nature genetics

Supplementary information

https://doi.org/10.1038/s41588-024-01947-9

Rare variant analyses in 51,256 type 2 diabetes cases and 370,487 controls reveal the pathogenicity spectrum of monogenic diabetes genes

In the format provided by the authors and unedited

Supplementary Note

Description of the discovery cohorts and Type 2 Diabetes definitions

UK Biobank

The UK Biobank (UKB) is a prospective cohort study with genetic and phenotypic data collected on approximately 500,000 individuals across the United Kingdom[1]. Participants agreed to provide detailed information about their lifestyle, environment, and medical history, biological samples (for genotyping and biochemical assays), to undergo measures, and to have their health followed (http://www.ukbiobank.ac.uk/). The UKB has obtained ethical approval covering this study from the National Research Ethics Committee (REC reference 11/NW/0382).

The UKB data was accessed through the application number 27892. UKB is composed of 46% males, 54% females, and ranges in age from 37 to 80 years old, with an average age of 57 years. 15% of the individuals are from genetically-inferred non-European ancestry, with 1.4% African/African American (AFA), 0.20% Admixed American (AMR), 1.8% Central South Asian (CSA), 0.55% East Asian (EAS), 0.3% Middle East (MID), and 10.8% other.

Genotype calling was performed by Affymetrix on two closely related purpose-designed arrays. ~50,000 participants were run on the Applied BiosystemsTM UK BiLEVE AxiomTM Array and the remaining ~450,000 were run on the Applied BiosystemsTM UK Biobank AxiomTM Array. The dataset combines results from both arrays and are 805,426 total markers in the released genotype data.

We used the set of variants (SNV=687,342) and samples (N=486,757) from the quality control procedure previously designed by the UKB to filter both arrays separately (UK BiLEVE and UK Biobank Axiom array)[2]. We then converted each filtered quality-controlled (QCed) array data to VCF format and performed phasing separately using SHAPEIT v4[3]. To ensure appropriate phasing of the X chromosome, we updated the array data to ensure males were encoded as diploid in the VCF file before running SHAPEIT v4.

After phasing, we split the data into chunks of about 25,000 samples each, randomly selected independently of ancestry or T2D status. We submitted the chunks to the TOPMed imputation server (https://imputation.biodatacatalyst.nhlbi.nih.gov/)[4] using the reference panel freeze 8[5], and once completed, we merged the imputed data using BCFTOOLS v1.20[6]. We filtered the imputed data to exclude variants with an Rsq below 0.3 within each chunk. When merging the chunks after imputation, all variants present in any of the imputed chunks were included in the fully merged data. After merging, we converted the VCF files to BGEN format for analysis with REGENIE[7].

UKB included a total of 27,323 T2D cases and 259,916 controls. We defined T2D cases and controls using an algorithm designed specifically for UKB[8]. Using this algorithm, we defined cases and controls using the following criteria:

Case selection criteria: 1) possible T2D, 2) probable T2D, or 3) had maximum HbA1c above 6.5 and were not possible or probable type 1 diabetes.

Control selection criteria: 1) not possible or probable type 1 or type 2 diabetes, 2) did not have a family history of diabetes, 3) had a maximum HbA1c below 5.7, and 4) were older than 45 years of age.

Mass General Brigham Biobank

Mass General Brigham Biobank (MGBB; formerly Partners HealthCare Biobank)[9] is a large repository of biospecimens and data linked to extensive electronic health records and survey data. Its objective is to support and enable translational research on genomic, environmental, biomarker, and family history associations with disease phenotypes. MGBB has enrolled more than 135,000 participants and has generated genomic data on more than 65,000 participants. MGBB consists of consented patients seen at various U.S. hospitals, including Massachusetts General Hospital, Brigham and Women's Hospital, McLean Hospital, and Spaulding Rehabilitation Hospital. Patients are recruited in the context of clinical care appointments at more than 40 sites. MGBB participants provide consent for the use of their samples and data in broad-based research. The approval for the analysis of MGB data was obtained from the MGB Institutional Review Board (study 2016P001018).

