Title

Multi-ancestry polygenic mechanisms of type 2 diabetes.

Journal

Nature Medicine, 30(4)

Authors

Smith, Kirk
Deutsch, Aaron
McGrail, Carolyn
et al.

Publication Date

2024-04-01

DOI

10.1038/s41591-024-02865-3

Peer reviewed

HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2024 October 01.

Published in final edited form as:

Nat Med. 2024 April ; 30(4): 1065–1074. doi:10.1038/s41591-024-02865-3.

Multi-ancestry Polygenic Mechanisms of Type 2 Diabetes

Kirk Smith[1,2,3,*], **Aaron J. Deutsch**[1,2,3,4,*], **Carolyn McGrail**[5], **Hyunkyung Kim**[1,2,3,6,7], **Sarah Hsu**[3,8], **Alicia Huerta-Chagoya**[3,2,1], **Ravi Mandla**[3,2,1], **Philip H. Schroeder**[3,2,1], **Kenneth E. Westerman**[3,4,8], **Lukasz Szczerbinski**[1,2,3,9,10], **Timothy D. Majarian**[3,8], **Varinderpal Kaur**[1,2,3], **Alice Williamson**[11,12], **Noah Zaitlen**[13,14,15], **Melina Claussnitzer**[1,2,3,4,16], **Jose C. Florez**[1,2,3,4], **Alisa K. Manning**[3,4,8], **Josep M. Mercader**[3,2,1,4], **Kyle J. Gaulton**[17], **Miriam S. Udler**[1,2,3,4]

[1.]Diabetes Unit, Endocrine Division, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

[2.]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

[3.]Programs in Metabolism and Medical & Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[4.]Department of Medicine, Harvard Medical School, Boston, MA, USA

[5.]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA, USA

[6.]Committee on Genetics, Genomics and Systems Biology, University of Chicago, Chicago, IL, USA

[7.]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA

[8.]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Boston, MA, USA

[9.]Department of Endocrinology, Diabetology and Internal Medicine, Medical University of Bialystok, Bialystok, Poland

[10.]Clinical Research Centre, Medical University of Bialystok, Bialystok, Poland

[11.]Precision Healthcare University Research Institute, Queen Mary University of London, London, UK

[12.]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK

Corresponding author: Miriam S. Udler, mudler@mgh.harvard.edu.

[*]contributed equally

Present address: TDM: Vertex Pharmaceuticals, Boston, MA, USA

Author contributions

Conceived and designed the study: KS, AJD, MSU

Conducted analysis: KS, AJD, CM, AH

Curated data: KS, AJD, HK, SH, RM, PHS, KEW, LS, TDM, VK, AW, AKM, JMM

Provided feedback on analysis: NZ, MC, JCF, AKM, JMM, KJG, MSU

Wrote initial draft: KS, AJD, MSU

All co-authors approved the final version of the paper

Competing Interests

The authors declare the following competing interests: TDM currently works for Vertex Pharmaceuticals. KS, MC, JCF, JMM, and MSU are currently part of a collaboration project between Broad Institute and Novo Nordisk. MSU is an unfunded collaborator with Nightingale and AstraZeneca. None of the other authors declare competing interests.

[13.]Department of Neurology, University of California, Los Angeles, Los Angeles, CA, USA

[14.]Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA, USA

[15.]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA

[16.]The Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[17.]Department of Pediatrics, University of California San Diego, San Diego, CA, USA

Abstract

Type 2 diabetes (T2D) is a multifactorial disease with substantial genetic risk, for which the underlying biological mechanisms are not fully understood. We identified multi-ancestry T2D genetic clusters by analyzing genetic data from diverse populations in 37 published T2D genome-wide association studies (GWAS) representing >1.4 million individuals. We implemented soft clustering with 650 T2D-associated genetic variants and 110 T2D-related traits, capturing known and novel T2D clusters with distinct cardiometabolic trait associations across two independent biobanks representing diverse genetic ancestral populations (African, *N*=21,906; Admixed American, *N*=14,410; East Asian, *N*=2,422; European, *N*=90,093; South Asian, *N*=1,262). The twelve genetic clusters were enriched for specific single-cell regulatory regions. Several of the polygenic scores derived from the clusters differed in distribution between ancestry groups, including a significantly higher proportion of lipodystrophy-related polygenic risk in East Asian ancestry. T2D risk was equivalent at a BMI of 30 kg/m^2 in the European subpopulation and 24.2 (22.9–25.5) kg/m^2 in the East Asian subpopulation; after adjusting for cluster-specific genetic risk, the equivalent BMI threshold increased to 28.5 (27.1–30.0) kg/m^2 in the East Asian group. Thus, these multi-ancestry T2D genetic clusters encompass a broader range of biological mechanisms and provide preliminary insights to explain ancestry-associated differences in T2D risk profiles.

Introduction

Type 2 diabetes (T2D) is a complex genetic disease mediated by multiple biological pathways. Ongoing efforts have focused on advancing precision medicine in diabetes by identifying unique clinical trajectories or treatment approaches based on diabetes subtype[1]. Multiple strategies have been applied to identify T2D disease pathways and subtypes, incorporating clinical, biomarker, and/or genomic data[2].

Previously, we used genetic data and implemented a soft clustering approach using Bayesian non-negative matrix factorization (bNMF) to identify five physiologically informed clusters of T2D genetic loci[3]. The bNMF procedure groups T2D loci in distinct, but sometimes overlapping, clusters that influence specific groups of clinical or biochemical phenotypic traits. We found two T2D clusters related to mechanisms of insulin deficiency and three related to insulin resistance. More recently, we developed a high-throughput pipeline to analyze a larger set of genome-wide association studies (GWAS) to increase our power to detect T2D clusters[4]. This approach recapitulated our previous five clusters and identified

five additional clusters. At least two other efforts aimed at identifying T2D genetic clusters also appeared to capture some of these clusters, based on the top traits and loci[5,6].

Here, we expand our analysis pipeline to investigate multi-ancestry cohorts. Previously, we focused on European genetic ancestry groups, due to limited data availability and methodological limitations when analyzing genetic data from diverse populations. We hypothesized that incorporating individuals from diverse populations could potentially explain ancestry-associated differences in T2D risk while avoiding the exacerbation of health disparities[7]. Thus, we now leverage recent multi-ancestry genetic studies[8–10] to investigate T2D genetic clusters in diverse ancestral populations. We recapture our previous T2D genetic clusters and identify three new clusters. We confirm that common pathways contribute to T2D risk across multiple ancestral populations and are distinctively enriched for tissue- and single-cell regulatory regions. Additionally, we describe associations of the genetic clusters with clinical phenotypes. Finally, we analyze ancestry-specific variation in genetic clusters to investigate why individuals in certain populations are more susceptible to T2D at a lower body mass index (BMI).

Results

Multi-ancestry approach yields twelve T2D genetic clusters

We previously developed a high-throughput pipeline to generate T2D clusters using T2D GWAS summary statistics[4]. Here, we expanded the set of input T2D GWAS to include participants with different ancestries (Supplementary Table 1) and updated our pipeline to account for varying allele frequencies across populations (Extended Data Fig. 1). We included 37 T2D GWAS representing over 1.4 million individuals across varied genetic ancestral backgrounds: African/African American (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), South Asian (SAS), or multi-ancestry (Supplementary Table 1). After performing quality control and removing correlated elements, we obtained a final set of 650 variants with independent genome-wide significant T2D associations (Supplementary Table 2) and 110 GWAS of T2D-associated traits (Supplementary Table 3), with which we applied our bNMF clustering algorithm.

We identified a total of twelve T2D multi-ancestry genetic clusters (Fig. 1, Table 1, Supplementary Table 4). Compared to our prior work[4], we identified three novel clusters (Lipodystrophy 2, Cholesterol, and Bilirubin), and we recaptured eight out of ten previously identified clusters, each now including more top-weighted variants (Extended Data Fig. 2). The two remaining clusters from our prior work (SHBG and LpA) collapsed into a single cluster, denoted here as SHBG-LpA.

The novel Lipodystrophy 2 cluster contained genetic determinants of lipid metabolism, liver dysfunction, and insulin resistance. The top-weighted traits included increased hepatic enzymes, increased Homeostatic Model Assessment for Insulin Resistance (HOMA-IR), and decreased insulin sensitivity index. The top-weighted loci included *PNPLA3* and *PPARG*, which regulate the accumulation of fatty acids in liver and adipose tissue[11,12]. Compared to our prior work, the previous single Lipodystrophy cluster appeared to split into two clusters (Extended Data Fig. 2), with traits related to body composition (such as increased visceral

adipose tissue) remaining in the Lipodystrophy 1 cluster and traits related to hepatic function moving to the novel Lipodystrophy 2 cluster.

