The power of TOPMed imputation for the discovery of Latino-enriched rare variants associated with type 2 diabetes.

Alicia Huerta-Chagoya, Philip Schroeder, Ravi Mandla, Aaron J. Deutsch, Wanying Zhu, Lauren Petty, Xiaoyan Yi, Joanne B. Cole, Miriam S. Udler, Peter Dornbos, Bianca Porneala, Daniel DiCorpo, Ching-Ti Liu, Josephine H. Li, Lukasz Szczerbiński, Varinderpal Kaur, Joohyun Kim, Yingchang Lu, Alicia Martin, Decio L. Eizirik, Piero Marchetti, Lorella Marselli, Ling Chen, Shylaja Srinivasan, Jennifer Todd, Jason Flannick, Rose Gubitosi-Klug, Lynne Levitsky, Rachana Shah, Megan Kelsey, Brian Burke, Dana M. Dabelea, Jasmin Divers, Santica Marcovina, Lauren Stalbow, Ruth J.F. Loos, Burcu F. Darst, Charles Kooperberg, Laura M. Raffield, Christopher Haiman, Quan Sun, Joseph B. McCormick, Susan P. Fisher-Hoch, Maria L. Ordoñez, James Meigs, Leslie J. Baier, Clicerio González-Villalpando, Maria Elena González-Villalpando, Lorena Orozco, Lourdes García-García, Andrés Moreno-Estrada, Mexican Biobank, Carlos A. Aguilar-Salinas, Teresa Tusié, Josée Dupuis, Maggie C.Y. Ng, Alisa Manning, Heather M. Highland, Miriam Cnop, Robert Hanson, Jennifer Below, Jose C. Florez, Aaron Leong and Josep M. Mercader

ELECTRONIC SUPPLEMENTAL MATERIAL (ESM)

ESM METHODS

Discovery sample

We aggregated data from 6 Latino cohorts with a total sample size of 18,885 individuals (8,150 cases and 10,735 controls): SIGMA1, SIGMA2, SIGMA3, MXBB, MGB and GERA cohorts (Figure 1, ESM Table 1).

The Slim Initiative for Genomic Medicine in the Americas (SIGMA) Cohorts [1–3] include individuals of Mexican or Mexican American origin recruited from both hospital- and population-based studies. The SIGMA study comprised of three different subsets. The SIGMA1 subset consists of 4,190 individuals, the SIGMA2 subset is composed of 3,730 individuals with a high proportion of Central and South America Native ancestry and the SIGMA3 subset includes 5,790 individuals. All human research was approved by the Comité de Ética e Investigación del Centro de Estudios en Diabetes, the Ethics and Research Committees of the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, the Ethics and Research Committes of the Instituto Nacional de Medicina Genómica and the Federal Commission for the Protection against Health Risks (COFEPRIS) (CAS/OR/CMN/113300410D0027-0577/2012).

The Mexican Biobank Cohort (MXBB)[4] includes population-based individuals from the 2000 Mexican National Health Survey conducted between the years 1999 and 2000 in Mexico. In the present study, we used a subset of 1,730 individuals for whom DNA genotyping data and type 2 diabetes definition were available. MXBB studies involving human participants were reviewed and approved by the Research Ethics Committee (Approval CI-1479) of the National Institute of Public Health, Mexico). The patients/ participants provided their written informed consent to participate in this study.

The Mass General Brigham (MGB) Biobank[5] maintains blood and DNA samples of patients seen at MGB hospitals, including the Massachusetts General Hospital, Brigham and Women’s Hospital, McLean Hospital and Spaulding Rehabilitation Hospital, all in the Boston area (U.S.). We analyzed a subset of 2,115 genetically identified Latino samples. MGB work was conducted with approval from the MGB Institutional Review Board (study 2016P001018). It acknowledges the Partners HealthCare System for support of the MGB biobank and MGB patients for providing samples, genomic data, and health information data.

The Genetic Epidemiology Research on Aging cohort (GERA)[6] includes over 100K adults who were members of the Kaiser Permanente Medical Care Plan in the Northern California Region of the U.S. In this study, we analyzed a subset of 1,330 genetically identified Latino samples.

For each cohort, we selected Latino samples based on their genetically estimated ancestry, as described below. We calculated the principal components (PCs) using common genetic variants (minor allele frequency [MAF]≥5%) using FlashPCA[7] on both self-reported Latino individuals and the 1000G Phase 3 samples. Then, we separately estimated the mean of the top 4 PCs on either Latino samples or each of the four continental superpopulations defined in the 1000G Project (*i.e.* European, African, East Asian, and South Asian). For MGB and GERA, we applied a first filtering step to remove Latino samples within 2 SD of the mean for PCs 1-4 for each superpopulation cluster. This step was not applied to SIGMA and MXBB cohorts since they showed more homogeneous genetic ancestries that fall on the expected cline of Native American and European ancestry. For all discovery cohorts, we applied a second step where outlier Latino samples lying more than 4 SD away from the mean PC for its own Latino cluster were also removed. For all cohorts, we estimated the genetic-based global ancestry using

Admixture[8]. After excluding high linkage disequilibrium (LD) regions, we performed LD pruning in each dataset, using a window 50kb, a shift of 5 variants at each step, and LD between variants $r^2>0.2$. We then merged non-ambiguous alleles with 1000Gp3 samples and ran ADMIXTURE on the output file assuming 1-10 five ancestral populations (K=1-10). We used a cross validation procedure to identify the best K. We used the default termination criteria for stopping the prior estimation algorithm, that is a log-likelihood increase by less than the convergence criterion $\in=10^{-4}$ between iterations. We ran the same procedure 10 times and found an average pairwise similarity>0.99 among the 10 runs in all cohorts. (ESM Fig. 1).

Genotyping, quality control and imputation

Genotyping was done using several commercially available genome-wide arrays. A subset of the SIGMA samples (N=9,520) underwent whole-exome sequencing (WES), which we integrated with the genotyping array data to improve the imputation backbone (Table S1). We applied pre-imputation quality control to each dataset separately. Variant quality checks included the exclusion of variants with 5% or more missing data, MAF<0.1%, a Hardy-Weinberg equilibrium p-value$<5\times10^{-10}$, palindromic single nucleotide variants (AT or CG), duplicates, as well as a case *vs.* control missingness difference *P*-value<0.00005. We also excluded samples with 2% or more missing data, that did not pass the sex check or that were closely related as estimated by the proportion of identity by descent (IBD>0.185) between pairs of individuals. Clean datasets were phased using SHAPEIT2[9] and used as input for imputation.

For comparison purposes, we imputed the phased haplotypes using both 1000G Phase 3 version 5[10] and TOPMed reference panels freeze 8[11] in each cohort, separately. We used Michigan imputation and TOPMed imputation servers to impute genotypes.[12–14] For chromosome X, we imputed non-pseudoautosomal regions from females and males separately.