We accessed Multi-Ethnic Genotyping Array (MEGA) genotyping data for 36K participants and Infimum Global Screening Array (GSA) genotyping data for 28K participants from MGBB, then QCed, phased, and imputed each of these array datasets separately. Briefly, the QC steps included filtering out genotyped variants based on MAF levels (<0.5%), missingness (<0.05), genotyping batch bias ($P < 5x10^{-5}$), and Hardy-Weinberg equilibrium ($P<1x10^{-10}$) and palindromic single nucleotide variants (AT or CG). An additional 2% of individuals were removed if their self-described sex did not match their genetic sex and if they had a high ratio of heterozygote variants. Filtering variants based on Hardy-Weinberg equilibrium and participants based on heterozygosity was performed in self-identified race and ethnicity subgroups. This clean dataset was then phased using Shapeit4 and imputed using the TOPMed reference panel freeze 8. The union of the imputed MEGA and GSA datasets were then merged together into a final dataset.

MGBB included a total of 6,623 T2D cases and 41,411 controls. T2D status was defined based on “curated phenotypes” developed by the MGBB Portal team using structured and unstructured electronic medical record data and clinical, computational, and statistical methods. Natural Language Processing was used to extract data from narrative text. Chart reviews by disease experts helped identify features and variables associated with particular phenotypes and were also used to validate the results of the algorithms. The process produced robust phenotype algorithms that were evaluated using metrics such as positive predictive value (PPV) and negative predictive value (NPV)[10].

Control selection criteria: 1) Individuals determined by the “curated disease” algorithm employed above to have no history of T2D with NPV of 99%, 2) Individuals at least age 55, and 3) Individuals with HbA1c less than 5.7.

Case selection criteria: 1) Individuals determined by the “curated disease” algorithm employed above to have T2D with PPV of 99%, and 2) Individuals at least age 30 given the higher rate of false positive diagnoses in younger individuals.

Genetic Epidemiology Research on Adult Health and Aging

Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort data was obtained through dbGaP under accession phs000674.v1.p1. Further information about the specific phenotypes (ICD-9-CM codes) included in GERA is available on its website on dbGaP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd004308).

The GERA Cohort was created by an RC2 Grand Opportunity grant that was awarded to the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics (AG036607; Schaefer/Risch, PIs). The RC2 project enabled genome-wide SNP genotyping (GWAS) to be conducted on a cohort of over 100 K adults who were members of the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC), and participating in its RPGEH. The resulting GERA cohort is composed of 42% males and 58% females and ranges in age from 18 to over 100 years old, with an average age of 63 years at the time of the RPGEH survey (2007).

This study included 60,710 individuals of European ancestry from the GERA cohort. We applied pre-imputation quality control protocol using PLINK v2.0[11]. Similar to MGBB, the QC steps for the GERA array data included filtering out variants based on MAF (<0.5%), missingness (<0.05), Hardy-Weinberg equilibrium ($P < 1x10^{-10}$), and palindromic single nucleotide variants (AT or CG). No individuals were removed due to discordance between self-described and genetically-inferred

sex. A total of 647,106 variants and 60,710 samples were included following quality control. For phasing and imputation, we performed the same steps described above for UKB.

GERA dataset included a total of 7,498 T2D cases and 53,212 controls.

Inclusion criteria: a) Eligible for RPGEH survey, b) ≥ 18 years of age at time of survey mailing (2007), c) KP Northern California Region enrollee for at least 2 years prior to survey, d) Consented to contribute biospecimen to RPGEH and returned saliva sample by cut-off date for GERA genotyping, e) Successfully genotyped (DQC ≥ 0.82; call rate ≥ 0.97) from extracted DNA, and d) Consented explicitly to have data deposited in NIH-maintained database.

Exclusion criteria: a) Subject requested withdrawal from study after DNA extraction and genotyping, b) Validity of link between biospecimen and study participant questionable because of genotype-phenotype discordance, e.g. gender.

A participant was coded as a T2D patient if he/she had at least two diagnoses within this disease category that had to be recorded on separate days. Diagnoses were obtained from patient encounters at Kaiser Permanente Northern California facilities from January 1, 1995 to March 15, 2013.