The novel Cholesterol cluster was associated with decreased LDL levels. The top-weighted locus was *APOE*, which mediates lipid metabolism[13]. Another top locus included rs5744672, located near the *POLK* and *HMGCR* loci. *HMGCR* encodes HMG-CoA reductase, which catalyzes the rate-limiting step of cholesterol synthesis. To support the hypothesis that the observed variation in cholesterol levels is mediated via *HMGCR*, we searched for expression quantitative trait loci (eQTL) in the Genotype-Tissue Expression Project (GTEx). We found a significant association between rs5744672 and *HMGCR* expression levels in skeletal muscle (normalized effect size = 0.14, $P = 2.9 \times 10^{-6}$), and there was strong linkage disequilibrium (LD) between rs5744672 and the top eQTL variant for *HMGCR* (rs3846662, $r^2 = 0.91$).

Finally, the novel Bilirubin cluster was associated with increased bilirubin levels. This cluster only included two variants, both located near the complex *UGT1A* locus, although the two variants were independent ($r^2 = 0.004$). The *UGT1A* locus encodes multiple enzymes in the UDP-glucuronosyltransferase family, which mediate excretion of bilirubin metabolites. We again characterized this genetic locus by searching for eQTLs in GTEx. We found that the top locus in the bilirubin cluster (rs887829) was significantly associated with *UGT1A3* expression in the liver (normalized effect size = 0.46, $P = 3.8 \times 10^{-14}$); once again, there was strong LD between rs887829 and the top eQTL for *UGT1A3* (rs869283, $r^2 = 0.58$).

T2D genetic clusters are shared across ancestry groups

To determine whether the genetic clusters were shared across populations, we repeated our clustering pipeline for each individual ancestry group. Our findings were similar; all groups had at least two clusters related to beta cell dysfunction and at least two related to insulin resistance (Extended Data Fig. 3, Supplementary Tables 5–8). We identified fewer clusters in the AFR and AMR groups, likely because the GWAS sample sizes were smaller, yielding fewer variants for the clustering algorithm. Subsequently, we focused our remaining analyses on the multi-ancestry clusters, since they included all variants from the ancestry-specific analyses and also included the trait GWAS with the largest sample sizes.

T2D genetic clusters capture distinct clinical associations

We tested associations between the multi-ancestry clusters and specific continuous traits or disease outcomes. To accomplish this, we implemented cluster-specific partitioned polygenic scores (pPS) using individual-level data, or if not available, using published GWAS summary statistics (GWAS-pPS) (see Methods; Supplementary Table 9).

To characterize glycemic physiology, we assessed the associations of clusters with glycemic traits (Homeostatic Model Assessment of β-cell function [HOMA-B], HOMA-IR, proinsulin, and corrected insulin response), using GWAS summary statistics included as inputs in the bNMF clustering algorithm (Fig. 2A, Supplementary Table 10). The Beta Cell 1, Beta Cell 2, and Proinsulin clusters were associated with decreased HOMA-B ($\beta = -0.002$ to -0.01, $P < 1 \times 10^{-4}$), suggesting a primary mechanism of insulin deficiency. The

Obesity, Lipodystrophy 1, and Lipodystrophy 2 clusters were associated with increased HOMA-IR (β = 0.005 to 0.007, $P < 1 \times 10^{-8}$), suggesting a primary mechanism of insulin resistance. Two clusters (Liver-Lipid and ALP Negative) were associated with both increased HOMA-IR (β = 0.005 to 0.007, $P < 0.02$) and fasting insulin adjusted for BMI (β = 0.0120 to 0.0128, $P < 9 \times 10^{-5}$), consistent with prior work suggesting a mechanism of insulin resistance; additionally, they were also now associated with decreased HOMA-B (β = −0.005 to −0.006, $P < 4 \times 10^{-3}$), potentially also indicating a more complex mechanism. The Hyper Insulin and Cholesterol clusters were not significantly associated with HOMA-B or HOMA-IR, but were associated with other variables suggestive of a mechanism of insulin resistance (fasting insulin, insulin sensitivity index adjusted for BMI). The final two clusters (Bilirubin and SHBG-LpA) were not clearly associated with insulin deficiency or resistance.

Next, to assess how the T2D clusters inform individual-level clinical differences, we calculated pPS in biobank participants. We performed our primary analysis in the All of Us cohort and replicated our findings in the Mass General Brigham (MGB) Biobank. Here, we present our findings from a meta-analysis of over 100,000 participants (including over 14,000 with T2D) from the two cohorts (Supplementary Table 11).

First, in biobank participants with and without T2D, we validated the relationship between cluster pPS and clinical measurements (Fig. 2B, Supplementary Tables 12–13). For example, we confirmed that the clusters had varied associations with lipid measurements: the Cholesterol pPS was associated with lower LDL (β = −0.07 standard deviations [SD] of LDL per SD of pPS, $P = 3.5 \times 10^{-73}$), the Liver/Lipid pPS with lower triglycerides (β = −0.08, $P = 4.3 \times 10^{-104}$), and the Lipodystrophy 1 pPS with lower HDL (β = −0.04, $P = 3.6 \times 10^{-33}$) and higher triglycerides (β = 0.09, $P = 1.2 \times 10^{-125}$).

We also analyzed measures of waist and hip circumference (only available in All of Us), as well as measures of subcutaneous adipose tissue (SAT) and visceral adipose tissue (VAT) (only available in ~9,000 MGB Biobank participants). The Lipodystrophy 1 pPS was associated with increased VAT/SAT ratio (β = 0.06, $P = 3.1 \times 10^{-12}$) and increased waist-hip ratio (β = 0.05, $P = 6.1 \times 10^{-39}$). In contrast, the Lipodystrophy 2 pPS had a weaker association with VAT/SAT and was not associated with waist-hip ratio. Findings were similar after adjusting for BMI as a covariate. We observed sex-specific effects for the Lipodystrophy 1 cluster; for example, in females the VAT/SAT ratio association was driven by an increased VAT ($\beta_{female} = 0.05$, $P = 3.5 \times 10^{-4}$) with no significant effect on SAT ($\beta_{female} = -0.003$, $P = 0.86$), whereas in males it was driven by a decreased SAT level ($\beta_{male} = -0.07$, $P = 8.7 \times 10^{-6}$) with no significant effect on VAT ($\beta_{male} = 0.01$, $P = 0.35$; Extended Data Fig. 4, Supplementary Table 13).

Second, we tested the association between pPS and cardiometabolic phenotypes (chronic kidney disease [CKD], hypertension [HTN], coronary artery disease [CAD], non-alcoholic fatty liver disease [NAFLD], diabetic retinopathy, and diabetic neuropathy) (Fig. 2C, Supplementary Tables 14–15). The Lipodystrophy 2 cluster was associated with increased NAFLD risk (OR = 1.24, $P = 1.0 \times 10^{-44}$), whereas the Liver-Lipid cluster was associated with decreased NAFLD risk (OR = 0.95, $P = 3.0 \times 10^{-4}$; Fig. 2C, Supplementary Table 15). The Cholesterol pPS was nominally associated with decreased CAD risk (OR = 0.97,

$P = 2.3 \times 10^{-3}$) in the individual-level data, and this negative association reached greater significance in the GWAS-pPS analysis (OR = 0.97, $P = 2.8 \times 10^{-27}$; Supplementary Table 10). Among individuals with T2D, we also analyzed the risk of two microvascular complications, diabetic retinopathy and diabetic neuropathy. No cluster was significantly associated with microvascular complications at the Bonferroni-adjusted threshold.

Third, we analyzed the extremes of each cluster. Across both individual-level cohorts, we found that 35% of all individuals with T2D had a pPS in the top decile of exactly one cluster, similar to the expected proportion under a binomial distribution with twelve independent clusters. Furthermore, individuals who fell in the top decile of a single cluster pPS had unique phenotypic differences (Supplementary Table 16). For example, among individuals with T2D, those in the top decile of the Cholesterol cluster pPS had a mean medication-adjusted LDL of 124.3 mg/dL, compared to 134.1 mg/dL for all other individuals ($P = 3.1 \times 10^{-3}$). Thus, for a substantial portion of the population, a single genetic cluster contributed more to T2D genetic risk than any other cluster did.