Imputation performance evaluation

We used the r^2 quality measure, as reported by the Michigan and TOPMed imputation servers. It calculates the ratio of the empirically observed variance of the allele dosage to the expected binomial variance.[15] We evaluated the performance of TOPMed and 1000G imputations by summarizing the chromosome-wise r^2 quality measure and the number of well-imputed ($r^2 \geq 0.8$) variants at different allele frequency thresholds. To further test if the quality of imputation was well-powered to detect low-frequency and rare variation without relying on the imputation server's quality measures, we leveraged available WES data from SIGMA3 cohort and estimated the proportion of the sequenced variants, for chromosome 22 only, that were well-imputed with TOPMed and 1000G panels at different WES allele frequency thresholds. Loss-of-Function (LoF) variants are of clinical interest as they disrupt protein-coding genes, potentially being disease-causal. Therefore, we used SnpEff[16] to annotate the WES variants for chromosome 22 only and estimated the percentage of well-imputed variants identified with TOPMed and 1000G imputations. We used CADD score[17] to predict the deleteriousness of variants. We classified the variants using the score cut-offs of 10 and 20, below which the variants were considered as benign.

We also calculated the effective sample size (Neff) needed to reach 80% statistical power to detect genome-wide significant associated signals ($\propto$=5×10^{-8}) at different effect sizes and allele frequencies covered by the imputations. Neff was calculated as derived by Grotzinger et al. [18]:

$$Neff = 4({}^{1}/_{N\ cases} + {}^{1}/_{N\ controls})$$

Type 2 diabetes association and meta-analysis

Type 2 diabetes association analyses were performed in each cohort with SNPTEST using the expectation maximization (*em*) method[19] for doing maximum likelihood estimation in a generalized linear model framework using genotype probabilities under the additive model. Models were adjusted for sex, age, body mass index (BMI) and 10 PCs to account for population structure. We ran additional models without adjusting for BMI. Only well-imputed variants ($r^2 \geq 0.5$) were meta-analyzed using the inverse of the corresponding squared standard errors in METAL.[20] The statistical significance threshold was set to $P<5\times10^{-8}$.

We performed LD-based clumping on the genome-wide significant variants to keep one representative variant per region of LD. We set an LD r^2=0.5 and a distance between variants of 250 kb. If the variant was located within a previously reported type 2 diabetes-related *locus*, we used a conditioning strategy to test for distinct signals. We conditioned on each of the previously reported SNPs within the *locus*. We performed conditional analyses in each cohort separately, and then meta-analyzed using the inverse of the variance of the effect estimates. If our lead variant showed evidence of residual type 2 diabetes association (p-value<5×10^{-5}) after conditioning on any previously reported variant within the *locus*, we considered our novel signal as distinct.

Variants with sub-genome-wide significance ($p<1\times10^{-6}$) that were only imputed with the TOPMed reference panel, showed increased frequency in the Latino population and were > 250 kb from other reported genome-wide significant variants from large consortia analyzing either European or East Asian populations[21, 22] were considered for further investigation.

Replication sample

Variants associated with type 2 diabetes risk at genome-wide and sub genome-wide significance were tested for replication in six independent cohorts described below (Table S2).

Progress in Diabetes Genetics in Youth (PRODIGY)

PRODIGY is the largest cohort for youth-onset type 2 diabetes with available GWAS data, comprising the Treatment Options for Type 2 Diabetes in Adolescents and Youth (TODAY)[23] and the SEARCH for Diabetes in Youth studies[24]. TODAY includes participants diagnosed with type 2 diabetes prior to 18 years of age and have documented BMI≥85th percentile at the time of diagnosis. Of note, the American Indian tribal nations that partnered on the TODAY Study elected not to participate in the genomics collection. SEARCH is a population-based prospective registry study launched in 2000 that ascertained diabetes in youth diagnosed at <20 years of age in the U.S. Overall, PRODIGY has shown consistent direction and size of effects for most genetic variants associated with type 2 diabetes in adults.[25] The TODAY and SEARCH protocols were approved by the institutional review boards of each participating institution. Participants provided written informed parental consent and child assent, including consent and assent specifically for genetic testing.

We identified 1,198 youth type 2 diabetes cases of Latino descent. As the control group, we used 1,805 diabetes-free adult Latino samples from the T2D-GENES (Type 2 Diabetes Genetics Exploration by Next-generation sequencing in multi-Ethnic Samples) and from the METS (Mexican Metabolic Syndrome) Cohort.[26] Genotypes from both datasets were merged and the pre-imputation quality control additionally included the exclusion of variants with genotyping array missingness difference (*p*-value<0.00005). Phasing was done as described above and genotypes were imputed to the TOPMed reference panel to ensure high-quality imputation. Type 2 diabetes association was tested under an additive model using SNPTEST and *em* method[19]. Models were adjusted for sex and 10 PCs to account for population structure. Given the case-control design of this replication cohort, age and BMI were not considered as covariates.

Cameron County Hispanic Cohort (CCHC)

CCHC is a population-based cohort of Mexican American individuals living in the U.S.[27] We selected 971 type 2 diabetes cases and 857 controls. We used high-quality TOPMed imputed genotypes (r^2>0.9) and tested for type 2 diabetes association under additive models adjusted for sex, age, BMI and 10 PCs. Human research was approved by the relevant Institutional Review Boards. All participants provided written informed consent. CCHC was approved by the Committee for the Protection of Human Subjects at the University of Texas Health Sciences Center at Houston, Human Research Protections Program at Vanderbilt University.

Urban American Indians and Arizona Pima Indians cohorts

We also used 851 type 2 diabetes cases and 2,191 controls from four groups of urban-dwelling American Indians living in or near Phoenix, Arizona, as well as 2,571 type 2 diabetes cases and 5,088 controls from a community of Pima Indians in Arizona, who participated in a longitudinal study of type 2 diabetes.[28] For both cohorts, imputation was done using Pima whole-genome sequences from 266 individuals. Variant rs1016378028, was directly genotyped using Taqman probes. type 2 diabetes association was tested by fitting additive models adjusted for sex, age, BMI and 5 PCs. In the Pima cohort, we additionally adjusted for birth year since exams took place over many years. In the Pima cohort, we used linear mixed models to account for estimated relatedness among individuals, whereas in the Urban American Indians cohort, we used a genomic control procedure. Human research was approved by the relevant Institutional Review Boards. All participants provided written informed consent.

Population Architecture using Genomics and Epidemiology (PAGE) study

The PAGE study aims to conduct genetic epidemiological research in ancestrally diverse populations within the U.S.[29] It includes four population-based cohorts with significant numbers of Latino participants: the Hispanic-Community Health Study/Study of Latinos (HCHS/SOL), the Women's Health Initiative (WHI), the Multiethnic Cohort (MEC) and the Icahn School of Medicine at Mount Sinai BioMe biobank in New York City (BioMe). Genotyped individuals self-identified as Hispanic/Latino, African American, Asian, Native Hawaiian, Native American or other. For this study, Latino samples were selected based on their genetically estimated ancestry following the same two-step filtering approach that we implemented with the discovery cohorts. We analyzed up to 6,761 type 2 diabetes cases and 5,747 controls. We used high-quality TOPMed imputed genotypes (r^2>0.9) and tested for type 2 diabetes association under additive models adjusted for sex, age, BMI and 14 PCs. Analyses were conducted in SUGEN[30] to account for relatedness within datasets. Human research was approved by the relevant Institutional Review Boards. All participants provided written informed consent.

All of Us Research Program

The All of Us Program aims to build a national longitudinal resource of multiple data types and biosamples from at least one million individuals in the U.S., with the main goal of broadly reflecting the diversity in the country.[31] We analyzed whole genome sequencing data from 8,958 genetically identified Admixed American/Latino individuals, of which 1,243 were type 2 diabetes cases and 7,715 were controls. We tested the type 2 diabetes association under additive models adjusted for sex, age, BMI, and 16 PCs. All participants consent to participate. The work described here was confirmed as meeting criteria for non-human subject research by the *AoU* Institutional Review Board. All methods were carried out in accordance with relevant guidelines and regulations.