The March 2013 ICD9-CM diagnoses used for the T2D category were:

a) 250.00 Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled.

b) 250.02 Diabetes mellitus without mention of complication, type II or unspecified type, uncontrolled.

c) 250.10 Diabetes with ketoacidosis, type II or unspecified type, not stated as uncontrolled.

d) 250.12 Diabetes with ketoacidosis, type II or unspecified type, uncontrolled.

e) 250.20 Diabetes with hyperosmolarity, type II or unspecified type, not stated as uncontrolled.

f) 250.22 Diabetes with hyperosmolarity, type II or unspecified type, uncontrolled.

g) 250.30 Diabetes with other coma, type II or unspecified type, not stated as uncontrolled.

h) 250.32 Diabetes with other coma, type II or unspecified type, uncontrolled.

i) 250.40 Diabetes with renal manifestations, type II or unspecified type, not stated as uncontrolled.

j) 250.42 Diabetes with renal manifestations, type II or unspecified type, uncontrolled.

k) 250.50 Diabetes with ophthalmic manifestations, type II or unspecified type, not stated as uncontrolled.

l) 250.52 Diabetes with ophthalmic manifestations, type II or unspecified type, uncontrolled.

m) 250.60 Diabetes with neurological manifestations, type II or unspecified type, not stated as uncontrolled.

n) 250.62 Diabetes with neurological manifestations, type II or unspecified type, uncontrolled.

o) 250.70 Diabetes with peripheral circulatory disorders, type II or unspecified type, not stated as uncontrolled.

p) 250.72 Diabetes with peripheral circulatory disorders, type II or unspecified type, uncontrolled.

q) 250.80 Diabetes with other specified manifestations, type II or unspecified type, not stated as uncontrolled.

r) 250.82 Diabetes with other specified manifestations, type II or unspecified type, uncontrolled.

s) 250.90 Diabetes with unspecified complication, type II or unspecified type, not stated as uncontrolled.

t) 250.92 Diabetes with unspecified complication, type II or unspecified type, uncontrolled.

The rest of subjects not coded as T2D patients were considered as controls.

All of Us Research Program

The All of Us Research Program (AoU) is a large, US-funded biobank developed to leverage the diversity of the US for facilitating and improving high-powered genetic and epidemiological studies[12]. We used the AoU short-read whole genome sequencing (srWGS) data from June 22, 2022, for 98,590 participants (release v6).

AoU provides a list of flagged individuals with poor srWGS data quality based on the number of variants, insertion/deletion ratios, number of insertions, number of deletions, heterozygosity of variants or indels, transition/transversion ratios, and number of variants not in gnomAD v3.1. Additionally, variants from srWGS were flagged as being low quality based on a GQ threshold of <20, a DP threshold of <10, an allele balance filter of <0.2, a low QUAL score (60 for SNPs and 69 for indels), or an ExcessHet score <54.69. These individuals and variants were removed from all subsequent analyses.

T2D cases and controls were defined using a modified version of the Northwestern T2D algorithm [13,14]. There are 6 conditions for a participant to be determined as having T2D. 1) A participant has no ICD-based diagnosis of T1D AND has an ICD-based T2D diagnosis AND no insulin prescription in an out-patient setting AND is prescribed non-insulin diabetes medications. 2) A participant has no ICD-based diagnosis of T1D AND has an ICD-based T2D diagnosis AND has

an insulin prescription in an out-patient setting AND has a non-insulin diabetes prescription prescribed before the insulin. 3) A participant has no ICD-based diagnosis of T1D AND has two separate ICD-based T2D diagnoses AND has an insulin prescription in an out-patient setting AND does not have a prescription for a non-insulin diabetes medication. 4) A participant has no ICD-based diagnosis of T1D AND has an ICD-based T2D diagnosis AND no insulin prescription in an out-patient setting AND is not prescribed a non-insulin diabetes medication AND has an abnormal glycemic lab. 5) A participant has no ICD-based diagnosis of T1D AND has no ICD-based T2D diagnosis AND has a prescription for a non-insulin diabetes medication AND has an abnormal glycemic lab. 6) A participant self-reports as having T2D in their intake survey. Abnormal glycemic labs for cases were defined as a fasting glucose $\geq$ 126 mg/dl or a hemoglobin A1c $\geq$ 6.5%.

