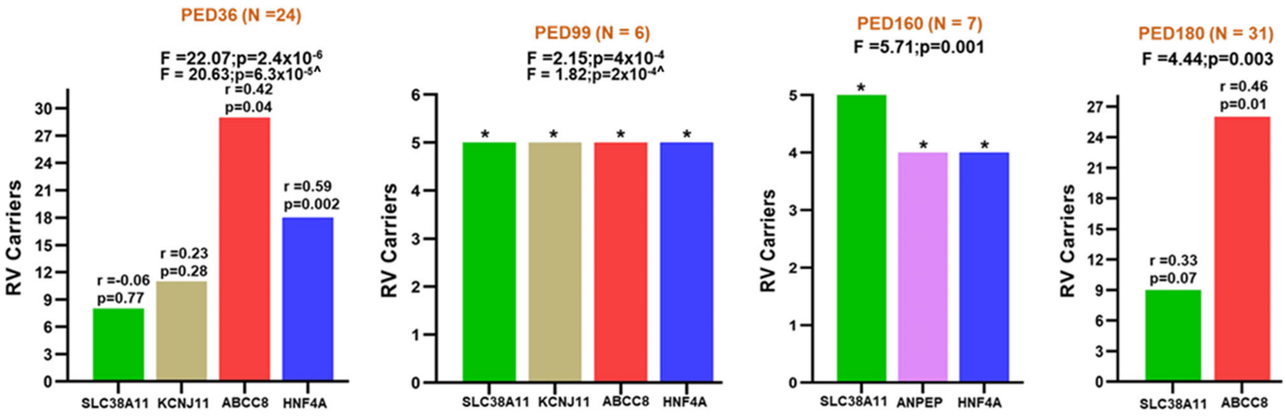


Fig. 3 | Bar graph representing the number of rare variant carriers in each family. The correlation coefficient (*r*) shows the correlation of rare variants with T2D. *(*r*) and *p*-value could not be determined due to the absence of controls in these pedigrees. ^Rare variant burden calculated excluding variants rs5216 and rs376706487 in PED36 and PED99.

($r = 0.42$, $p = 0.04$) in families180 and 36, respectively (Fig. 3). The enrichment of RVs in MODY and other genes was also observed in Ped 99, but their correlation with T2D could not be computed because none of the participants was nondiabetic. We hypothesize that the number of RVs in the known MODY genes, rather than specific RVs, may increase the burden for T2D in these families. Contrastingly, Ped 332, with 12 living members affected with T2D, showed no variant in the MODY genes associated with T2D. However, the T2D-associated C allele in a known common variant (rs5219/ Lys23Glu) did show suggestive evidence of association with T2D in this family (with %T2D in carriers: 100%, LOD score 0.52 at $p = 0.001$) (Fig. 4). RVs rs373044286 in *HK1*, rs148713144 in *DDX50*, and rs148567888 in *ANPEP* were also segregated in this family; however, none of these variants showed evidence for association with T2D.

Lastly, several RVs in the solute carrier family 38 members 11 (*SLC38A11*) correlated with T2D in these pedigrees (i.e., AIDHS/SDS). A RV (rs548846751, G>T) was found in pedigrees 99 and 160 (Supplementary Fig. 2), and two new missense variants, rs140708593 (Arg116Gln) and rs139693718 (Phe47Ser) in six members of Ped 180, and 100% of carriers had T2D (Fig. 5). These individuals also had RV burdens from *ABCC8*, *HNF4A*, and the common variant rs5219 (Lys23Glu) of the *KCNJ11* gene. The new missense variants detected in *SLC38A11* were prevalent in these families and a few sporadic cases of this Punjabi population. In total, 28 RVs were associated with an increased risk for T2D in first-degree relatives, showing incomplete penetrance in other sibships (Figs. 1, 4, Supplementary Fig. 2, Supplementary Table 5). Of the 28 variants, 20 RVs were from MODY genes.