Replication in non-Latino populations

Since one of our novel variants, rs2891691, is also prevalent among African (MAF=16%) and East Asian populations (MAF=7.6%), we tested its replication in both ancestries. For East Asian ancestry, we leveraged publicly available summary statistics from Vujkovic et al., 2020 [32] (46,511 T2D cases and 169,776 T2D controls, Neff=142,087). For African ancestry, the largest publicly available summary statistics dataset is from Vujkovic et al., 2020 [32] ((31,446 T2D cases and 56,092 T2D controls, Neff=55,217)). To increase the sample size and statistical power, we leveraged 4 additional datasets of African ancestry for which he had available individual-level data. The 4 cohorts are: the UKBB (Neff=2,516), the MGB[5] (Neff=1,078), the GERA [6] (N=1,563) and the All of Us [31] (N=6,395). Except for the All of Us cohort that has whole exome sequencing data, we performed quality control and imputed each cohort to the TOPMed panel, as described for the Latino analyses. We used high quality imputed variants to perform a type 2 diabetes GWAS in each cohort, separately. Then, we meta-analyzed the results using the inverse of the corresponding squared standard errors with METAL. The Neff of the T2D GWAS African meta-analysis was 66,769.

To aggregate the rs2891691 allelic effects across Latino, African, and East Asian ancestries, we performed a fixed-effects meta-analysis. Additionally, to allow for heterogeneity in allelic effects correlated with ancestry, which is not accommodated with a fixed-effects meta-analysis, we used MR-MEGA software [33]. It implements a multi-ancestry meta-regression approach to model the allelic effects as a function of axes of the genetic variation, which are derived from a matrix of mean pairwise allele frequency differences between GWAS. Specifically for this analysis, we did not include the All of Us cohort, as we did not have available GWAS data (ESM Fig. 3).

Association with type 2 diabetes-related phenotypes

Given the lack of large-scale publicly available biobanks with Latino samples that may allow for better characterization of our novel signals, including those occurring at a low frequency, we assembled a collection of cohorts to perform QTL analyses focused on 46 glycemic, anthropometric and lipid traits. In addition to 5 of the Latino cohorts analyzed in the type 2 diabetes meta-analysis (*i.e.* SIGMA1, SIGMA2, SIGMA3, MXBB and MGB Biobank), we included three extra cohorts, which we also imputed to the TOPMed panel as described above: the METS Cohort, the Mexican Hypertriglyceridemia (MHTG) Cohort, as well as the genetically identified Latino samples from the UK Biobank (UKBB).[34] The MHTG study was reviewed and approved by the Ethics and Research Committees from the Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran and UCLA Institutional Review Board MIRB1. The UK Biobank has obtained ethical approval covering this study from the National Research Ethics Committee (REC reference 11/NW/0382).

We also analyzed the Nightingale Nuclear Magnetic Resonance-based panel of 168 metabolomic biomarkers in Latino samples from the UKBB. The panel provides measures spanning multiple pathways, including lipoprotein lipids, fatty acids, amino acids, ketone bodies and glycolysis metabolites.

QTL analyses were conducted using high-quality TOPMed imputed genotypes and cleaned phenotypes from non-pregnant Latino adults. We used inverse rank normal transformation when the normality assumption was not met. Association analyses were done with a maximum of 26,400 adult Latino individuals depending on the trait, of whom 19,459 were diabetes-free. We used SNPTEST and *em* method to run linear regressions assuming additive genetic models in each cohort, separately. Models were adjusted for sex, age, BMI, and 10 PCs. If the outcome was available in >1 cohort, we meta-analyzed the results using the fixed-effects inverse variance method.

Credible sets

For each novel variant, we identified the set of variants with 99% probability of containing the causal variant. We used a Bayesian method[35], considering variants in LD with the lead variant (r^2>0.1). We calculated LD using genetic data from 1,996 Hispanic/Latino samples from TOPMed Freeze 5b. For each region, an approximate Bayes factor (ABF) was calculated for each variant as follows:

$$ABF = \sqrt{1 - r\, e^{(\frac{rz^2}{2})}}$$

where $r = {}^{0.04}/_{(SE^2 + 0.04)}$ and $z = {}^{\beta}/_{SE}$

The β and the SE are the estimated effect size and the corresponding standard error, respectively, that result from testing the variant association under a logistic regression model.

The posterior probability for a variant being causal is equal to its ABF divided by the sum of all ABF values for the *locus*. Large values of the Bayes factor correspond to strong evidence for association. Therefore, variants are ranked by ABF in decreasing order, and the cumulative probability is calculated starting at the top of the list until the value exceeds 99%.

Genomic annotation of the 99% credible set variants

We used the 99% credible sets for each novel signal to annotate their genomic effect using the VEP[36] (GRCh38.p7) and SNPNEXUS[37] applications. This allowed us to gather information about gene and protein proximity, the effect of non-synonymous coding variants on protein function (SIFT, PolyPhen), non-coding variants scoring (CADD, FunSeq2), and the occurrence of regulatory elements.

We used GTEx V8[38] to assess the influence of the variants in gene-level expression, as well as the TIGER Portal[39] for evaluating the gene-level expression in pancreatic islets and the Islet Gene View[40] for assessing the gene co-expression in human islets. We also assessed individual variant associations with a variety of common phenotypes and diseases using the Common Metabolic Disease Knowledge Portal (cmdgenkp.org. 2021 Nov 15), as well as other resources, including web servers such as UK Biobank imputed with TOPMed (https://pheweb.org/UKB-TOPMed/) and FinnGen (https://r6.finngen.fi/). We also assessed gene-phenotype associations using the Genebass browser[41], a resource of exome-based association statistics across 281,852 individuals with exome sequence data from the UKBB.[42]

To obtain more evidence implicating the variants or their closest genes, with any disease or biological process, we used the Open Targets Platform, which aggregates a variety of resources and scores the collected information to contextualize and weigh the underlying evidence.[43]

Expression of genes near novel variants

We also assessed the expression levels of the genes ± 500 kb around the novel signals in human islets under different conditions pertaining to type 1 diabetes and type 2 diabetes. We examined the following datasets: Type 1 diabetes dataset (FACS-purified β cells obtained from donors with type 1 diabetes) downloaded from GSE121863 (n=4 type 1 diabetes versus 12 controls)[44]; type 2 diabetes samples gathered and integrated from three cohorts: one generated by T2DSystems (TIGER Portal), and the other two downloaded from the Gene Expression Omnibus (GEO) database under accession numbers GSE159984 and GSE50244 (a total of n= 47 type 2 diabetes cases versus 228 controls)[39, 45, 46]; brefeldin A-exposed human islets (0.1 µg/mL for 24h) downloaded from GEO under accession number GSE152615 (n=4)[47]; cytokine-exposed human islets (exposed to interferon-γ (1,000 U/ml) and interleukin-1β (50 U/ml) for 48 h or to interferon-α (2000 U/mL) for 2h, 8h, or 18h, n=5-6 per condition) from GSE108413 and GSE133221[48, 49]; human islets exposed to palmitate (0.5 mM), high glucose (22.2 mM) and palmitate plus high glucose for 48h (n=3-5 per condition) from GSE159984[45]. Gene expression differences between groups were assessed using *p*-values and adjusted *p*-values (Benjamini Hochberg correction) determined by the Wald test using the DESeq2 pipeline.[50] Transcript per million (TPM) was normalized by Salmon 1.4.0.[51]