Controls individuals without diabetes were defined if they had 1) at least one in-person healthcare provider visit, 2) at least one glucose measurement available, 3) no abnormal glycemic lab values, 4) no ICD codes or survey responses for any type of diabetes or related conditions, 5) no prescriptions of diabetes medications. We further excluded controls younger than 45.

In total, we identified 9,812 T2D cases and 15,948 controls with complete genetic data and covariate data (age, sex, and BMI and principal components), which we used in our discovery analysis.

Description of the replication cohorts for the identified rare variants and Type 2 Diabetes definitions

Genetic Epidemiology Research on Adult Health and Aging

The GERA replication cohort (GERA_REP) included individuals from African American, Admixed American, and East Asian ancestry[15] who were not previously included in the discovery GWAS. Because of the multi-ethnic nature of the cohort, different ethnic-specific arrays, based on the Affymetrix Axiom Genotyping System, were employed to enhance genome-wide coverage. The number of SNPs and SNP content varied by array, with SNP content designed to maximize the coverage of low frequency and more common variants specific to the different race-ethnicity while also maximizing coverage of the whole genome and including new SNPs from sequencing projects and SNPs with established associations with disease phenotypes.

The custom array genotyped data was array-wise QCed and imputed to the TOPMed freeze 8 reference panel. Variant quality checks included the removal of variants with 5% or more missing data, MAF < 0.1%, a Hardy-Weinberg equilibrium p-value < 5×10^{-10}, palindromic single nucleotide variants (AT or CG), and duplicates. We also excluded samples with 2% or more missing data.

Clean datasets were phased using SHAPEIT2 and used as input for imputation. We merged the imputed data using IMMERGE[16] to keep all variants present in any of the three imputed chunks with Rsq $\geq$ 0.3. After merging, we converted the VCF files to BGEN format for analysis with REGENIE.

We applied the same T2D definition as with GERA EUR and tested the association of the novel variants with T2D as a binary outcome and included age, sex, BMI, the top 10 PCs, and the imputation batch. In total, GERA_REP included 2,737 T2D cases and 9,270 controls.

All of Us Research Program

The AoU replication cohort (AoU_REP) included non-overlapping individuals and non-family members (pairwise kinship score > 0.1) to those included in the discovery meta-analysis as they were released in the posterior v7 data freeze. We used the AoU srWGS for a total of 17,693 independent T2D cases and 28,918 controls that were QCed and analyzed to test the association with T2D as previously described.

Geisinger MyCode cohort

The Geisinger MyCode Community Health Initiative (GEISINGER) is a biorepository of blood, serum and DNA samples that are linked to electronic health records for the purpose of broad research use. Participants were recruited from Geisinger, an integrated healthcare system located in central and northeastern Pennsylvania. The study sample consisted of 172,366 individuals with exome sequence and microarray genotype data. Available clinical data includes clinical diagnoses, procedures, medications, and laboratory results, and are updated daily[17]. The cohort is 61% female and consists of 93% genetically inferred EUR, 3% AFA, 1.5% AMR, 0.3% SAS, 0.3% EAS, and 1.7% other ancestries. The mean (SD) age at enrollment is 40.3 (1s8.9) years, range 0 to 99. Participation in MyCode is done through written consent under Geisinger Institutional Review Board Study #2016-0269. The results reported here were determined by the Geisinger IRB to meet the criteria for “Non-human subject research” as defined in 45CFR46.102(e). All research was performed in accordance with relevant guidelines and regulations.