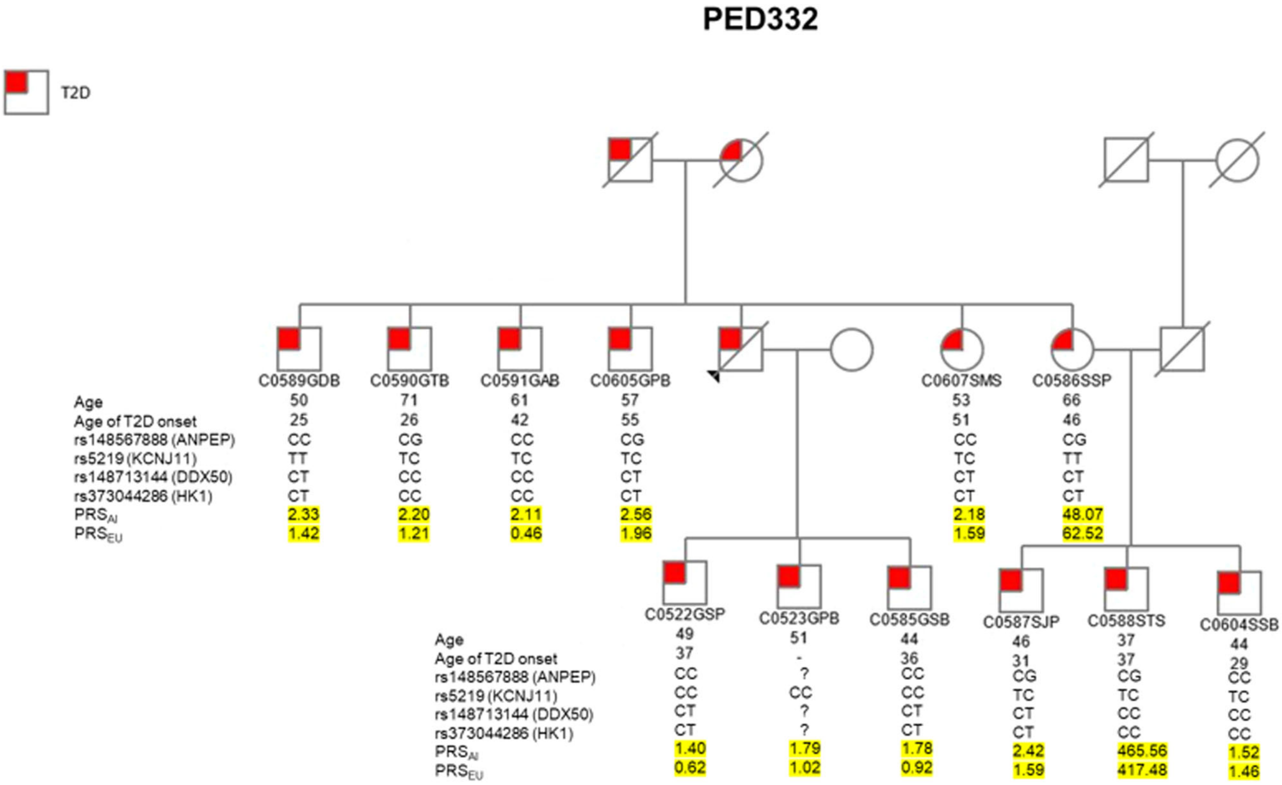

PED332
T2D
C0589GDB
C0590GTB
C0591GAB
C0605GPB
C0607SMS
C0586SSP
Age
Age of T2D onset
rs148567888 (ANPEP)
rs5219 (KCNJ11)
rs148713144 (DDX50)
rs373044286 (HK1)
PRS_AI
PRS_EU
C0589GDB: 50, 25, CC, TT, CT, CT, 2.33, 1.42
C0590GTB: 71, 26, CG, TC, CC, CC, 2.20, 1.21
C0591GAB: 61, 42, CC, TC, CC, CC, 2.11, 0.46
C0605GPB: 57, 55, CG, TC, CT, CT, 2.56, 1.96
C0607SMS: 53, 51, CC, TC, CT, CT, 2.18, 1.59
C0586SSP: 66, 46, CG, TT, CT, CT, 48.07, 62.52
C0522GSP
C0523GPB
C0585GSB
C0587SJP
C0588STS
C0604SSB
Age
Age of T2D onset
rs148567888 (ANPEP)
rs5219 (KCNJ11)
rs148713144 (DDX50)
rs373044286 (HK1)
PRS_AI
PRS_EU
C0522GSP: 49, 37, CC, CC, CT, CT, 1.40, 0.62
C0523GPB: 51, -, ?, CC, ?, ?, 1.79, 1.02
C0585GSB: 44, 36, CC, CC, CT, CT, 1.78, 0.92
C0587SJP: 46, 31, CG, TC, CT, CT, 2.42, 1.59
C0588STS: 37, 37, CG, TC, CC, CC, 465.56, 417.48
C0604SSB: 44, 29, CC, TC, CC, CC, 1.52, 1.46

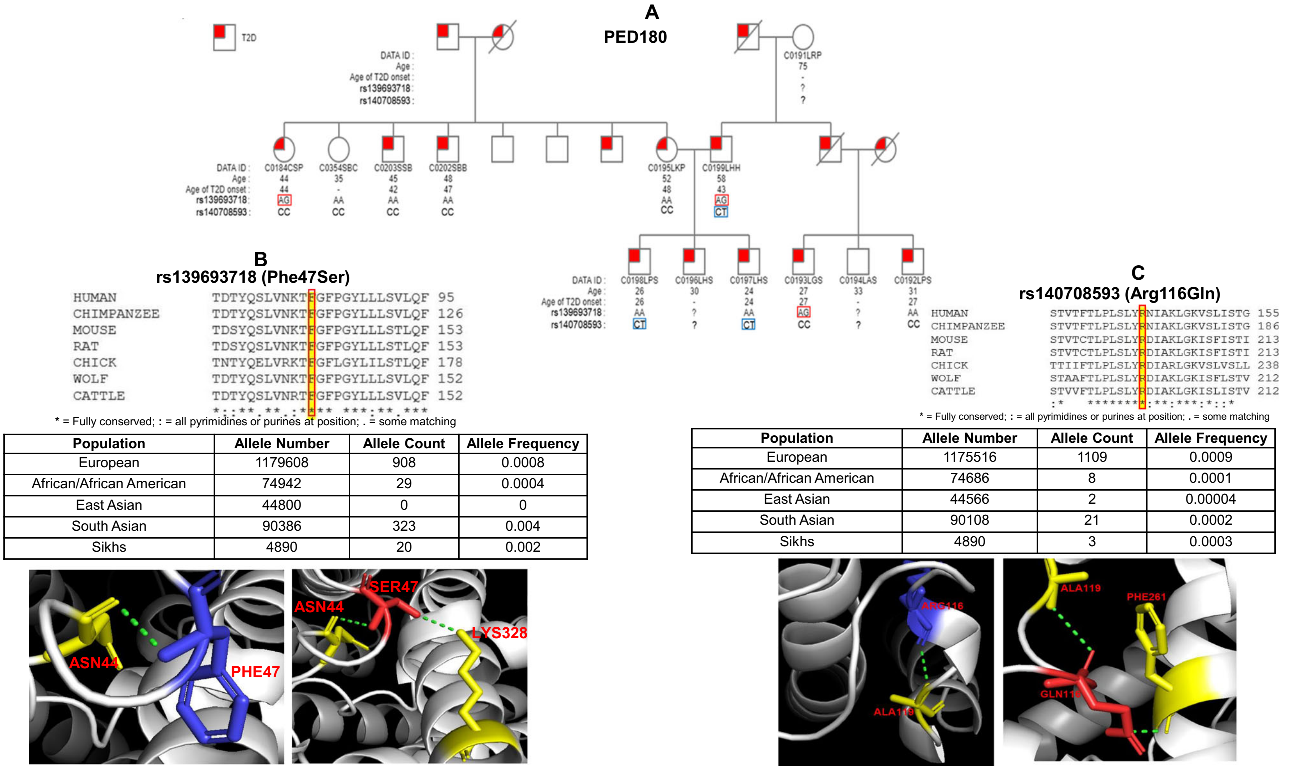
Fig. 4 | A Sikh family (PED332) showing rare variants burden in *ANPEP*, *DDX50*, and *HK1* genes. Despite having a high load of T2D, no rare variant was detected from MODY genes in this family. Highlighted fonts depict polygenic risk scores using Asian Indian (PRS_{AI}) and European (PRS_{EU}) models.

Some RVs were not found in people of other Indian ethnic groups; these are private variants restricted to a few Punjabi families like rs2063404685 (*HNF4A*; a 5'UTR variant that is prevalent in Ped 160 and no one else carried this variant (Supplementary Fig. 2B and Supplementary Table 4)). However, a rare intronic variant in the *ABCC8* gene (rs376706487) correlated strongly with the T2D phenotype of people in the Punjabi family (Ped 99) (with 100% penetrance) and was also in people of Ped 50 from Jaipur and associated with T2D except for three T2D case and also found in people of 3 more families from Jaipur with 24 total carriers, more than 30% were diabetic while 69% of the remaining carriers were below age 40 and did not have T2D (Supplementary Fig. 3). Additionally, the population-specific missense RV in the *ANPEP* (rs148567888; Gln467His) identified in individuals of one Sikh family (Ped 332) was detected in 59 individuals in the Jaipur families, and 36% of carrier parents had T2D; however, the young offspring carriers ($n = 4$) were nondiabetic (Supplementary Fig. 5). Incidentally, the other two EEG people were monomorphic or had one or two carriers. We also performed the single variant association analysis of all RVs found to be segregated with T2D in the family (discovery) cohort in the unrelated replication cohort of 4602 individuals (2574 cases and 2028 controls). Of the 28 most interesting variants, 24 were found in the replication case-control cohort. However, no significant association was detected with T2D in any of the variants (Supplementary Table 2). Because most of these variants are ultra rare or rare, we would need 25,877 to 94,658 cases at alpha 0.05 to identify RV association with T2D with MAF ranging from 0.001 to 0.0005, respectively, with 80% power.

Next, we performed the gene-centric analysis to estimate the joint effects of the RV burden on T2D in the entire dataset of the people from the Punjabi cohort. Gene-burden analysis revealed the highest RV burden contributed by the *HNF4A* variants (CMC with T^2 test ($p = 0.0003$) and CMC regression $p = 0.0002$) and SKAT-O ($p = 0.0012$) followed by *KCNJ11*/*ABCC8* (CMC with T^2 test ($p = 0.0061$) and CMC regression $p = 0.0055$). At the same time, a marginally significant burden on T2D was observed in CMC T^2 and CMC regression analysis in the new gene *SLC38A11* (Table 2). We further evaluated the RV burden by each pedigree discussed above. The highest RV burden of multiple genes was observed in Ped 36 ($F = 22.07$, $p = 2.36E-06$), followed by Ped 160 ($F = 5.71$; $p = 0.0012$) and Ped 180 ($F = 4.44$; $p = 0.0028$) of AIDHS/SDS (Table 3).