Polygenic scores

Polygenic scoring using single ancestry summary statistics and LD reference panels was calculated via Bayesian Regression and Continuous Shrinkage priors as implemented in PRS-CS.[52] We used the UKBB LD reference panel and GWAS summary statistics from European[22], East Asian[21] and Latin American populations. GWAS Latino summary statistics were calculated using a meta-analysis with five of the discovery cohorts (*i.e.* SIGMA1, SIGMA2, SIGMA3, MGB, and GERA). Five phi shrinkage priors (Φ) were used (*i.e.* 1, 10^{-2}, 10^{-4}, 10^{-6} and the one automatically estimated from the data). Then, we used the estimated posterior SNP effect sizes for each ancestry to calculate the PSs in a training cohort (*i.e.* MXBB).

To evaluate the performance of the PSs, we first fitted a simple model that included sex, age and 10 PCs to account for population stratification. We then fitted models that additionally included the PS standardized scores. We calculated the variance explained in type 2 diabetes status for each model using the Nagelkerke pseudo r^2. We repeated the same strategy by comparing the individuals above different percentile cutoffs with those in the interquartile of the PS. The best shrinkage prior was selected based on the larger incremental r^2 of type 2 diabetes status in the MXBB training cohort. Then, the selected model was tested in a target cohort (*i.e.* the METS Cohort).

Given that the ancestry-specific PSs were not highly correlated (r^2<0.3), we also used PRS-CSx[53], a novel method that improves cross-population polygenic prediction by integrating GWAS summary statistics from multiple populations. For a given shrinkage prior, PRS-CSx returns posterior SNP effect size estimates for each discovery population, which we used to calculate the cross-population PS. First, we fitted a linear regression combining the standardized scores for each population in the MXBB training cohort, as follows:

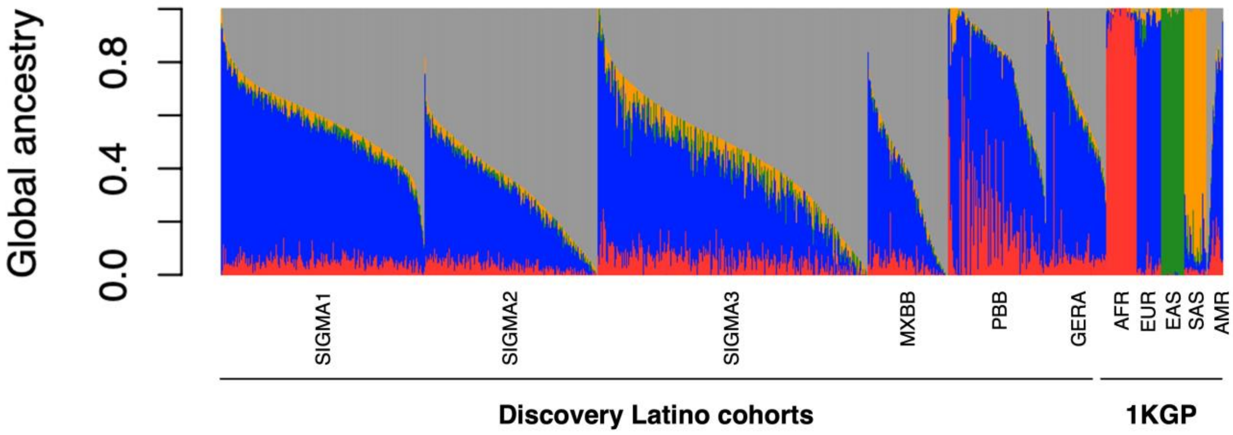
$$y \sim B_{\emptyset,Pop1} PRS_{\emptyset,Pop1} + \ B_{\emptyset,Pop2} PRS_{\emptyset,Pop2} + \cdots + \ B_{\emptyset,K} PRS_{\emptyset,K}$$

where y is type 2 diabetes status, $B_{\Phi,Pop}$ is the regression coefficient for a given phi shrinkage prior and population and $PS_{\Phi,Pop}$ is the standardized PS for a given phi shrinkage prior and population. The phi shrinkage prior and corresponding regression coefficients for the linear combination of PS that maximizes the incremental Nagelkerke pseudo R^2 were used in the METS target cohort to evaluate the performance of the cross-population PS.

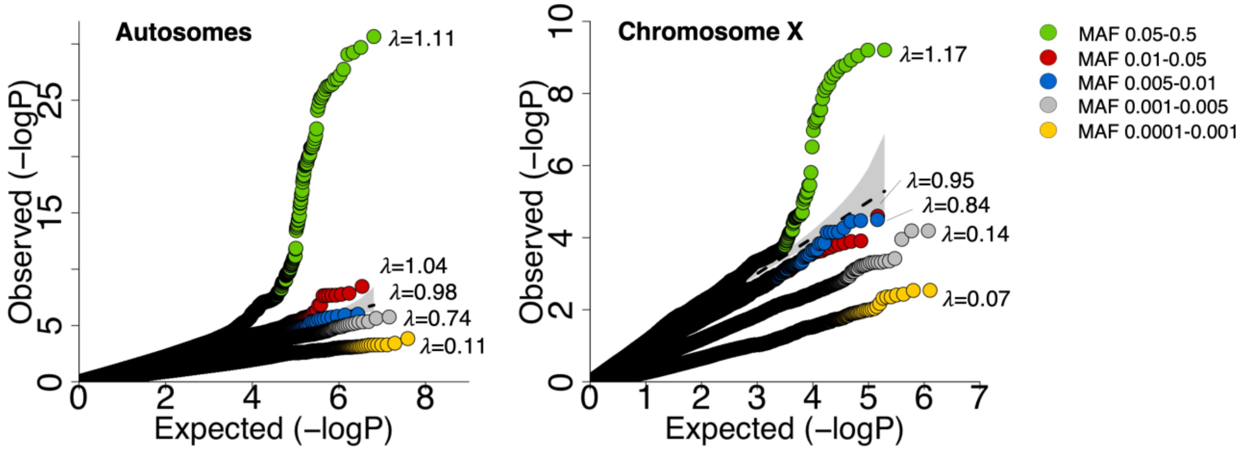
$$PRSCSx = \hat{B}_{\emptyset,Pop1}PRS_{\emptyset,Pop1} + \hat{B}_{\emptyset,Pop2}PRS_{\emptyset,Pop2} + \cdots + \hat{B}_{\emptyset,K}PRS_{\emptyset,K}$$

We also calculated the AUC either for the covariates sex, age and 10 PCs or the cross-population PS plus the covariates. The DeLong test was used to assess the difference between AUCs. We then calculated the OR per standard deviation in the cross-population PS, adjusting for sex, age and 10 PCs. Finally, we identified the high-risk individuals at the top 2.5%, 5% and 10% of the cross-population PS distribution and calculated the OR of the high-risk individuals *versus* the interquartile distribution.