T2D cluster loci are enriched in relevant tissues

To further explore biological mechanisms of the multi-ancestry T2D genetic clusters, we assessed for epigenomic evidence of transcriptional activity across a wide array of human tissues. We examined single-cell epigenomic data in CATLAS[14], as well as tissue-specific data generated by the Roadmap Epigenomics Consortium[15]. We found cluster-specific enrichment of epigenomic annotations in biologically relevant tissues (Fig. 3, Supplementary Table 17). The Beta Cell 1 cluster was enriched for epigenomic annotations in a diverse range of cell types, including pancreatic beta cells, while the Beta Cell 2 cluster was specifically characterized by pancreatic islet cell enrichment in alpha, beta, gamma, and delta cells (False Discovery Rate [FDR] < 0.01) (Fig. 3A). The associations between the two Beta Cell clusters and pancreatic islets were also captured in the Roadmap analysis (Fig. 3B). Meanwhile, the Liver-Lipid cluster was enriched in fetal hepatoblasts (FDR < 0.01), while the Lipodystrophy 1 and 2 clusters were enriched in adipose tissue (FDR < 0.01). These findings confirm that the genetic clusters capture variants with distinct effects in specific tissue types, and these effects relate to suspected disease mechanisms.

Distribution of T2D genetic clusters differs by ancestry

Next, we assessed whether the multi-ancestry T2D genetic clusters had varying contributions to overall T2D genetic risk in different populations. We applied principal component analysis to classify individuals by genetic ancestry in both All of Us and MGB Biobank[16]. After confirming that the distribution of T2D genetic clusters was similar in both cohorts, we performed a meta-analysis of both biobanks. For all individuals, we calculated pPS using the multi-ancestry T2D genetic clusters.

We found that the cluster-specific distribution of T2D genetic risk differed according to genetic ancestry. For example, the median Beta Cell 1 pPS was highest in the AFR ancestry group, whereas the median Lipodystrophy 1 and 2 pPS were highest in the EAS ancestry group (Fig. 4A; Extended Data Fig. 5). Furthermore, within each ancestry group, we found varied proportions of T2D genetic risk attributable to each cluster (Extended Data Fig. 6).

For example, 12.7% of the total T2D genetic risk across all clusters was present in the Lipodystrophy 1 cluster for the EAS ancestry group, which was significantly higher than the other groups (AFR: 8.1%, AMR: 9.6%, EUR: 9.0%, SAS: 9.6%; $P < 10^{-300}$, one-way ANOVA).

Lipodystrophy 1 and 2 clusters modulate T2D-BMI relationship

We then investigated whether ancestry-specific variation in T2D genetic risk resulted in phenotypic differences between ancestry groups. Given that individuals with East Asian genetic ancestry develop T2D at lower BMI levels[17], which may be due to a tendency for metabolically unhealthy or "lipodystrophic" fat distribution[18], we hypothesized that this phenomenon could be partly explained by genetics. We focused on the multi-ancestry Lipodystrophy 1 and 2 clusters, which were associated with decreased BMI but increased VAT/SAT ratio and T2D risk (Fig. 2, Supplementary Tables 12, 13). After classifying individuals by genetic ancestry, we first analyzed the relationship between BMI and T2D risk in both All of Us and MGB Biobank separately; then, we performed a meta-analysis of both cohorts together.

Within each ancestry group, we calculated BMI thresholds with equivalent T2D risk and found they were largely consistent with prior reports[19–22]. For example, at a BMI of 30 kg/m^2 (typically used to define obesity), the risk of T2D within the EUR ancestry group was 11.7%. However, to achieve the same risk of T2D, the corresponding BMI cutoff varied in other groups: AFR, 25.2 kg/m^2 (95% confidence interval 24.7–25.7); AMR, 23.7 kg/m^2 (23.1–24.3); EAS, 24.2 kg/m^2 (22.9–25.5); and SAS, 20.8 kg/m^2 (19.4–22.2) (Fig. 4B). After adjusting for the Lipodystrophy 1 and 2 pPS, to achieve the same risk of T2D as an individual with a BMI of 30 kg/m^2 in the EUR ancestry group, the corresponding BMI cutoffs were: AFR, 25.1 kg/m^2 (24.6–25.6); AMR, 25.0 kg/m^2 (24.4–25.5); EAS, 28.5 kg/m^2 (27.1–30.0); and SAS, 22.0 kg/m^2 (20.6–23.4) (Fig. 4C). Thus, by accounting for cluster-specific pPS, the difference in T2D risk-equivalent BMI thresholds between the EAS and EUR ancestry groups decreased by approximately 4 kg/m^2.

We confirmed that the difference between T2D risk-equivalent BMI thresholds was primarily driven by the Lipodystrophy 1 and 2 clusters, as our findings were similar when we adjusted for pPS from all 12 clusters simultaneously. Furthermore, our findings were similar after restricting the population to individuals with similar Lipodystrophy 1 pPS (Extended Data Fig. 7A–B), and the relationship between Lipodystrophy 1 pPS and T2D risk remained strong even within the EUR ancestry group (Extended Data Fig. 7C). In addition, we did not find evidence for artifacts in ancestry-specific variants that could explain significant variation in cluster pPS across ancestry groups (Extended Data Fig. 8).

Next, we investigated the mechanism by which the Lipodystrophy 1 and 2 pPS impacted the BMI threshold for T2D risk. In the subset of ~9,000 MGB Biobank participants with available VAT and SAT measurements, we investigated the relationship between T2D risk, VAT/SAT ratio, and triglyceride levels. We found that the Lipodystrophy 1 and 2 pPS partly explained the relationship between VAT/SAT ratio and T2D risk across subpopulations, as well as the relationship between BMI and triglyceride levels (Extended Data Fig. 9).

Potential clinical application of T2D genetic clusters

In current clinical practice, providers frequently use ancestry-based BMI ranges when counseling patients about T2D prevention and treatment; for example, obesity has been defined by a BMI > 30 kg/m^2 in White individuals and > 27.5 kg/m^2 in certain Asian populations[23]. While applying subpopulation-based normal ranges is intended to provide individualized care, such practices remain controversial due to both accuracy-related and ethical concerns[24]. To illustrate a potential clinical application of the T2D genetic clusters, we used the All of Us participants to develop individualized BMI thresholds that do not rely on a person's race or genetic ancestry (see Methods). When we uniformly selected all individuals with a BMI of 30±10%, the T2D risk (adjusted for age and sex) was 24.6% in the EAS group compared to 11.5% in the EUR group (P = 0.013). If instead we selected all individuals with a BMI equivalent to their population-level risk threshold ±10%, the adjusted T2D risk was 14.6% for the EAS group and 11.1% for the EUR group (P = 0.38). Finally, when we selected all individuals with a BMI equivalent to the individual-level risk threshold ±10%, the adjusted T2D risk was 9.5% for the EAS group and 10.3% for the EUR group (P = 0.82). Thus, we demonstrated that cluster pPS can be applied in a clinical setting to determine an individual's target BMI level, regardless of their genetic ancestry.

Discussion

In this study, we assembled a diverse set of GWAS to analyze 650 independent T2D-associated variants and 110 relevant traits, and we identified twelve potential T2D genetic clusters. By including genetic variants from multiple ancestry groups, we validated and expanded on our prior T2D clustering work, which focused on European populations and included only 323 variants[3,4]. We confirmed the existence of eight previously identified T2D genetic clusters, and we found that the previously defined SHBG and Lipoprotein A clusters[4] merged into a single cluster. A new cluster, which we denoted as Lipodystrophy 2, split from the previous Lipodystrophy cluster, and we identified novel clusters associated with cholesterol and bilirubin. These clusters were significantly enriched in regulatory genomic regions in both bulk tissue and single cell epigenomic datasets, implicating tissues and cells consistent with predicted disease mechanisms. Additionally, we characterized the clinical features of the genetic clusters, notably including sex-specific analyses as well as association results with NAFLD.

One of the new clusters, the Cholesterol cluster, captures the complex relationship between T2D, CAD, and LDL cholesterol. While multiple genetic loci confer increased risk for both T2D and CAD[25,26], this cluster of T2D risk alleles was associated with reduced LDL levels and decreased CAD risk. The cluster included an eQTL for *HMGCR*, the target of statin medications, which lower LDL cholesterol and CAD risk, but are also known to increase T2D risk[27]. Hence, this cluster supports the notion that a subset of individuals have a genetic mechanism causing divergent CAD and T2D risk[28].