We also evaluated the impact of intronic RVs on gene function and regulation using bioinformatic analysis. The RVs in the *HNF4A* were the most associated with T2D in family 36, showing the highest correlation ($r = 0.59$; $p = 0.002$) (Fig. 3). The rs376287515 splice-site variant in the *HNF4A* was predicted to be damaging with a pathogenicity probability of 73% using RegSNP-intron[48]. The intronic RV in *ABCC8* (rs117727754) represses the activity of TFs Pax4, Sp1, and Nf-κβ with scores of −10.52; $p = 3.0 \times 10^{-7}$; −4.90, 1.9×10^{-4}; and −2.91; $p = 5.3 \times 10^{-4}$, respectively (Fig. 6A). DNA-RNA binding of mutant (mut1 and 2) and wildtype (WT) carriers of the rs117727754 using the DNA pull-down assay and immunostaining revealed reduced binding of TF (Nf-kβ) from 2.6-fold to 4.2-fold and of Pax4 from 2.1 to 2.3-fold in heterozygous mutation carriers compared to the homozygous WT carriers (Fig. 6C, D). Because these are RVs, no homozygous mutant carriers were found in our cohorts. Interestingly, the same variant results in the interaction of *ABCC8* with the *NCR3LG1* gene by forming a chromatin loop (Fig. 6B). Another intronic RV rs150373196 in *HNF4A* gene represses the activity of two TFs (Nfatc1 (−3.64; $p = 3.7 \times 10^{-4}$) and Irf4 (−3.37; $p = 1.2 \times 10^{-3}$)), and enhances the activity of Creb1 (2.10; $p = 1.9 \times 10^{-4}$), and Mef2d (2.26; $p = 2.1 \times 10^{-3}$). The *HNF4A* variant also forms a chromatin loop with *TTPAL*, *SERINC3*, and *PKIG* genes (Supplementary Fig. 6A, B).

Lastly, we evaluated if the families with a high prevalence of T2D and high RV burden also have high PRS. Interestingly, most members of these families showed a low burden of PRS both in PRS_{AI} and PRS_{EU} models. With a few exceptions, the patients carrying a high load of RVs had a low PRS score in these families (Figs. 1, 4, and Supplementary Fig. 2). However, the average PRS (mean ± SE) in this Punjabi population ($n = 4602$) was (61.7 ± 2.8 in T2D cases and 12.0 ± 0.8 in non-T2D controls) using the


A
PED180
T2D
DATA ID :
Age :
Age of T2D onset :
rs139693718 :
rs140708593 :
C0191LRP
75
-
?
?
DATA ID : C0184CSP C0354SBC C0203SSB C0202SBB
Age : 44 35 45 48
Age of T2D onset : 44 - 42 47
rs139693718 : AG AA AA AA
rs140708593 : CC CC CC CC
C0195LKP 52 48 AA CC
C0199LHH 58 43 AG CT
DATA ID : C0198LPS C0196LHS C0197LHS C0193LGS C0194LAS C0192LPS
Age : 26 30 24 27 33 31
Age of T2D onset : 26 - 24 27 - 27
rs139693718 : AA ? AA AG ? AA
rs140708593 : CT ? CT CC ? CC
B
rs139693718 (Phe47Ser)
HUMAN TDTYQSLVNKTFGFPGYLLLSVLQF 95
CHIMPANZEE TDTYQSLVNKTFGFPGYLLLSVLQF 126
MOUSE TDSYQSLVNKTFGFPGYLLLSTLQF 153
RAT TDSYQSLVNKTFGFPGYLLLSTLQF 153
CHICK TNTYQELVRKTFGFLGYLILSTLQF 178
WOLF TDTYQSLVNKTFGFPGYLLLSVLQF 152
CATTLE TDTYQSLVNRTFGFPGYLLLSVLQF 152
*:.:**.**.:.****.***:**.***
* = Fully conserved; : = all pyrimidines or purines at position; . = some matching
C
rs140708593 (Arg116Gln)
HUMAN STVTFTLPLSLYRNIAKLGKVSLISTG 155
CHIMPANZEE STVTFTLPLSLYRNIAKLGKVSLISTG 186
MOUSE STVTCTLPLSLYRDIAKLGKISFISTI 213
RAT STVTCTLPLSLYRDIAKLGKISFISTI 213
CHICK TTIIFTLPLSLYRDIARLGKVSLVSLL 238
WOLF STAAFTLPLSLYRDIAKLGKISFLSTV 212
CATTLE STVVFTLPLSLYRDIAKLGKISLISTV 212
:*   ********:**:***:*::*
* = Fully conserved; : = all pyrimidines or purines at position; . = some matching
ASN44
PHE47
SER47
ASN44
LYS328
ARG116
ALA119
ALA119
PHE261
GLN116

Population	Allele Number	Allele Count	Allele Frequency
European	1179608	908	0.0008
African/African American	74942	29	0.0004
East Asian	44800	0	0
South Asian	90386	323	0.004
Sikhs	4890	20	0.002

Population	Allele Number	Allele Count	Allele Frequency
European	1175516	1109	0.0009
African/African American	74686	8	0.0001
East Asian	44566	2	0.00004
South Asian	90108	21	0.0002
Sikhs	4890	3	0.0003

Fig. 5 | Detection and characterization of two missense RVs in *SLC38A11* gene segregating with T2D. **A** Pedigree of a Sikh family (PED180) shows the representation of new rare damaging variants rs139693718 (Phe47Ser) and rs140708593 (Arg116Gln) in *SLC38A11* gene; **B** Sequence alignment reveals absolute conservation of rs139693718 at position Phe47Ser of the *SLC38A11* gene across species; allele frequency of Phe47Ser among different populations and the wildtype residue (blue) forms hydrogen bonds with Asparagine at position 44. However, the mutant residue (red) Serine forms hydrogen bonds not only with Asparagine 44 but also with Lysine at position 328. **C** Sequence alignment reveals absolute conservation of rs140708593 at position Arg116Gln of the *SLC38A11* gene across species; allele frequency of Arg116Gln among different populations and the wildtype residue (blue) forms hydrogen bonds with Alanine at position 119. However, the mutant residue (red) Glutamine forms hydrogen bonds with Alanine119 and an additional hydrogen bond with Phenylalanine at position 261. Only Ped#180 was carrying these variants.

Table 2 | Gene-centric analysis showing high RV burden on T2D in the selected candidate genes

Method	*SLC38A11*		*KCNJ11/ABCC8*		*HNF4A*	
	P	Permuted *P*	*P*	Permuted *P*	*P*	Permuted *P*
CMC with Hotelling T squared test	0.029	0.037	0.006 0.003*	0.005 0.002*	0.0003	0.0002
CMC with regression	0.030	0.044	0.005 0.002*	0.007 0.002*	0.0002	0.0002
SKAT-O	0.140	–	0.068 0.025*	–	0.001	–

*Rare variant burden calculated excluding less common variants rs5216 and rs376706487.