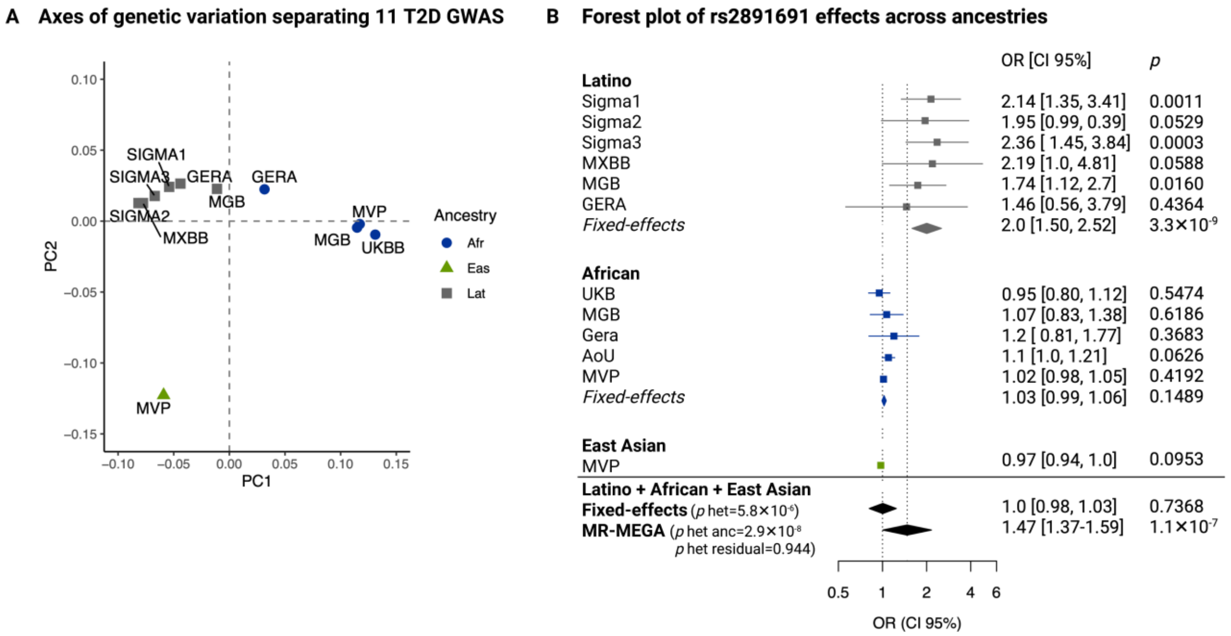
ESM FIGURES



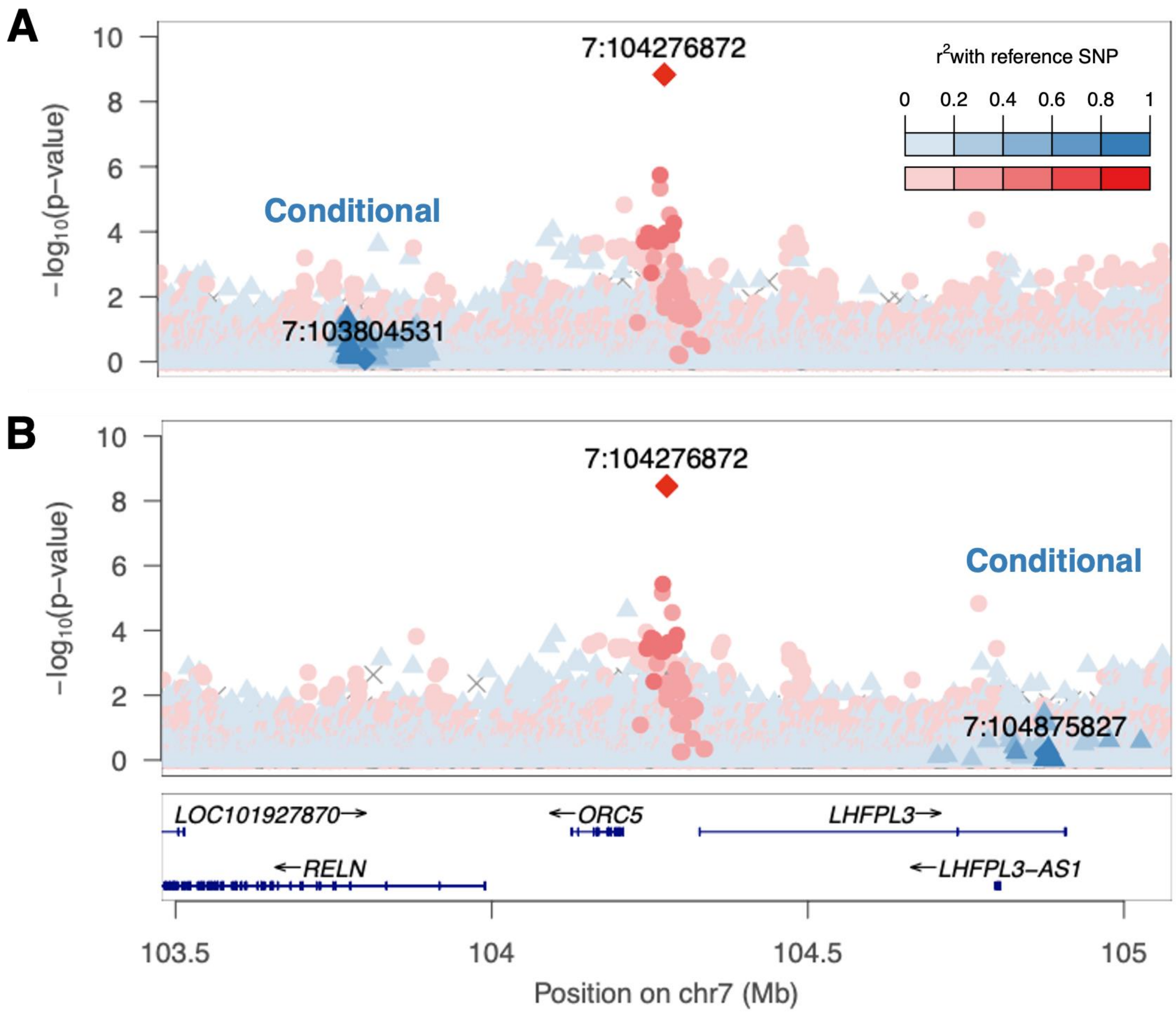
ESM Fig. 1. Individual global ancestry estimation in the discovery Latino cohorts. Global ancestry proportions were estimated after merging genotypes with 2,504 individuals from the 1000G phase 3, using ADMIXTURE software at K=5. Colors represent the super-continental ancestries: African (red), European (blue), East Asian (green) and South Asian (yellow). Grey color represents Admixed American ancestry.



ESM Fig. 2. QQ plots of association statistics in GWAS. Plots show the calibration under the null and enrichment of T2D-associated variants in the tail for autosomes and chromosome X, stratifying by minor allele frequency. There was a minimal inflation of test statistics for variants with an allele frequency of 0.05 or higher, as indicated by a slight early departure between observed and expected *P*-values of the QQ plots (lambda=1.1). For low-frequency or rare variants, we observed deflated QQ plots, which is expected given the large sample sizes needed to reach statistical power.