Aside from *HMGCR*, our multi-ancestry T2D genetic clusters confirmed the role of multiple genetic variants encoding proteins that serve as drug targets. For example, the Lipodystrophy 1 and 2 clusters included rs17036160 near *PPARG*, the target of thiazolidinediones, which

promote insulin sensitivity. Furthermore, the Hyper Insulin cluster included rs10305420 near *GLP1R*, the target of GLP1 receptor agonists, which potentiate insulin secretion.

The biological significance of the novel Bilirubin cluster is unclear. The top locus included an eQTL for *UGT1A3*, which mediates bilirubin metabolism. Although the cluster suggests a positive correlation between bilirubin levels and T2D risk, previous epidemiologic studies have demonstrated a negative association[29]. In addition, bile acid sequestrants may be used for treatment of T2D[30]; however, the link between bile acid sequestrants and serum bilirubin levels is uncertain.

The Lipodystrophy cluster from our previous work split into two clusters, Lipodystrophy 1 and 2 (Extended Data Fig. 2). The pPS for both clusters were significantly associated with the classic "lipodystrophy-like" phenotype of increased triglycerides, insulin resistance, and VAT/SAT ratio, but decreased HDL and BMI (Supplementary Table 10, 12, 13). Both cluster pPS were also significantly associated with increased risk of HTN, CAD, CKD, and NAFLD (Supplementary Table 15) and were epigenetically enriched in adipocytes (Fig. 3B). However, the Lipodystrophy 1 cluster was driven more by traits and loci related to body composition (e.g. gluteofemoral and visceral fat measures; *COBLL1*[31], *FAM13A*[32]), whereas Lipodystrophy 2 was driven by liver-related loci and phenotypes (e.g. ALT, AST; *PNPLA3*[11], ERLIN1[33]). Indeed, only the Lipodystrophy 1 pPS (not the Lipodystrophy 2 pPS) was associated with increased VAT and increased waist-hip ratio (Extended Data Fig. 4, Supplementary Table 13).

We also demonstrated how the Lipodystrophy clusters can help explain the heterogeneity of T2D across populations. Individuals from various self-identified non-White populations are more susceptible to T2D at lower BMIs, compared to self-identified White individuals[17,21], and many authors have suggested that population-specific BMI thresholds should be used to define obesity[19,20,22]. For example, certain guidelines suggest that individuals should be screened for diabetes if their BMI is ≥25 kg/m^2, or if they identify as Asian and their BMI is ≥23 kg/m^2 [34]. Some studies have suggested that population-level differences in adipose tissue distribution may explain the varied relationship between BMI and T2D risk[18]. In practice, however, applying varied guidelines according to race and ethnicity raises ethical concerns about perpetuating structural racism, and accuracy may be limited for individuals who identify with more than one race[35].

In this study, after classifying individuals by genetic ancestry, we confirmed prior observations[17,21] that individuals in the EUR ancestry group had the lowest risk of T2D at all BMI strata. Furthermore, we demonstrated that variation in the BMI-T2D relationship is at least partially explained by variation in the Lipodystrophy 1 and 2 genetic clusters (Fig. 4A). After adjusting for the Lipodystrophy 1 and 2 pPS, the difference in T2D risk-equivalent BMI thresholds between the EAS and EUR ancestry groups was reduced by about 4 kg/m^2 (Fig. 4B, C). These findings represent a potential step toward developing individualized, genetically informed BMI recommendations. For instance, as genetic information becomes more widely available, decision support tools may incorporate an individual's Lipodystrophy 1 and 2 pPS to help clinicians recommend an individualized target BMI for preventing or treating diabetes.

Notably, however, the Lipodystrophy 1 and 2 pPS were not markedly elevated in the other groups (AFR, AMR, and SAS) compared to the EUR subpopulation, so adjusting for these scores did not substantially affect T2D risk estimates. In particular, across the BMI spectrum, T2D risk was strikingly elevated in the SAS group compared to other subgroups, consistent with prior observations that individuals with South Asian ancestry have elevated T2D risk despite lower BMI[36]. Hypothesized mechanisms for this observation include decreased insulin secretion, decreased lean muscle mass, or ectopic fat deposition in lean muscle tissue[37]. It is possible that the multi-ancestry genetic clusters did not adequately account for genetic risk in the SAS group, as the input GWAS had relatively low representation of individuals with SAS ancestry. Thus, further work is needed to investigate genetic and non-genetic factors affecting the BMI-T2D relationship in SAS ancestry and other populations.

In parallel to our current study, the T2D Global Genomics Initiative (T2DGGI) also investigated T2D genetic clusters in multi-ancestry GWAS[6]. The T2DGGI clustering approach used genetic variants from a single, large multi-ancestry GWAS. In contrast, the study presented here used genetic variants from multiple ancestry-specific and multi-ancestry GWAS. Furthermore, the T2DGGI approach used a hard clustering method, compared to the soft clustering method presented here. Nevertheless, both studies demonstrated high degrees of similarity in several clusters, including the Beta Cell, Obesity and Lipodystrophy 1 clusters (Extended Data Fig. 10). Each study also included certain clusters that were not captured by the other study, and further downstream analyses will be necessary to determine the relative utility of both results.

Although our results demonstrate the importance of analyzing diverse ancestral populations, our findings were limited by the availability of genetic data. In particular, the largest available GWAS primarily include individuals with European genetic ancestry, underscoring the necessity of broadening genetic research across the globe. Likewise, the majority of individuals in our biobank analyses had European genetic ancestry, limiting our ability to make inferences in other populations. Finally, while our analysis linked genetic loci to disease mechanisms, experimental models are needed for functional validation, and the cluster pPS are not yet able to definitively assign an individual to a specific genetic subtype.

Overall, we demonstrated that similar patterns of T2D genetic clusters occur across multiple populations. Using a multi-ancestry approach, we identified novel clusters that help to elucidate the complex relationship between BMI, T2D, CAD, and NAFLD. We also demonstrated how genetic variation across ancestry groups can cause differences in body fat composition, thereby altering T2D risk. To advance the care of patients with diabetes, current and future studies may focus on precision medicine strategies to target specific biological mechanisms highlighted by the T2D genetic clusters.

Methods

Our research complied with all relevant ethical regulations. The research protocol was approved by the Mass General Brigham Institutional Review Board.

Pipeline for input variant–trait association matrix for clustering

The pipeline's data preprocessing steps are detailed in the flowchart shown in Extended Data Fig. 1. For the multi-ancestry clusters, GWAS-significant ($P < 5 \times 10^{-8}$) variants were extracted from a diverse set of T2D GWAS (Supplementary Table 1), including studies performed in European, East Asian, African, Admixed American, South Asian and mixed cohorts. After removing indels and variants found in the major histocompatibility complex (MHC) region, variants underwent five independent iterations of LD-pruning (LD $r^2 < 0.05$, MAF < 0.001), one for each population's reference panel. Variants were only retained if found to be independent in all five populations. If any of the pruned variants had high-missingness across the trait GWAS, was multi-allelic or was ambiguous (A/T, C/G), then it was replaced with a high-LD ($r^2 > 0.8$) proxy variant. As a final check, the variants were queried in the largest multi-ancestry T2D GWAS and were removed if they had $P > 0.05$ or if there were discrepancies in the noted risk alleles. The final set of 650 T2D-associated variants is shown in Supplementary Table 2.

For the traits included in the clustering, we compiled an extensive list of 165 continuous phenotypes GWAS and allowed the pipeline to determine which were relevant to the T2D variants (Supplementary Table 3). We prioritized sex-specific and multi-ancestry GWAS; however, if those were not available for a specific trait, then European-based GWAS were used. Traits were filtered out if their median sample size was below 5,000 or if their minimum P value for the final variant set was not Bonferroni-significant ($P_{min} > 0.05/650$ variants). Finally, we removed highly correlated traits ($R > 0.80$), prioritizing traits by their maximum variant-trait association (Supplementary Table 4C). With this final set of variants and traits (650 variants x 110 traits), we generated a matrix of standardized and scaled z-scores, which had been aligned to the T2D risk-increasing alleles. To fill in any remaining missing variant-trait associations in this final matrix, we used z-scores from proxies (LD $r^2 > 0.5$) where possible, and otherwise assigned the trait's median value.

The ancestry-specific clusters were generated using the same general steps; however, the input T2D GWAS were limited to studies where the cohort matched the population of interest. For the African and Admixed American clusters, the T2D P value threshold was lowered to $P < 5 \times 10^{-6}$, to account for the less powerful GWAS. The variants were pruned in a single iteration, using the appropriate reference panel for each population. For the traits, ancestry-specific GWAS were prioritized, followed by multi-ancestry and European-based summary statistics (Supplementary Table 3).