PRS_{AI}. It was (37.3 ± 1.6 in T2D cases and 7.4 ± 0.3 in non-T2D controls) using the PRS_{EU}[53]. Despite having a high prevalence of diabetes, patients in the family cohort had significantly lower PRS_{AI} (20.3 ± 7.0 vs. 61.7 ± 2.8; $p = 4.4 \times 10^{-12}$) and PRS_{EU} (12.5 ± 4.3 vs. 37.3 ± 1.6; $p = 8.2 \times 10^{-12}$) compared to T2D cases from the unrelated cohort. Even in the subset of families with high RV load, the average PRS in cases was substantially lower compared to cases from the unrelated cohort PRS_{AI} (16.8 ± 11.4 vs. 61.7 ± 2.8; $p = 3 \times 10^{-6}$) and PRS_{EU} (14.2 ± 9.3 vs. 37.3 ± 1.6; $p = 1 \times 10^{-3}$) in both PRS scores driven from people of Asian Indian and European studies (Fig. 7).

Discussion

This study examined ten genetic regions of interest (identified by positional cloning or GWAS approaches) to determine if the low frequency or RVs are associated with T2D in Sikh families because the common variants of GWAS signals did not show a strong association with T2D. Independent replication/validation was conducted using data from other Sikh people (unrelated) and families from three other EEG people. These candidate gene regions were chosen based on the GWAS and association studies in people from SA[8,23]. Of all the genes analyzed, most RVs found in three known genes (i.e., *HNF4A*-MODY-1, *ABCC8*-MODY-12, and *KCNJ11*-MODY-13) showed association with T2D in a handful of families. Previous studies have suggested that MODY genes may overlap with the adult T2D phenotype. However, the role of these genes in T2D is not well characterized and has been underreported, particularly in families with a high prevalence of diabetes[23,58]. MODY diabetes is a dominant form of a group of non-immune diabetes subtypes, each caused by dysfunction in a single gene. MODY phenotype presents in teens or early adulthood, and mutational changes affect the development or the function of β-cells[59]. There are 14 classified MODY genes known to cause different forms of MODY diabetes. This study did not aim to investigate all MODY genes but to screen RVs under the selected GWAS/candidate genes identified in people from SA GWAS and meta-analysis studies. Incidentally, the common variant signal (rs4812829) in the *HNF4A* (MODY-1) was associated with T2D GWAS in SA people[23]. However, this association could not be verified in our families ($p = 0.76$). On the other hand, multiple ultra RVs in *HNF4A* were commonly present in selected Sikh pedigrees, particularly in Ped 36, which also had the highest RV burden ($F = 22.07$, $p = 2.36 \times 10^{-6}$) in multiple MODY genes (*HNF4A*, *KCNJ11-ABCC8*) and *ANPEP* associated with T2D.

ABCC8 and *KCNJ11* genes are located next to each other on chromosome 11p15.1. The encoded proteins of both *KCNJ11* (Kir6.2) and *ABCC8* (SUR1) are subunits of the ATP-sensitive potassium (KATP) channel. Earlier studies suggest that T2D patients with *ABCC8* mutations respond better to sulphonylurea treatment than other antidiabetic drugs or insulin[60]. A rise in glucose affects ATP levels and leads to the closure of the KATP channel, which activates the voltage-dependent calcium channel, stimulating the influx of calcium and releasing insulin. Mutations in these genes influence the regulation of the KATP channel to cause hypo- or hyperglycemia by channel activating and deactivating effects in transitory or permanent neonatal diabetes[61]. Thus, similar effects on insulin secretion or dysfunction may be expected in these families carrying RV load of the *KCNJ11* and *ABCC8* genes in which two variants (rs376706487 in *ABCC8* and rs41282930 in *KCNJ11*) were in complete LD (Supplementary Fig. 1).

Table 3 | Rare variant burden by each pedigree

Family	Genes	F	P
PED36	*SLC38A11, KCNJ11, ABCC8, ANPEP, HNF4A*	22.07 20.63*	2.36E-06 6.25E-05*
PED99	*SLC38A11, HK1, KCNJ11, ABCC8, HNF4A*	2.15 1.82*	0.0004 0.0002*
PED160	*SLC38A11, ANPEP, HNF4A*	5.71	0.001
PED180	*SLC38A11, ABCC8*	4.44	0.003
PED332	*DDX50, HK1*	2.96	0.025
PED324	*ABCC8*	2.73	0.114
PED277/278	*KCNJ11*	0.03	0.858

*Rare variant burden calculated excluding less common variants rs5216 and rs376706487 in PED36 and PED99.

The rs41282930 (*KCNJ11*) is a population-specific rare missense variant that segregated with T2D in four families of Sikh people (Ped 99, 36, 277, and 278), supporting the enrichment of very rare SNPs due to inbreeding and no other people from non-SA study has reported this variant. However, we could not confirm the prevalence or association of rs41282930 with T2D in the other three EEG people because of the lack of genotype data for this variant. The rs41282930 results in an amino acid change from Ser to Cys at codon 385, which would create a more hydrophobic residue (cysteine), which makes an additional bond with ILE384 that would disturb correct folding as predicted by bioinformatic analysis (Fig. 2D). The phosphorylation of the Ser385 residue of the Kir6.2 subunit of the KATP (potassium channel) by AMPK (AMP-activated protein kinase) may play a role in insulin secretion by disrupting the activation of the calcium channel[62,63]. A Glu227Lys variant in *KCNJ11* was detected in a MODY family from people of France with incomplete penetrance[64]. Another variant in *KCNJ11* at codon 42 (Cys42Arg) was associated with MODY and the later onset of T2D in Japanese people; both variants were not found in our population[65]. We also did not detect the Val169Ile found in South Indian people and Arg114Trp in European people associated with MODY 1 diabetes (*HNFA4*) in our families[66,67]. However, those families carrying RVs from multiple MODY genes (*HNFA4*, *KCNJ11-ABCC8*) may manifest T2D due to insulin secretory defects[68] and may suggest an oligogenic model of inheritance of rare forms of T2D caused by MODY or other genes.