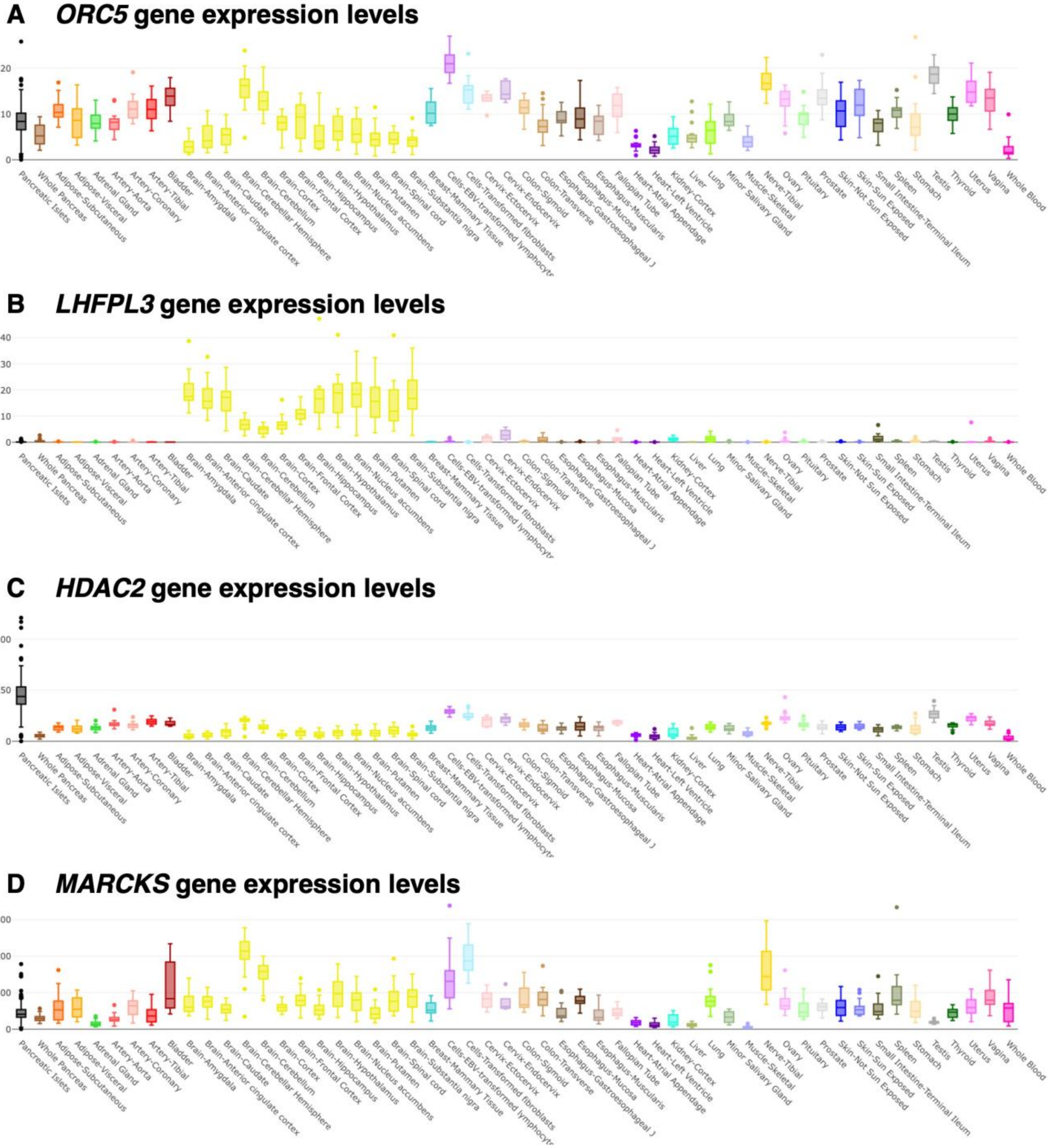


ESM Fig. 3. Meta-analysis of the association of rs2891691 with type 2 diabetes across ancestries across Latino, African, and East Asian ancestries, where the variant is present. a. Forest plot of rs2891691 allelic effects across each ancestry, as well as the fixed-effects and MR-MEGA multi-ancestry meta-regression effects across ancestries. b. Axes of genetic variation separating the 11 T2D GWAS used to model the multi-ancestry allelic effects of rs2891691 with type 2 diabetes.

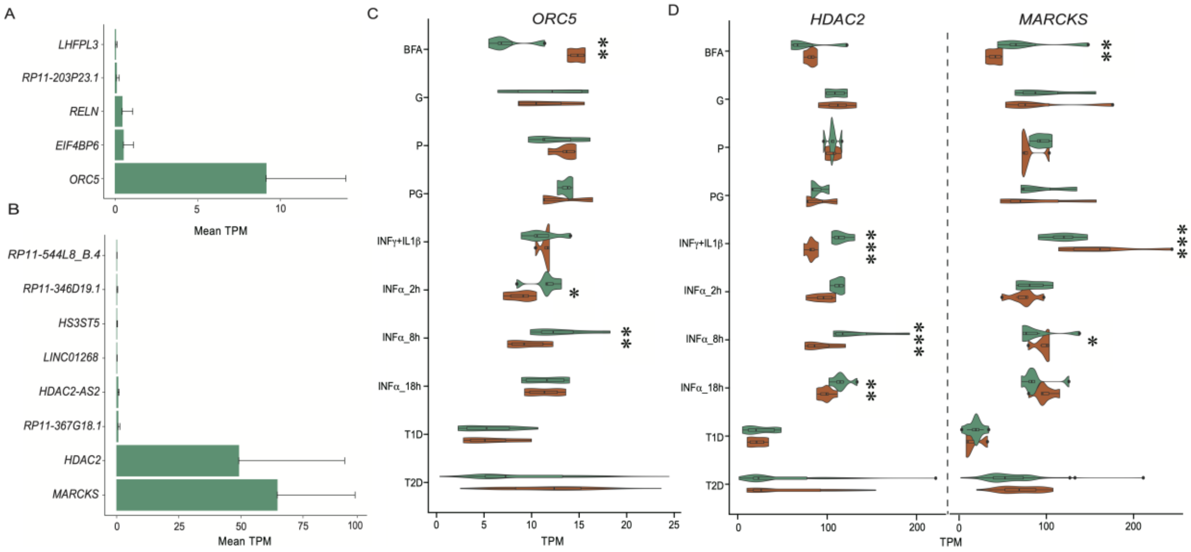


ESM Fig. 4. Regional plot of the association at novel *ORC5/LHFPL3 locus* conditional on two nearby T2D-associated signals. Two type 2 diabetes-associated signals near the rs2891691 variant have been reported in Europeans. One is located within the upstream *RELN* gene (a) (rs39328, b38 chr7:103,804,531)[22], while the other is located within the downstream *LHFPL3* gene (b) (rs73184014, b38 chr7:104,875,827)[32]. rs2891691 was not in LD with either of these variants (r^2<0.0006), and neither of them was associated in our Latino meta-analysis (*p*=0.30 and *p*=0.07, respectively). After conditioning on each of them, rs2891691 remained significant (OR

[95% CI] =2.01 [1.59-2.53], $p=3.1\times10^{-9}$ and OR [95% CI] =2.01 [1.59-2.53], $p=3.5\times10^{-9}$, respectively). Red color intensity indicates r^2 to the novel variant rs2891691. Blue color intensity indicates r^2 to the conditional variants.