Statistical comparison of cluster overlap

To compare different versions of the T2D genetic clusters, we focused on the cluster weights assigned to the T2D-related traits. For each pair of clusters, we calculated the Pearson correlation coefficient (R) between each set of trait cluster weights. We compared the multi-ancestry clusters generated in this study to the T2D genetic clusters identified in our prior studies[3,4]. We also compared the ancestry-specific clusters to the multi-ancestry clusters and to the T2D genetic clusters from our prior study[3].

To compare the multi-ancestry clusters and the T2D clusters generated by the T2DGGI study[6], we focused on the genetic loci included in each cluster, since the clustering method used by the T2DGGI study did not assign traits to specific clusters or generate cluster weights. First, we matched genetic variants included in the T2DGGI clusters to a corresponding high-LD variant ($r^2 > 0.5$) from our multi-ancestry clusters. By doing so, we were able to transfer our variant weights to the T2DGGI clusters. We then assessed the correlation between genetic variant weights across the T2DGGI and multi-ancestry clusters using the Wilcoxon rank-sum test.

Calculation of partitioned polygenic scores

We created partitioned polygenic scores (pPS) by calculating a weighted sum of the genetic variants in each cluster. We used individual-level data when possible; when unavailable, we used GWAS summary statistics. To calculate GWAS-partitioned pPS, we extracted the genetic variants from summary statistics of GWAS for specific traits. We combined the variants using inverse-variance weighted fixed effects meta-analysis, weighting each variant according to its GWAS effect size. We chose GWAS for several key glycemic traits (such as disposition index, proinsulin, and fasting insulin) as well as for measures of adipose tissue distribution or cardiometabolic outcomes. In addition, we calculated individual-level pPS using genotype data from two external biobanks: the All of Us research program[38] and the Mass General Brigham (MGB) Biobank[39]. For individual-level pPS, we weighted the genetic variants according to the cluster weights generated by the bNMF algorithm. We only included those variants with a weight above 0.7802, a threshold that was calculated to maximize the signal-to-noise ratio, as described in Kim *et al.*[4] For these analyses, we also calculated a total genetic risk score (GRS) using the effect sizes of all 650 T2D variants.

Biobank Analyses

For individual-level data, we performed a meta-analysis of two datasets. Each dataset was independent of the GWAS cohorts used to generate the clusters. Informed consent was obtained from all participants in both datasets. We complied with all relevant ethical regulations when analyzing genetic data from human research participants. Individuals were not compensated for participation in this study.

All of Us: Analysis of the All of Us cohort[38] was approved by an institutional Data Use and Registration Agreement between MGB and the All of Us Research Program (study protocol 2020P002213). We used the All of Us Controlled Tier Dataset v6. Full details on the demographic distribution of the dataset are provided in Supplementary Table 11. Individuals were classified as having type 2 diabetes if they were identified by an algorithm from Northwestern University as part of the Electronic Medical Records and Genomics (eMERGE) consortium[40], or if they self-identified as having type 2 diabetes on the All of Us participant survey[41]. The eMERGE algorithm classifies individuals based on diagnosis codes, medication prescriptions, and laboratory values. All individuals who were not classified as having type 2 diabetes were labeled as controls; however, individuals were excluded from the control group if they were less than 30 years old, or if they ever had a hemoglobin A1c greater than or equal to 6.5%. Other phenotypes were defined as described in Supplementary Table 14. Sex was derived from medical records.

MGB Biobank: We used clinical and genomic data from the MGB Biobank[39], which was current as of October 2022. Analysis of the MGB Biobank was approved by the MGB IRB (study protocol 2016P001018). Full details on the demographic distribution of the dataset are provided in Supplementary Table 11. Type 2 diabetes was defined using a phenotype algorithm developed by the MGB Biobank[39], with a set positive predictive value of 0.95. Once again, individuals were excluded from the control group if they were less than 30 years old, or if they ever had a hemoglobin A1c greater than or equal to 6.5%. Other phenotypes were defined as described in Supplementary Table 14. Sex was derived from medical records.

Statistics and reproducibility

Detailed methods related to the clustering algorithm are provided under “Pipeline for input variant-trait association matrix for clustering”. In particular, no statistical method was used to predetermine sample size. We assembled the largest available GWAS for T2D and related traits. Variants and traits were excluded according to specified criteria (e.g. high missingness, high correlation with other traits) that would diminish the effectiveness of the clustering algorithm.

Detailed methods related to individual-level analyses are provided under “Biobank Analyses”. No statistical method was used to predetermine sample size. In both All of Us and MGB Biobank, we started with all participants who consented to the study and who had genomic data available. Individuals were excluded from the control group if they were less than 30 years old, as they could still develop T2D later in life. Additionally, individuals were excluded from the control group if they ever had a hemoglobin A1c greater than or equal to 6.5%, as these individuals may have been misclassified and likely had diabetes. All analyses were retrospective, and individuals were not prospectively allocated to experimental groups. Therefore, the experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

All statistical analyses were performed using the following software packages: R version 4, Python version 3.

Assessment of transcriptional activity

We analyzed transcriptional activity of genetic loci using two databases of epigenomic information. For our primary analysis, we used CATLAS, a resource that maps regions of accessible chromatin across the human genome at single-cell resolution[14]. CATLAS uses ATAC-Seq to identify over 1 million candidate cis-regulatory elements across more than 200 distinct human cell types (both adult and fetal cells). As a secondary analysis, we used information from the Roadmap Epigenomics Consortium, which includes maps of regulatory elements for over 100 tissue types at the bulk tissue level[15]. To assess for enrichment of epigenomic annotations, we first defined 99% credible sets for each locus. To do this, we calculated approximate Bayes factors (aBFs) for all variants within a 500 kb window that had $r^2 \geq 0.1$ with the index variant, as described previously[42]. We calculated a posterior probability for each variant by dividing the aBF by the sum of all aBFs in the credible set. Next, within each cluster, we overlapped credible set variants with cell

type genomic annotations and calculated the cumulative posterior probability (cPPA) for each annotation. We used a permutation test to assess the significance for annotations in each cluster. For each cluster, we permuted locus and cell type annotations and recalculated the cPPA based on shuffled labels. After performing 10,000 permutations, we compared the observed cPPA to the permuted background using a one-tailed test to determine the significance of each annotation. We corrected for multiple tests and defined statistically significant enrichment at q value thresholds of 0.1 and 0.001.

Determination of genetic ancestry

We performed principal component analysis to uncover population stratification in each dataset (MGB Biobank and All of Us). Measurements that capture genetic similarity (such as principal components) are preferred when performing genomics research. However, due to privacy restrictions, we were unable to combine genomic data from both datasets to generate a single set of principal components. Therefore, we used principal component data to apply population descriptor labels at the level of continental ancestry, acknowledging that these labels are imprecise. We used a random forest classifier model to assign participants in each biobank to one of six continental ancestry groups (African, Admixed American, East Asian, European, Middle Eastern, or South Asian), following the method of the Pan-UK Biobank[16]. For any given individual, if the probability of each ancestry group was less than 50%, then the individual's genetic ancestry was left as "unclassified". The total number of individuals in each genetic ancestry group is listed in Supplementary Table 11. We excluded any population with fewer than 500 individuals in a given dataset; therefore, the Middle Eastern ancestry group was excluded from downstream analyses.

Individual-level cluster associations with clinical phenotypes

After generating individual-level pPS, we analyzed the association of the pPS with various clinical phenotypes, using linear regression (for continuous outcomes) or logistic regression (for binary outcomes). We analyzed all associations in a meta-analysis of both biobanks (MGB Biobank and All of Us), using a random effects model. Each regression model was adjusted for the following covariates: age, sex, and genetically inferred ancestry. Certain regression models were also adjusted for type 2 diabetes status and/or BMI, as noted. A subset of analyses was performed separately for female or male participants only; these analyses included age and genetically inferred ancestry as covariates. Clinical measurements that were not normally distributed were log-transformed to obtain a normal distribution[43]; following previous studies, these measurements included BMI and triglycerides[4,44].