DNA Samples from 173,585 MyCode participants were exome sequenced using IDT exome capture/Illumina HiSeq. Of those, 172,366 were genotyped using Infinium OmniExpress Exome array (Illumina), and GSA-24v1-0 array (Illumina) through a collaboration between Geisinger and the Regeneron Genetics Center (Dewey FE et al). Rare coding variants for this study have

genotype quality ≥ 30, read depth ≥ 15, allelic depth ≥ 20, and strand bias p-value ≥ 10%. For array-based genotyping data, SNPs with minor allele frequency ≤ 1%, significant deviation ($p \leq 1x10^{-15}$) from Hardy-Weinberg Equilibrium, and site-level missingness ≥ 1% were removed. Array-based genotypes were imputed to the TOPMED reference panel (97,256 deeply sequenced genomes) with a GRCh38 build using the TOPMED Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov), which employed Eagle v2.4 and Minimac4 as the phasing and imputation algorithm, respectively. All candidate non-coding rare variants in this study have an imputation info score ≥0.3. A set of LD-pruned (200 variant windows, 50 variant sliding windows, and $r^2 < 0.25$), high quality (info score ≥0.3) common variants (919,227) were applied for PCA (PLINK v2.0).

Diabetes was defined using outpatient ICD9/10 codes for diabetes, hbA1c, fasting blood glucose, and antihyperglycemic medications (other than biguanides). Laboratory definition for diabetes used thresholds defined by the American Diabetes Association. In absence of other evidence, individuals with IC9/10 codes for diabetes for at least two encounters more than 3 months apart were included individuals with only an ICD9/10 code for diabetes were excluded from the study. T2D cases were defined 1) having a ratio T1D/T2D ICD9/10 count $\leq$0.50, and 2) the following exclusion: a. on insulin medication only and has $\geq$2 insulin medication prescriptions, or b. 1 ICD9/10 code for insulin pump and has $\geq$2 insulin medication prescriptions, or c. on insulin medication only and diabetes duration $\geq$3 years and has $\geq$2 insulin medication prescriptions, or d. has >1 ICD9/10 for insulin pump. Controls were defined using a modified version of the Northwestern T2D algorithm. Controls were individuals $\geq$45 years of age without an ICD9/10 code for diabetes or antihyperglycemic medications including biguanides, and no HbA1c $\geq$ 5.6% or random glucose $\geq$ 200 mg/dl or fasting glucose $\geq$ 100 mg/dl. The index age for controls was calculated by the subtraction of the last active date (primarily last encounter date) and the birth date. The index age for cases was the age of the first recorded evidence of diabetes.

Genomic-determined sex, index age, and ten major principal components (PCs) were included as covariates in the association study using REGENIE v3.4.1. Only high-quality common variants were used to fit the whole genome regression model for the dichotomous or quantitative traits, and a set of genomic predictions were produced as output of REGENIE Step 1. The candidate rare variants are tested for association in REGENIE Step 2 with Firth logistic regression model to fit the dichotomous traits. Rank Inverse Normal Transformation was applied to the quantitative phenotypes in both Step 1 and Step 2.

REFERENCES

1. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
2. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
3. Delaneau, O., Zagury, J.F., Robinson, M.R., Marchini, J.L. & Dermitzakis, E.T. Accurate, scalable and integrative haplotype estimation. *Nat Commun* **10**, 5436 (2019).
4. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
5. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021).
6. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).
7. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097-1103 (2021).
8. Eastwood, S.V. *et al.* Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PLoS One* **11**, e0162388 (2016).
9. Karlson, E.W., Boutin, N.T., Hoffnagle, A.G. & Allen, N.L. Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. *J Pers Med* **6**(2016).
10. Yu, S. *et al.* Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc* **22**, 993-1000 (2015).
11. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
12. All of Us Research Program, I. *et al.* The "All of Us" Research Program. *N Engl J Med* **381**, 668-676 (2019).
13. Hripcsak, G. *et al.* Facilitating phenotype transfer using a common data model. *J Biomed Inform* **96**, 103253 (2019).
14. Jennifer Pacheco, W.T. Type 2 Diabetes Mellitus. (PheKB, 2012).
15. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285-95 (2015).
16. Zhu, W. *et al.* IMMerge: merging imputation data at scale. *Bioinformatics* **39**(2023).

17. 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).