ESM Fig. 5. Expression levels of genes around the two Latino T2D-associated leading variants. Multiple tissue and human islets expression levels (TPM) from GTEx and TIGER portals.



ESM Fig. 6. Expression levels of genes around the two Latino T2D-associated leading variants in human islets under different conditions. a. Expression levels (mean TPM+SD) of genes ± 500 Kb to *ORC5* lead variant in human islets under control condition. b. Expression levels (mean TPM+SD) of genes ± 500 Kb to *HDAC2* lead variant in human islets under control condition. c. Expression level of *ORC5* and d. *HDAC2* and *MARCKS* in human islets under different conditions (in orange, compared to control in green) and in islets of patients with T1D or T2D. Asterisks show the adjusted (Benjamini Hochberg correction) *p* value * <0.05; ** <0.01; *** <0.001. BFA brefeldin A, G high glucose, P palmitate, PG palmitate + high glucose; specific conditions and data sources are provided in the Methods section.

ESM CONTRIBUTORS OF MEXICAN BIOBANK

Leadership

Andrés Moreno Estrada	Advanced Genomics Unit, CINVESTAV Irapuato, Mexico
Lourdes García García	National Institute of Public Health, Cuernavaca, Mexico
Adrian V.S. Hill	The Jenner Institute, University of Oxford, United Kingdom

Co-PIs

Selene Fernández Valverde	Advanced Genomics Unit, CINVESTAV Irapuato, Mexico
Teresa Tusié Luna	National Institute of Medical Sciences, Mexico City, Mexico
Carlos Aguilar Salinas	National Institute of Medical Sciences, Mexico City, Mexico
Celia Alpuche Aranda	National Institute of Public Health, Cuernavaca, Mexico

Alexander J. Mentzer	Wellcome Centre for Human Genetics, Oxford, United Kingdom

MXB Analysis Working Group

Mashaal Sohail	Advanced Genomics Unit, CINVESTAV Irapuato, Mexico
Consuelo Quinto Cortes	Advanced Genomics Unit, CINVESTAV Irapuato, Mexico
María José Palma Martínez	Advanced Genomics Unit, CINVESTAV Irapuato, Mexico
Carmina Barberena Jonas	Advanced Genomics Unit, CINVESTAV Irapuato, Mexico
Santiago G. Medina Muñoz	Advanced Genomics Unit, CINVESTAV Irapuato, Mexico
Luis Pablo Cruz Hervert	National Institute of Public Health, Cuernavaca, Mexico
Leticia Ferreyra Reyes	National Institute of Public Health, Cuernavaca, Mexico
Guadalupe Delgado Sánchez	National Institute of Public Health, Cuernavaca, Mexico
Hortensia Moreno Macías	Metropolitan Autonomous University, Mexico City, Mexico

Sergio Adrian Cortés	Wellcome Centre for Human Genetics, Oxford, United Kingdom
Kathryn Auckland	Wellcome Centre for Human Genetics, Oxford, United Kingdom
Amanda Chong	Wellcome Centre for Human Genetics, Oxford, United Kingdom
Genevieve Wojcik	John Hopkins University, Baltimore, United States
Christopher R. Gignoux	University of Colorado, Denver, United States
Alexander Ioannidis	Stanford University, United States

ESM REFERENCES

1. Williams Amy AL, Jacobs Suzanne SBR, Moreno-Macías H, et al (2014) Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature 506(7486):97–101. https://doi.org/10.1038/nature12828
2. Estrada K, Aukrust I, Bjørkhaug L, et al (2014) Association of a low-frequency variant in HNF1A with type 2 diabetes in a latino population the SIGMA Type 2 Diabetes Consortium. JAMA - Journal of the American Medical Association 311(22):2305–2314. https://doi.org/10.1001/jama.2014.6511
3. Mercader JM, Liao RG, Bell AD, et al (2017) A Loss-of-Function Splice Acceptor Variant in IGF2 Is Protective for Type 2 Diabetes. Diabetes 66(11):2903–2914. https://doi.org/10.2337/db17-0187

4. Sepúlveda J, Tapia-Conyer R, Velásquez O, et al (2007) Diseño y metodología de la Encuesta Nacional de Salud 2000. Salud Publica Mex 49(Supplement 3):427–432

5. Karlson EW, Boutin NT, Hoffnagle AG, Allen NL (2016) Building the partners healthcare biobank at partners personalized medicine: Informed consent, return of research results, recruitment lessons and operational considerations. J Pers Med 6(1):1–11. https://doi.org/10.3390/jpm6010002

6. Banda Y, Kvale MN, Hoffmann TJ, et al (2015) Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics 200(4):1285–1295. https://doi.org/10.1534/genetics.115.178616

7. Abraham G, Qiu Y, Inouye M (2017) FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics 33(17):2776–278. https://doi.org/10.1093/bioinformatics/btx299

8. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9):1655–1664. https://doi.org/10.1101/gr.094052.109

9. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9(2):179–181. https://doi.org/10.1038/nmeth.1785

10. Sudmant PH, Rausch T, Gardner EJ, et al (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526(7571):75–81. https://doi.org/10.1038/nature15394

11. Kowalski MH, Qian H, Hou Z, et al (2019) Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet 15(12):e1008500. https://doi.org/10.1371/journal.pgen.1008500

12. Taliun D, Harris D, Kessler M, et al (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590(7845):290–299. https://doi.org/10.1101/563866

13. Das S, Forer L, Schönherr S, et al (2016) Next-generation genotype imputation service and methods. Nat Genet 48(10):1284–1287. https://doi.org/10.1038/ng.3656

14. Fuchsberger C, Abecasis GR, Hinds DA (2015) Minimac2: Faster genotype imputation. Bioinformatics 31(5):782–784. https://doi.org/10.1093/bioinformatics/btu704

15. Liu Q, Cirulli ET, Han Y, Yao S, Liu S, Zhu Q (2015) Systematic assessment of imputation performance using the 1000 Genomes reference panels. Brief Bioinform 16(4):549–562. https://doi.org/10.1093/BIB/BBU035

16. Cingolani P, Platts A, Wang LL, et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6(2):80–92. https://doi.org/10.4161/FLY.19695

17. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47(D1):D886–D894. https://doi.org/10.1093/NAR/GKY1016

18. Grotzinger AD, Fuente J de la, Privé F, Nivard MG, Tucker-Drob EM (2023) Pervasive Downward Bias in Estimates of Liability-Scale Heritability in Genome-wide Association Study Meta-analysis: A Simple Solution. Biol Psychiatry 93(1):29–36. https://doi.org/10.1016/j.biopsych.2022.05.029

19. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39(7):906–913. https://doi.org/10.1038/ng2088

20. Willer CJ, Li Y, Abecasis GR (2010) METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26(17):2190–2191. https://doi.org/10.1093/bioinformatics/btq340

21. Spracklen CN, Horikoshi M, Kim YJ, et al (2020) Identification of type 2 diabetes loci in 433,540 East Asian individuals. Nature 582(7811):240–245. https://doi.org/10.1038/s41586-020-2263-3

22. Mahajan A, Taliun D, Thurner M, et al (2018) Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet 50(11):1505–1513. https://doi.org/10.1038/s41588-018-0241-6