For validation tests that confirmed known associations between cluster pPS and variables used in the clustering algorithm, we did not use multiple test correction to denote statistical significance. Of note, clinical phenotypes were not directly used in the clustering algorithm, but several traits that were included in the clustering (i.e. glucose, hemoglobin A1c, creatinine, cystatin C, systolic blood pressure, and diastolic blood pressure) can define T2D, CKD, and HTN. In contrast, tests with the remaining phenotypes (CAD, NAFLD, diabetic retinopathy, and diabetic neuropathy) revealed associations with cluster pPS. For these discovery tests, we defined statistical significance using a Bonferroni-adjusted threshold of $0.05/(K \times N)$, where K represents the number of clusters tested and N represents the number

of phenotypes tested. For individual-level testing, we excluded any binary outcome in which fewer than 500 participants met the outcome in either biobank; therefore, ischemic stroke was excluded from downstream analyses.

For patients taking lipid-lowering medications, we adjusted lipid levels for medication use as described previously[45]. In particular, total cholesterol was divided by 0.8, LDL by 0.7, and triglycerides by 0.85. Due to the low frequency of individuals taking non-statin lipid-lowering medications (e.g. bile acid sequestrants), we did not adjust for these medications. In addition, we did not adjust HDL levels for medication use due to the lack of a clear quantitative relationship, although statins are generally felt to cause a modest increase in HDL levels.

Calculation of ancestry-specific BMI cutoff values

We determined ancestry-specific BMI cutoff values with equivalent risk of type 2 diabetes as described previously[17], except we used log transformation of BMI rather than fractional polynomials, following standard statistical practices[43]. For the outcome measure, we used the probability of type 2 diabetes generated from a logistic regression model, rather than type 2 diabetes incidence, as we were unable to reliably ascertain new diagnoses of type 2 diabetes in the biobank cohorts. We fitted a logistic regression model of type 2 diabetes status versus log(BMI), adjusted for age, sex, and genetic ancestry group. We determined the predicted probability of type 2 diabetes for an individual with European genetic ancestry and a BMI of 30. For each ancestry group, we calculated the BMI that would yield the same predicted probability of type 2 diabetes. Then, we repeated this process after adjusting the logistic regression model for the specified cluster pPS values. All tests were performed in a meta-analysis of MGB Biobank and All of Us, using a random effects model.

Analysis of body composition metrics

For participants in MGB Biobank, we used image-based body composition metrics derived from a machine learning algorithm[46,47]. This algorithm quantifies the cross-sectional areas of muscle, subcutaneous adipose tissue (SAT), and visceral adipose tissue (VAT), as measured in abdominal computed tomography (CT) imaging at the level of the L3 vertebral body. For participants in All of Us, we used measurements of waist and hip circumference, which were measured for most participants at the time of enrollment.

Development of genetically-informed BMI thresholds

To identify personalized BMI thresholds, we analyzed individual-level data in the All of Us cohort. First, we randomly divided the All of Us participants into a training cohort (70%) and a validation cohort (30%). In the training cohort, we used a regression model to capture T2D risk, adjusting for age, sex, BMI, Lipodystrophy 1 pPS, and Lipodystrophy 2 pPS. Then, in the validation cohort, we calculated the BMI threshold that conferred equivalent T2D risk for each ancestry group as well as for each individual.

Data Availability

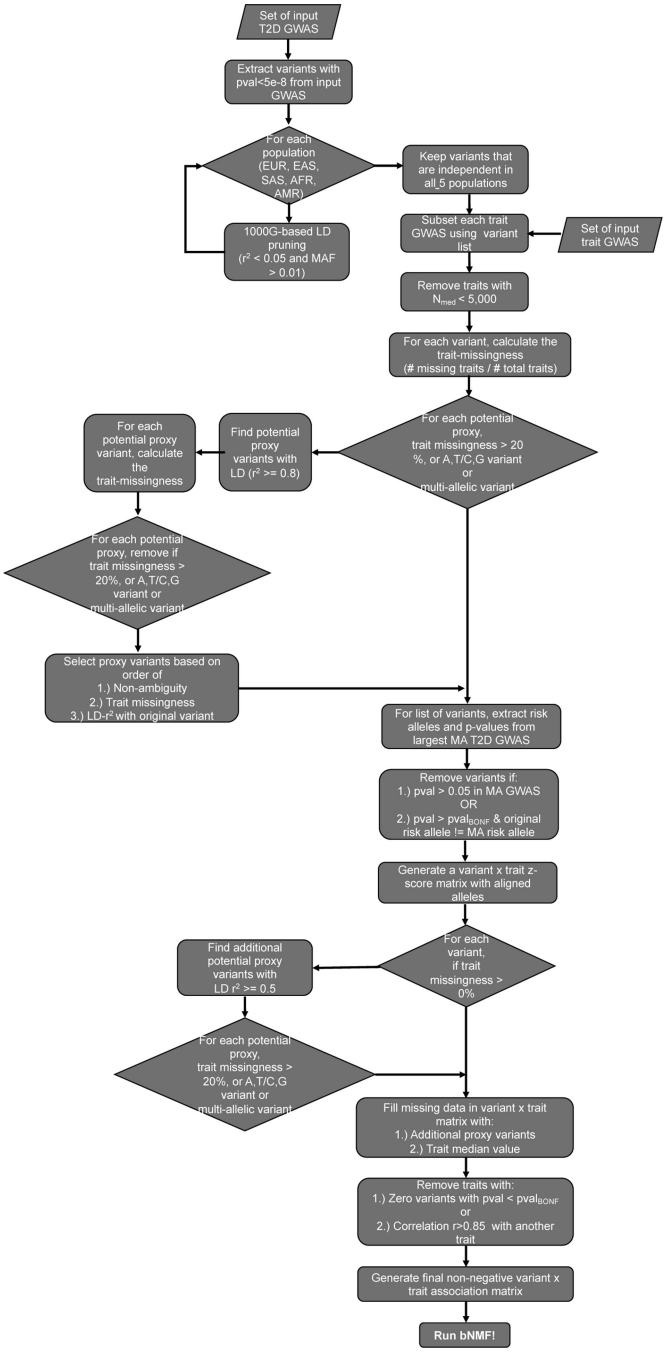
All referenced GWAS summary statistics are publicly available and are cited in Supplementary Tables 1, 3 and 9. Researchers can apply to access individual-level data

in the All of Us program (researchallofus.org). Individual-level data in the Mass General Brigham biobank are only available with approval from the Mass General Brigham Institutional Review Board. Databases of epigenomic activity are available online for CATLAS (https://catlas.org) and Roadmap (https://egg2.wustl.edu/roadmap/).