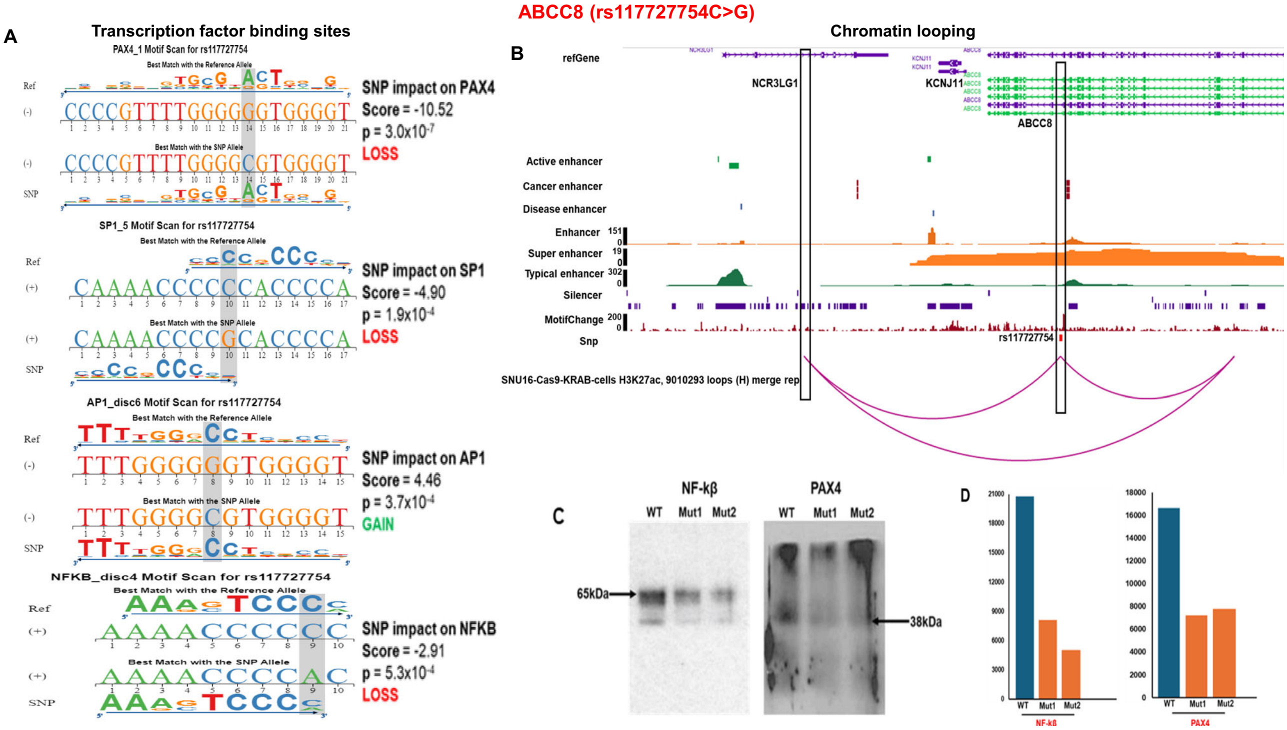


Fig. 6 | The rare intronic variant rs117727754 detected in the *ABCC8* gene was present in the transcription factor binding site. **A** Rare variant represses the activity of transcription factors Pax4, Sp1, and NF-kβ with scores of −10.52, −4.90, and −2.91, respectively, and enhances the activity of Ap1 (Score = 4.46). **B** The *ABCC8* variant also forms a chromatin loop with *NCR3LG1* gene. **C** The Pax4-ABCC8 and NF-kβ-ABCC8 pull-down assay using biotinylated probes: DNA-transcription factor (RNA) binding in Western blots showing the binding of NF-kβ -ABCC8 and Pax4-ABCC8 among two mutant (CG) human carriers vs. wildtype (CC) carriers. **D** Bar graphs representing the decreased intensity (2.6-fold) for Mut1 and (4.2-fold) Mut2 for Nf-kβ, and (2.3-fold) for Mut1 and (2.1-fold) for Mut2 compared to the WT band for the Pax4 transcription factor. The bands were quantified and analyzed using ImageJ.

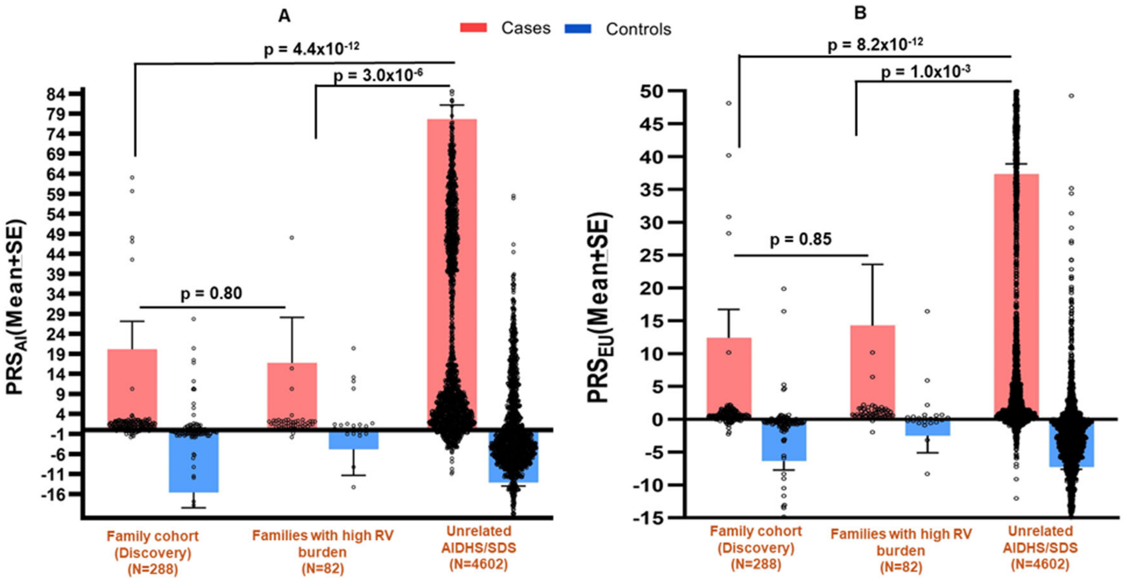


Fig. 7 | Bar charts representing the distribution of polygenic risk scores. PRS_{AI} (**A**) and PRS_{EU} (**B**) showing significant differences between family cohort (Discovery) ($N = 288$), families with high rare variant burden from the selected GWAS candidate genes ($N = 82$) vs. the unrelated individuals of AIDHS/SDS ($N = 4602$) among T2D cases. Errors bars represent the standard error and individual data points are added.

In addition, we identified RVs in two new genes, *SLC38A11* and *ANPEP*, which could be potential new candidates for T2D in these populations. The presence of two very rare LoF RVs in *SLC38A11* was only detected in one family (Ped 180) matching with T2D phenotype (Fig. 5), and intronic variants showing a strong correlation with T2D in families 99, 160, and 36 support the role of *SLC38A11* in T2D. Notably, the *SLC38A11* falls under the T2D GWAS locus *GRB14*[23] and is an amino acid transporter predicted to be involved in amino acid transmembrane transporter activity. Recent studies have reported common variants in the *SLC38A11* association with BMI, waist-to-hip ratio, and triglycerides, but its contribution to T2D pathophysiology is currently unknown[69]. The association of *ANPEP* with T2D has been reported in the Gene Expression Omnibus database of pancreatic beta cells[70], and plasma levels of *ANPEP* increased in patients with prostate cancer[71]. Interestingly, the detection of ultra-rare missense variant in the *ANPEP* (Gln467His; rs148567888) identified in people from one Sikh family was more prevalent in people from Agarwal (Jaipur) families (Supplementary Fig. 4); both Sikhs and Agarwals are from Northern India. At the same time, this variant was monomorphic in people from the other two EEGs from Southern India (Chettiars and Reddys). Our T2D families came from four geographically and ethnically distinct EEGs: North (Punjab and Rajasthan) and South (Tamil Nadu and Andhra Pradesh) India. As expected, several differences in the T2D phenotype and other traits were identified in these families from different EEGs (Table 1), as reported earlier[8,29].