23. Haymond M, Anderson B, Barrera P, et al (2007) Treatment options for type 2 diabetes in adolescents and youth: a study of the comparative efficacy of metformin alone or in combination with rosiglitazone or lifestyle intervention in

adolescents with type 2 diabetes. Pediatr Diabetes 8(2):74–87. https://doi.org/10.1111/J.1399-5448.2007.00237.X

24. SEARCH Study Group (2004) SEARCH for Diabetes in Youth: a multicenter study of the prevalence, incidence and classification of diabetes mellitus in youth. Control Clin Trials 25(5):458–471. https://doi.org/10.1016/J.CCT.2004.08.002

25. Srinivasan S, Chen L, Todd J, et al (2021) The First Genome-Wide Association Study for Type 2 Diabetes in Youth: The Progress in Diabetes Genetics in Youth (ProDiGY) Consortium. Diabetes 70(4):996–1005. https://doi.org/10.2337/db20-0443

26. Arellano-Campos O, Gómez-Velasco D v., Bello-Chavolla OY, et al (2019) Development and validation of a predictive model for incident type 2 diabetes in middle-aged Mexican adults: The metabolic syndrome cohort. BMC Endocr Disord 19(1):41. https://doi.org/10.1186/s12902-019-0361-8

27. Fisher-Hoch SP, Rentfro AR, Salinas JJ, et al (2010) Socioeconomic Status and Prevalence of Obesity and Diabetes in a Mexican American Community, Cameron County, Texas, 2004-2007. Prev Chronic Dis 7(3):A53. https://doi.org/10.13016/vtrw-onkt

28. Nair AK, Sutherland JR, Traurig M, et al (2018) Functional and association analysis of an Amerindian-derived population-specific p.(Thr280Met) variant in RBPJL, a component of the PTF1 complex. European Journal of Human Genetics 26:238–246. https://doi.org/10.1038/s41431-017-0062-6

29. Wojcik GL, Graff M, Nishimura KK, et al (2019) Genetic analyses of diverse populations improves discovery for complex traits. Nature 570(7762):514–518. https://doi.org/10.1038/S41586-019-1310-4

30. Lin DY, Tao R, Kalsbeek WD, et al (2014) Genetic association analysis under complex survey sampling: The hispanic community health study/study of latinos. Am J Hum Genet 95(6):675–688. https://doi.org/10.1016/J.AJHG.2014.11.005/ATTACHMENT/7D325123-F24E-4BE3-9C95-83221AC8FD37/MMC1.PDF

31. All of Us Research Program Investigators, Denny J, Rutter JL, et al (2019) The “All of Us” Research Program. New England Journal of Medicine 381(7):668–676. https://doi.org/10.1056/NEJMSR1809937/SUPPL_FILE/NEJMSR1809937_APPENDIX.PDF

32. Vujkovic M, Keaton JM, Lynch JA, et al (2020) 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(7):680–691. https://doi.org/10.1038/s41588-020-0637-y

33. Mägi R, Horikoshi M, Sofer T, et al (2017) Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum Mol Genet 26(18):3639–3650. https://doi.org/10.1093/hmg/ddx280

34. Ahola-Olli A v., Mustelin L, Kalimeri M, et al (2019) Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four

Finnish cohorts. Diabetologia 62(12):2298–2309. https://doi.org/10.1007/s00125-019-05001-w

35. The Wellcome Trust Case Control Consortium, Maller JB, McVean G, et al (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet 44(12):1294. https://doi.org/10.1038/NG.2435

36. McLaren W, Gil L, Hunt SE, et al (2016) The Ensembl Variant Effect Predictor. Genome Biology 2016 17:1 17(1):1–14. https://doi.org/10.1186/S13059-016-0974-4

37. Oscanoa J, Sivapalan L, Gadaleta E, Dayem Ullah AZ, Lemoine NR, Chelala C (2020) SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). Nucleic Acids Res 48(W1):W185–W192. https://doi.org/10.1093/NAR/GKAA420

38. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45(6):580–585. https://doi.org/10.1038/NG.2653

39. Alonso L, Piron A, Morán I, et al (2021) TIGER: The gene expression regulatory variation landscape of human pancreatic islets. Cell Rep 37(2):109807. https://doi.org/10.1016/J.CELREP.2021.109807

40. Asplund O, Storm P, Chandra V, et al (2022) Islet Gene View-a tool to facilitate islet research. Life Sci Alliance 5(12):1–17. https://doi.org/10.26508/lsa.202201376

41. Karczewski KJ, Solomonson M, Chao KR, et al (2021) Systematic single-variant and gene-based association testing of 3,700 phenotypes in 281,850 UK Biobank

exomes. medRxiv 2021.06.19.21259117. https://doi.org/10.1101/2021.06.19.21259117

42. Gagliano Taliun SA, VandeHaar P, Boughton AP, et al (2020) Exploring and visualizing large-scale genetic associations by using PheWeb. Nature Genetics 2020 52:6 52(6):550–552. https://doi.org/10.1038/s41588-020-0622-5

43. Ochoa D, Hercules A, Carmona M, et al (2021) Open Targets Platform: supporting systematic drug–target identification and prioritisation. Nucleic Acids Res 49(D1):D1302–D1310. https://doi.org/10.1093/NAR/GKAA1027

44. Russell MA, Redick SD, Blodgett DM, et al (2019) HLA class II antigen processing and presentation pathway components demonstrated by transcriptome and protein analyses of islet β-cells from donors with type 1 diabetes. Diabetes 68(5):988–1001. https://doi.org/10.2337/DB18-0686/-/DC1

45. Marselli L, Piron A, Suleiman M, et al (2020) Persistent or Transient Human β Cell Dysfunction Induced by Metabolic Stress: Specific Signatures and Shared Gene Expression with Type 2 Diabetes. Cell Rep 33(9):108466. https://doi.org/10.1016/J.CELREP.2020.108466

46. Fadista J, Vikman P, Laakso EO, et al (2014) Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc Natl Acad Sci U S A 111(38):13924–13929. https://doi.org/10.1073/PNAS.1402665111

47. Bone R, Oyebamiji O, Talware S, et al (2020) A Computational Approach for Defining a Signature of β-Cell Golgi Stress in Diabetes. Diabetes 69(11):2364–2376. https://doi.org/10.2337/DB20-0636

48. Colli ML, Ramos-Rodríguez M, Nakayasu ES, et al (2020) An integrated multi-omics approach identifies the landscape of interferon-α-mediated responses of human pancreatic beta cells. Nature Communications 2020 11:1 11(1):1–17. https://doi.org/10.1038/s41467-020-16327-0

49. Gonzalez-Duque S, Azoury ME, Colli ML, et al (2018) Conventional and Neo-antigenic Peptides Presented by β Cells Are Targeted by Circulating Naïve CD8+ T Cells in Type 1 Diabetic and Healthy Donors. Cell Metab 28(6):946-960.e6. https://doi.org/10.1016/J.CMET.2018.07.007

50. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 2014 15:12 15(12):1–21. https://doi.org/10.1186/S13059-014-0550-8

51. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 2017 14:4 14(4):417–419. https://doi.org/10.1038/nmeth.4197

52. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun 10(1):1776. https://doi.org/10.1038/s41467-019-09718-5

53. Ruan Y, Lin Y-F, Feng Y-CA, et al (2022) Improving polygenic prediction in ancestrally diverse populations. Nat Genet 54(5):573–580. https://doi.org/10.1038/S41588-022-01054-7