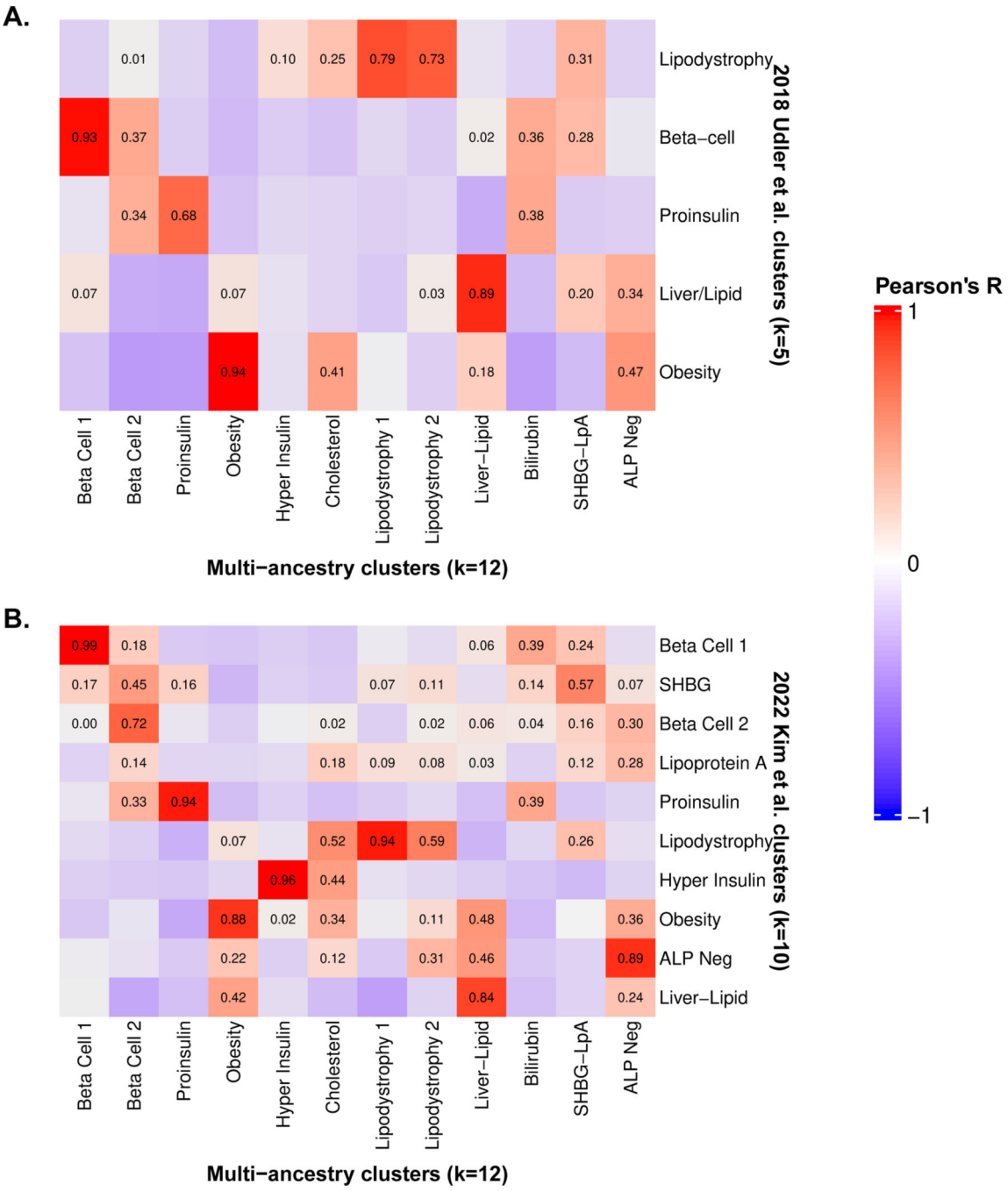
Code Availability

Code for variant pre-processing, bNMF clustering, and basic visualizations is available at https://github.com/gwas-partitioning/bnmf-clustering.

Extended Data

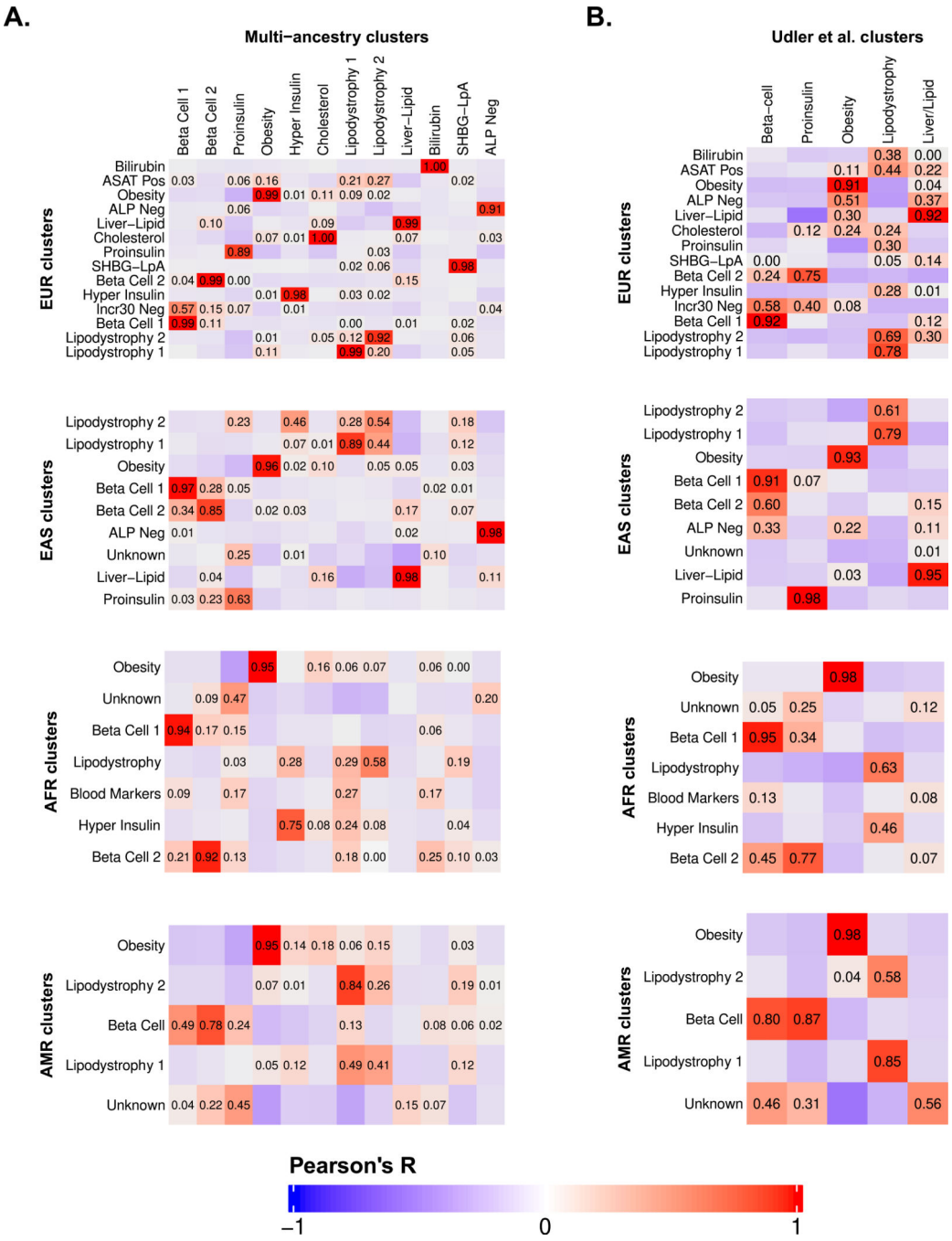


Extended Data Fig. 1.
Overview of high-throughput bNMF pipeline for multiancestry (MA) clusters.



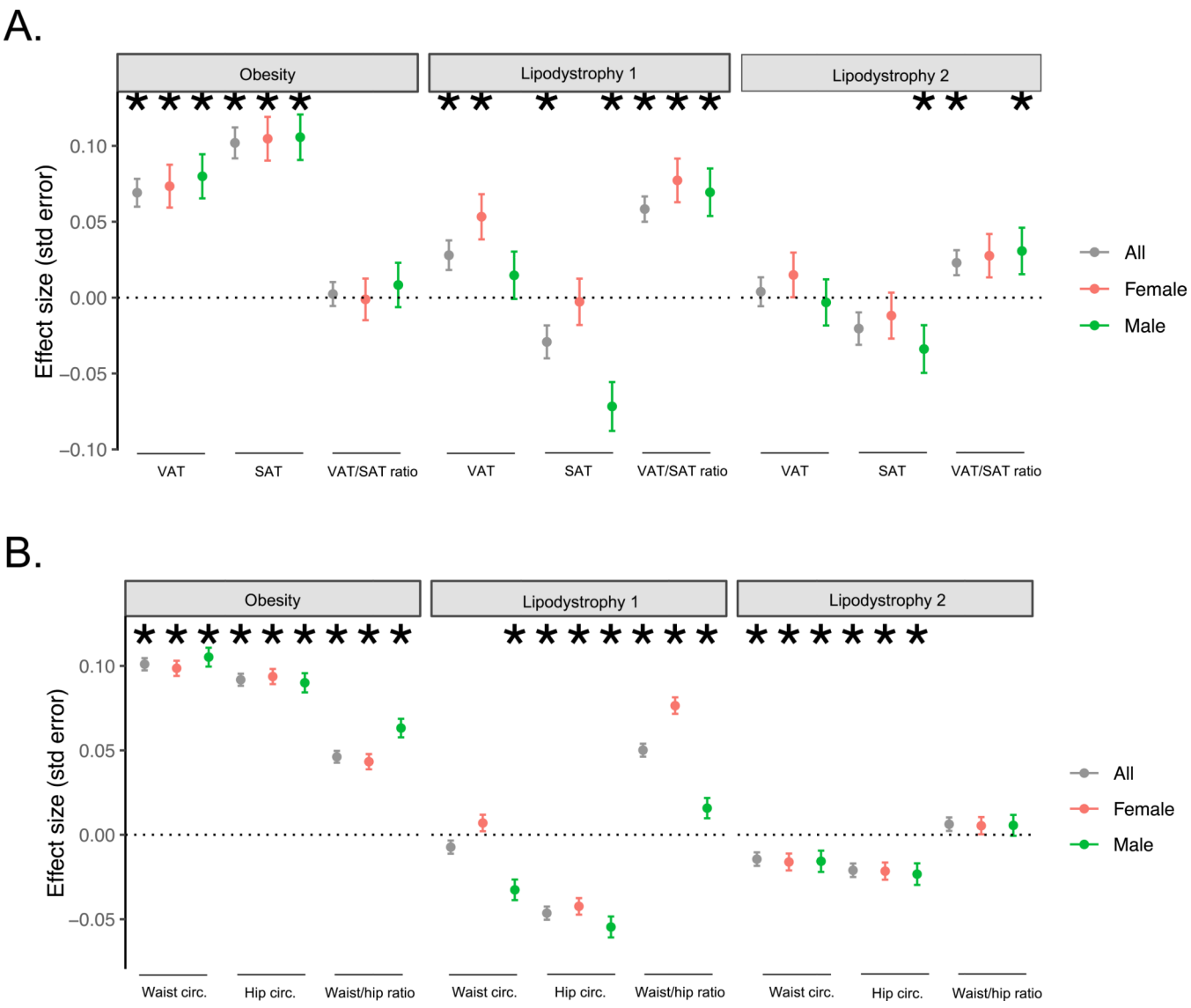
Extended Data Fig. 2.