Intronic RV in the *ABCC8* gene (rs117727754) in the Ped 36 family represses the activity of three TFs (Pax4, Sp1, and Nf-κβ), and the highest score was found in the Pax4 (score −10.52). The significantly reduced binding of TFs (Pax4 and Nf-kβ) with *ABCC8* in the DNA pull-down assay further supports the functional role of this RV in T2D. The Pax4 or Paired Box 4 is a key TF involved in the embryonic development of the pancreatic islets of Langerhans and thus has an essential role in insulin-secreting β cell lineage. Knockout of Pax4 in animal models leads to the absence of β cells, resulting in a significant increase in glucagon-producing α cells, leading to neonatal deaths[72]. The Nf-κβ is known to play a key role in the pathogenesis of vascular complications of diabetes. Persistent hyperglycemia in T2D activates Nf-κβ and triggers the expression of various cytokines, chemokines, and cell adhesion molecules[73]. Additionally, the *NCR3LG1* gene (located upstream of *ABCC8* and *KCNJ11*) controls their activity and helps in the insulin secretion pathway. However, the chromatin loop interaction due to RV rs117727754 in the *ABCC8* (Fig. 6B) could be repressing the insulin secretion pathway impacted by *NCR3LG1*[74], which agrees with our hypothesis of defects in insulin secretion in these families. The intronic RV (rs150373196) in the *HNF4A* gene in the Ped 36 family enhances the activity of a known TF-Creb. The Creb phosphorylation triggers the cascade for the expression of enzymes involved in gluconeogenesis. Studies have shown that Creb is over-activated by glucagon and many other factors contributing to the elevation of fasting glucose levels in diabetes[75,76]. Most of the TFs enhanced or repressed by the intronic RVs have a role in T2D. Similarly, most of the carriers of these variants in the Ped 36 family have T2D, which might suggest the role of these variants in dysregulating T2D-based TFs.

Most of the identified RVs have only recently been added to GnomeAD. The role of these variants with T2D is currently unknown in any population except three RVs (rs41282930, rs1800467, and rs5216) have been reported in two small studies published in Asian Indians and Indo-Trinidad populations[77,78]. The highest rate of RV prevalence occurred in families with a strong history of T2D, showing dominant transmission patterns like that of Mendelian genes. The most intriguing finding of this study was that, despite carrying a high prevalence of T2D and high RV burden, most members of these families showed a lower average burden of common variant-derived PRS models (PRS_{AI} and PRS_{EU}) constructed from extensive T2D GWAS meta-analysis studies (Fig. 7). The lack of a common variant burden may suggest an oligogenic origin of T2D, which the PRS does not explain. The RV enrichment of *SLC38A11*, *ANPEP*, and three MODY genes calls for an in-depth evaluation of these families through whole-genome sequencing.

Lastly, as we know, the T2D etiology is highly heterogeneous due to its multiple roots of origin; population heterogeneity due to genetic background and environmental differences poses further challenges in underestimating the complexity of the T2D phenotype[79,80]. Such families from genetically homogenous EEGs can provide additional insights into the molecular mechanism by identifying different molecular subtypes of T2D, as suggested by this study. The limitations include the genetic information reported in this study being limited to only a few GWAS loci, the RVs information from the remaining part of the human genome, and other ~400 confirmed GWAS genes missing. Based on the observed RV load of three MODY genes in T2D families, investigations of other MODY genes are warranted and may identify rare subtypes of T2D. Moreover, the RVs of these target gene regions were primarily available in the discovery cohort of Sikh EEG people, and only 118 variants were genotyped in the people of the remaining EEGs. Also, we used imputed SNPs for replication analysis in our case-control cohort, which could have resulted in the poor replication of these RVs. The imputation algorithms historically work poorly for RVs and, more specifically, in populations like ours due to the absence of a population-based reference panel[81]. Besides, the high prevalence of RVs in the family pedigrees is not unexpected, as the offspring inherit 50% of the genome from each parent. However, the enrichment of the RVs from gene loci already implicated in causing MODY diabetes is intriguing, specifically in those T2D pedigrees with low PRS. Moreover, this study shows an enrichment of intronic variants with functional consequences, which necessitates deeper evaluations of such families from understudied populations for potential novel therapeutics. Further, our data revealed the presence of a north-south gradient in the RV association for *ANPEP* and *ABCC8* genes with T2D in people of Punjabi and Agarwal (Jaipur) families, both from the northern region. These findings support earlier published genetic studies on Indian populations reflecting founder effects due to the endogamous nature, multiple cast groups, or separation due to geographic distance among the Indian communities[9,82].

In summary, our data indicates that the abundance of noncoding RVs from various MODY and other new genes contributes significantly to the development of T2D in certain families with a low common variant burden, as indicated by PRS, suggesting an oligogenic mode of inheritance. It is also possible that RVs from polygenic susceptibility genes play a role in causing a high genetic burden for T2D. Since these families were only sequenced for a limited number of target gene regions, a thorough evaluation using whole-genome sequencing is needed to clarify the genetic factors contributing to the high burden of T2D in these families. These efforts will improve disease classification and assist with developing personalized medicine approaches to advance risk prediction and treatment options.

Reporting summary

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

Data availability

The source data for Fig. 7 is in Supplementary Table 6. Data files with variant information for constructing the genome-wide PRS are available in the supplementary data files as Table 7 for Asian Indian people and Table 8 for European people. The additional information on PRS is available through our previous manuscript (Rout et al.[53]). The genome-wide genotype data associated with the Punjabi Sikh study with phenotype data has already been submitted to DbGap https://www.ncbi.nlm.nih.gov/gap/advanced_search/?TERM=sanghera. The data from these family studies containing sensitive participant information for further analysis will be made available at reasonable request through collaborations through the corresponding author.

Received: 28 November 2023; Accepted: 23 January 2025;
Published online: 22 February 2025

References

1. Sanghera, D. K. & Blackett, P. R. Type 2 diabetes genetics: beyond GWAS. *J. Diabetes Metab.* **3**, 6948 (2012).
2. L'Heveder, R. & Nolan, T. International Diabetes Federation. *Diab. Res. Clin. Pract.* **101**, 349–351 (2013).
3. Chen, L., Magliano, D. J. & Zimmet, P. Z. The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives. *Nat. Rev. Endocrinol.* **8**, 228–236 (2011).
4. McKeigue, P. M., Pierpoint, T., Ferrie, J. E. & Marmot, M. G. Relationship of glucose intolerance and hyperinsulinaemia to body fat pattern in south Asians and Europeans. *Diabetologia* **35**, 785–791 (1992).
5. Ogurtsova, K. et al. IDF Diabetes Atlas: global estimates for the prevalence of diabetes for 2015 and 2040. *Diab. Res. Clin. Pract.* **128**, 40–50 (2017).
6. Nakagami, T. et al. Age, body mass index and Type 2 diabetes-associations modified by ethnicity. *Diabetologia* **46**, 1063–1070 (2003).
7. Goyal, S. & Sanghera, D. K. Genetic and non-genetic determinants of cardiovascular disease in south Asians. *Curr. Diab. Rev.* **17**, e011721190373 (2021).
8. Saxena, R. et al. Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India. *Diabetes* **62**, 1746–1755 (2013).
9. Indian Genome Variation, C. Genetic landscape of the people of India: a canvas for disease gene exploration. *J. Genet.* **87**, 3–20 (2008).
10. Cai, L. N. et al. Genome-wide association analysis of type 2 diabetes in the EPIC-InterAct study. *Sci. Data* **7**, 393 (2020).
11. 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).
12. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**, 212–219 (2009).
13. Shuldiner, A. R. & Pollin, T. I. Genomics: variations in blood lipids. *Nature* **466**, 703–704 (2010).
14. Vujkovic, M. et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691 (2020).
15. King, A., Wu, L., Deng, H. W., Shen, H. & Wu, C. Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease. *BMC Med.* **20**, 385 (2022).
16. Fritsche, L. G. et al. On cross-ancestry cancer polygenic risk scores. *PLoS Genet.* **17**, e1009670 (2021).
17. DeForest, N. & Majithia, A. R. Genetics of type 2 diabetes: implications from large-scale studies. *Curr. Diab. Rep.* **22**, 227–235 (2022).
18. Marquez-Luna, C. et al. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).
19. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
20. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
21. Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
22. Wang, K. et al. Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.* **86**, 730–742 (2010).
23. 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).
24. Saxena, R. et al. Genome-wide association study identifies variants in casein kinase II (CSNK2A2) to be associated with leukocyte telomere length in a Punjabi Sikh diabetic cohort. *Circ. Cardiovasc. Genet.* **7**, 287–295 (2014).
25. Sapkota, B. R. et al. Genome-wide association study of 25(OH) Vitamin D concentrations in Punjabi Sikhs: results of the Asian Indian diabetic heart study. *J. Steroid Biochem. Mol. Biol.* **158**, 149–156 (2016).
26. Sanghera, D. K. et al. The Khatri Sikh Diabetes Study (SDS): study design, methodology, sample collection, and initial results. *Hum. Biol.* **78**, 43–63 (2006).
27. Sanghera, D. K. et al. Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk. *BMC Med. Genet.* **9**, 59 (2008).
28. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diab. Care* **27**, S5–S10 (2004).
29. Venkatesan, V. et al. Burden of Type 2 diabetes and associated cardiometabolic traits and their heritability estimates in endogamous ethnic groups of India: findings from the INDIGENIUS Consortium. *Front. Endocrinol.* **13**, 847692 (2022).
30. Sanghera, D. K. et al. Targeted sequencing of candidate genes of dyslipidemia in Punjabi Sikhs: population-specific rare variants in

GCKR promote ectopic fat deposition. *PLoS ONE* **14**, e0211661 (2019).
31. Sapkota, B. R. & Sanghera, D. K. A rare missense variant in the milk fat globule-EGF factor 8 (MFGE8) increases T2DM susceptibility and cardiovascular disease risk with population-specific effects. *Acta Diabetol.* **57**, 733–741 (2020).
32. Been, L. F. et al. Replication of association between a common variant near melanocortin-4 receptor gene and obesity-related traits in Asian Sikhs. *Obesity* **18**, 425–429 (2010).
33. Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* **16**, 1215 (1988).
34. Goyal, S. et al. APOC3 genetic variation, serum triglycerides, and risk of coronary artery disease in Asian Indians, Europeans, and other ethnic groups. *Lipids Health Dis.* **20**, 113 (2021).
35. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
36. Sui, H., Chen, Q. & Imamichi, T. A pull-down assay using DNA/RNA-conjugated beads with a customized competition strategy: an effective approach to identify DNA/RNA binding proteins. *MethodsX* **7**, 100890 (2020).
37. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
38. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
39. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
40. Purcell, S., Cherny, S. S. & Sham, P. C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–150 (2003).
41. Rout, M. et al. Polygenic risk score assessment for coronary artery disease in Asian Indians. *J. Cardiovasc Transl. Res.* **17**, 1086–1096 (2024).
42. Venselaar, H., Te Beek, T. A., Kuipers, R. K., Hekkelman, M. L. & Vriend, G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinform.* **11**, 548 (2010).
43. Li, J. et al. VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Res.* **46**, D1039–D1048 (2018).
44. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
45. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
46. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
47. Shin, S., Hudson, R., Harrison, C., Craven, M. & Keles, S. atSNP Search: a web resource for statistically evaluating influence of human genetic variation on transcription factor binding. *Bioinformatics* **35**, 2657–2659 (2019).
48. Lin, H. et al. RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biol.* **20**, 254 (2019).
49. Zhou, Q. et al. ChromLoops: a comprehensive database for specific protein-mediated chromatin loops in diverse organisms. *Nucleic Acids Res.* **51**, D57–D69 (2023).
50. Abdullah, N., Attia, J., Oldmeadow, C., Scott, R. J. & Holliday, E. G. The architecture of risk for type 2 diabetes: understanding Asia in the context of global findings. *Int. J. Endocrinol.* **2014**, 593982 (2014).
51. Sanghera, D. K., Sapkota, B. R., Aston, C. E. & Blackett, P. R. Vitamin D status, gender differences, and cardiometabolic health disparities. *Ann. Nutr. Metab.* **70**, 79–87 (2017).
52. Grarup, N., Sandholt, C. H., Hansen, T. & Pedersen, O. Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia* **57**, 1528–1541 (2014).
53. Rout, M. et al. Assessing the prediction of type 2 diabetes risk using polygenic and clinical risk scores in South Asian study populations. *Ther. Adv. Endocrinol. Metab.* **14**, 20420188231220120 (2023).
54. Choi, S. W., Mak, T. S. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
55. Bejar, C. A. et al. A Bidirectional Mendelian Randomization Study to evaluate the causal role of reduced blood vitamin D levels with type 2 diabetes risk in South Asians and Europeans. *Nutr. J.* **20**, 71 (2021).
56. Collister, J. A., Liu, X. & Clifton, L. Calculating polygenic risk scores (PRS) in UK Biobank: a practical guide for epidemiologists. *Front. Genet.* **13**, 818574 (2022).
57. O'Connor, M. J. et al. Recessive genome-wide meta-analysis illuminates genetic architecture of type 2 diabetes. *Diabetes* **71**, 554–565 (2022).
58. Bansal, V. et al. Spectrum of mutations in monogenic diabetes genes identified from high-throughput DNA sequencing of 6888 individuals. *BMC Med.* **15**, 213 (2017).
59. Naylor R, Knight Johnson A, del Gaudio D. Maturity-Onset Diabetes of the Young Overview. 2018 May 24. In: Adam MP, Feldman J, Mirzaa GM, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2025. Available from: https://www.ncbi.nlm.nih.gov/books/NBK500456/.
60. Zwaveling-Soonawala, N. et al. Successful transfer to sulfonylurea therapy in an infant with developmental delay, epilepsy and neonatal diabetes (DEND) syndrome and a novel ABCC8 gene mutation. *Diabetologia* **54**, 469–471 (2011).
61. Proks, P. et al. Molecular basis of Kir6.2 mutations associated with neonatal diabetes or neonatal diabetes plus neurological features. *Proc. Natl Acad. Sci. USA* **101**, 17539–17544 (2004).
62. Chang, T. J. et al. Serine-385 phosphorylation of inwardly rectifying K + channel subunit (Kir6.2) by AMP-dependent protein kinase plays a key role in rosiglitazone-induced closure of the K(ATP) channel and insulin secretion in rats. *Diabetologia* **52**, 1112–1121 (2009).
63. Bowman, P. et al. Effectiveness and safety of long-term treatment with sulfonylureas in patients with neonatal diabetes due to KCNJ11 mutations: an international cohort study. *Lancet Diab. Endocrinol.* **6**, 637–646 (2018).
64. Antcliff, J. F., Haider, S., Proks, P., Sansom, M. S. & Ashcroft, F. M. Functional analysis of a structural model of the ATP-binding site of the KATP channel Kir6.2 subunit. *EMBO J.* **24**, 229–239 (2005).
65. Yorifuji, T. et al. The C42R mutation in the Kir6.2 (KCNJ11) gene as a cause of transient neonatal diabetes, childhood diabetes, or later-onset, apparently type 2 diabetes mellitus. *J. Clin. Endocrinol. Metab.* **90**, 3174–3178 (2005).
66. Chapla, A. et al. Maturity onset diabetes of the young in India—a distinctive mutation pattern identified through targeted next-generation sequencing. *Clin. Endocrinol.* **82**, 533–542 (2015).
67. Mirshahi, U. L. et al. Reduced penetrance of MODY-associated HNF1A/HNF4A variants but not GCK variants in clinically unselected cohorts. *Am. J. Hum. Genet.* **109**, 2018–2028 (2022).
68. Gloyn, A. L. et al. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568–572 (2003).
69. Crocco, P. et al. The genetic variability of members of the SLC38 family of amino acid transporters (SLC38A3, SLC38A7 and SLC38A9) affects susceptibility to type 2 diabetes and vascular complications. *Nutrients* **14**, 4440 (2022).
70. Ding, L., Fan, L., Xu, X., Fu, J. & Xue, Y. Identification of core genes and pathways in type 2 diabetes mellitus by bioinformatics analysis. *Mol. Med. Rep.* **20**, 2597–2608 (2019).

71. Sorensen, K. D. et al. Prognostic significance of aberrantly silenced ANPEP expression in prostate cancer. *Br. J. Cancer* **108**, 420–428 (2013).
72. Ko, J., Fonseca, V. A. & Wu, H. Pax4 in health and diabetes. *Int. J. Mol. Sci.* **24**, 8283 (2023).
73. Suryavanshi, S. V. & Kulkarni, Y. A. NF-kappabeta: a potential target in the management of vascular complications of diabetes. *Front. Pharm.* **8**, 798 (2017).
74. Bevacqua, R. J. et al. CRISPR-based genome editing in primary human pancreatic islet cells. *Nat. Commun.* **12**, 2397 (2021).
75. Benchoula, K., Parhar, I. S., Madhavan, P. & Hwa, W. E. CREB nuclear transcription activity as a targeting factor in the treatment of diabetes and diabetes complications. *Biochem. Pharm.* **188**, 114531 (2021).
76. Dalle, S., Quoyer, J., Varin, E. & Costes, S. Roles and regulation of the transcription factor CREB in pancreatic beta -cells. *Curr. Mol. Pharm.* **4**, 187–195 (2011).
77. Boodram, L. G., Miyake, K., Hayes, M. G., Bell, G. I. & Cockburn, B. N. Association of the KCNJ11 variant E23K with type 2 diabetes in Indo-Trinidadians. *West Indian Med. J.* **60**, 604–607 (2011).
78. Phani, N. M. et al. Population specific impact of genetic variants in KCNJ11 gene to type 2 diabetes: a case-control and meta-analysis study. *PLoS ONE* **9**, e107021 (2014).
79. Been, L. F. et al. A low frequency variant within the GWAS locus of MTNR1B affects fasting glucose concentrations: genetic risk is modulated by obesity. *Nutr. Metab. Cardiovasc. Dis.* **22**, 944–951 (2012).
80. Blackett, P. R. & Sanghera, D. K. Genetic determinants of cardiometabolic risk: a proposed model for phenotype association and interaction. *J. Clin. Lipidol.* **7**, 65–81 (2013).
81. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7**, 16 (2015).
82. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).

Acknowledgements

We warmly thank the participants of the AIDHS/SDS, TNFDS, JFDS, and NFDS and are grateful for their contribution to this study. This study was supported by grants from the Indian Council of Medical Research [ICMR] (India) Project: No. 55/6/2/Indo-US/ 2014-NCD-II and the National Institute of Diabetes and Digestive and Kidney Diseases [NIDDK], US National Institutes of Health [NIH] grant R21DK105913. The genotyping of AIDHS/SDS cohorts was also partly supported by NIDDK R01DK082766, R01DK118427, Dr. Geoffrey Altshuler Children Hospital Foundation Endowment funds, and the Presbyterian Health Foundation grants.

Author contributions

M.R. and C.E.A. performed data analysis. M.M. and A.K.G. helped in the DNA pull-down assay. S.R., G.S.W., and J.R.S. contributed to recruitment and phenotyping for AIDHS/SDS, and P.R.B. helped in clinical phenotyping and contributed to the manuscript review. The clinical data and biospecimen sample collection, processing, and data management of TNFDS, JFDS, and NFDS were supervised by S.P., C.N., U.R., R.G., K.M., D.K., A.P., and R.D. through direct contributions from V.V., T.K., D.R., S.S., S.L., K.S., M.S., R.R., P.V., P.R., N.S., and J.E. D.K.S. contributed to the study design, genotyping, and phenotyping as a cohort PI of AIDHS/SDS, and wrote the manuscript; R.D. and C.E.A. helped edit the manuscript. All authors read and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-00750-9.

Correspondence and requests for materials should be addressed to Dharambir K. Sanghera.

Peer review information *Communications Medicine* thanks Gautam Pandey and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

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

Madhusmita Rout[1], Deepika Ramu[2], Mendez Mariana[3], Teena Koshy[2], Vettriselvi Venkatesan[2], Juan C. Lopez-Alvarenga[4], Rector Arya[5], Umarani Ravichandran[6], Surendra K. Sharma[7], Sailesh Lodha[8], Amaresh Reddy Ponnala[9], Krishna Kumar Sharma[10], Mahaboob Vali Shaik[11], Roy G. Resendez[5], Priyanka Venugopal[2], Parthasarathy R[2], Noelta S[2], Juliet A. Ezeilo[5], Marcio Almeida[12], Juan Paralta[12], Srinivas Mummidi[5], Chidambaram Natesan[2], Narinder K. Mehra[13], Jai Rup Singh[14], Gurpreet S. Wander[15], Sarju Ralhan[15], Piers R. Blackett[1], John Blangero[12], Krishna M. Medicherla[16], Sadagopan Thanikachalam[2], Thyagarajan Sadras Panchatcharam[2], Dileep Kumar K[11], Rajeev Gupta[8], Solomon Franklin D. Paul[2], Asish K. Ghosh[3], Christopher E. Aston[1], Ravindranath Duggirala[5] & Dharambir K. Sanghera[1,17,18,19,20] ✉

[1]Department of Pediatrics, College of Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [2]Department of Human Genetics, Sri Ramachandra Institute of Higher Education and Research (Deemed to be University), Chennai, India. [3]Stephenson Cancer Center, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [4]Department of Population Health & Biostatistics, University of Texas Rio Grande Valley (UTRGV), Harlingen, TX, USA.

[5]Department of Health and Behavioral Sciences, Texas A&M University-San Antonio, San Antonio, TX, US. [6]Department of Medicine, Rajah Muthiah Medical College Hospital, Annamalai University, Chidambaram, India. [7]Department of Endocrinology, Galaxy Specialty Centre, Jaipur, India. [8]Departments of Preventive Cardiology, Internal Medicine and Endocrinology, Eternal Heart Care Centre and Research Institute, Mount Sinai New York Affiliate, Jaipur, India. [9]Department of Endocrinology, Krishna Institute of Medical Sciences (KIMS) Hospital, Nellore, India. [10]Department of Pharmacology, Lal Bahadur Shastri College of Pharmacy, Rajasthan University of Health Sciences, Jaipur, India. [11]Department of Endocrinology, Narayana Medical College and Hospital, Nellore, India. [12]Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley (UTRGV), Brownsville, TX, USA. [13]All India Institute of Medical Sciences and Research, New Delhi, India. [14]Guru Nanak Dev University, Amritsar, India. [15]Hero Dayanand Medical College and Heart Institute, Ludhiana, India. [16]Birla Institute of Scientific Research, Jaipur, India. [17]Department of Pharmaceutical Sciences, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [18]Department of Physiology, College of Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [19]Oklahoma Center for Neuroscience, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [20]Harold Hamm Diabetes Center, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA.
✉e-mail: Dharambir-sanghera@ouhsc.edu