The multi-ancestry clusters recapture several key pathways that were identified in our previous papers.



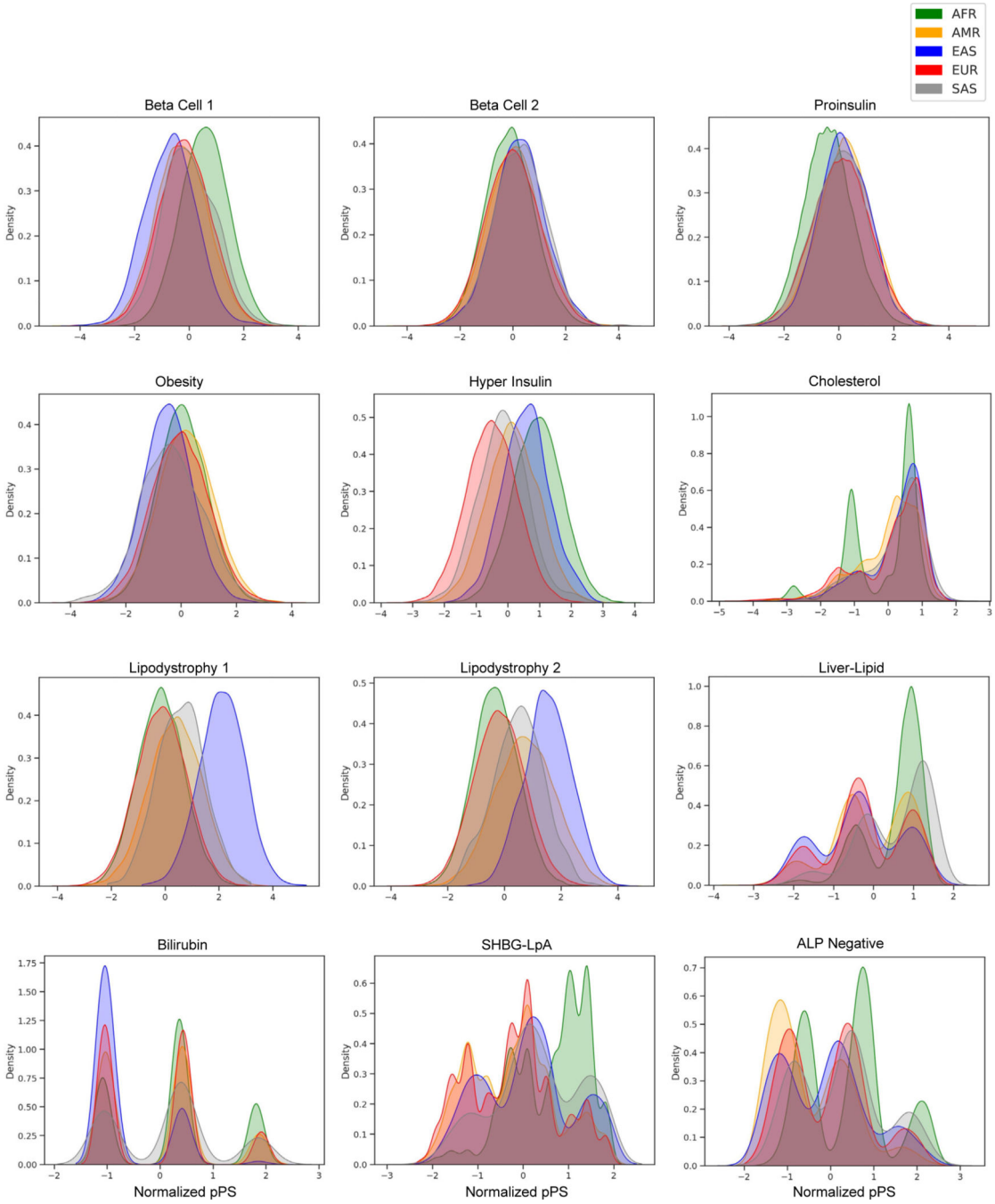
Extended Data Fig. 3.

Common T2D genetic clusters are shared across individual ancestry groups.



Extended Data Fig. 4.

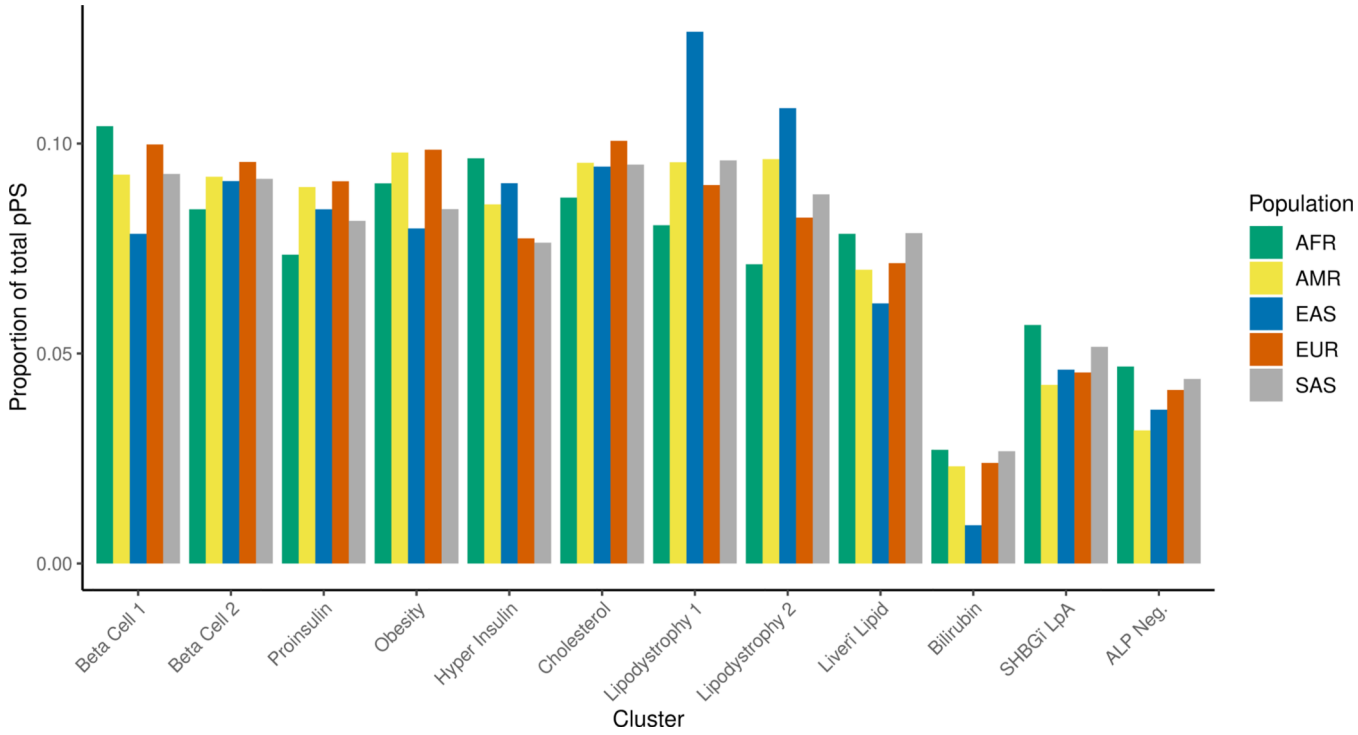
Sex-stratified association of multi-ancestry T2D genetic clusters with anthropometric traits.



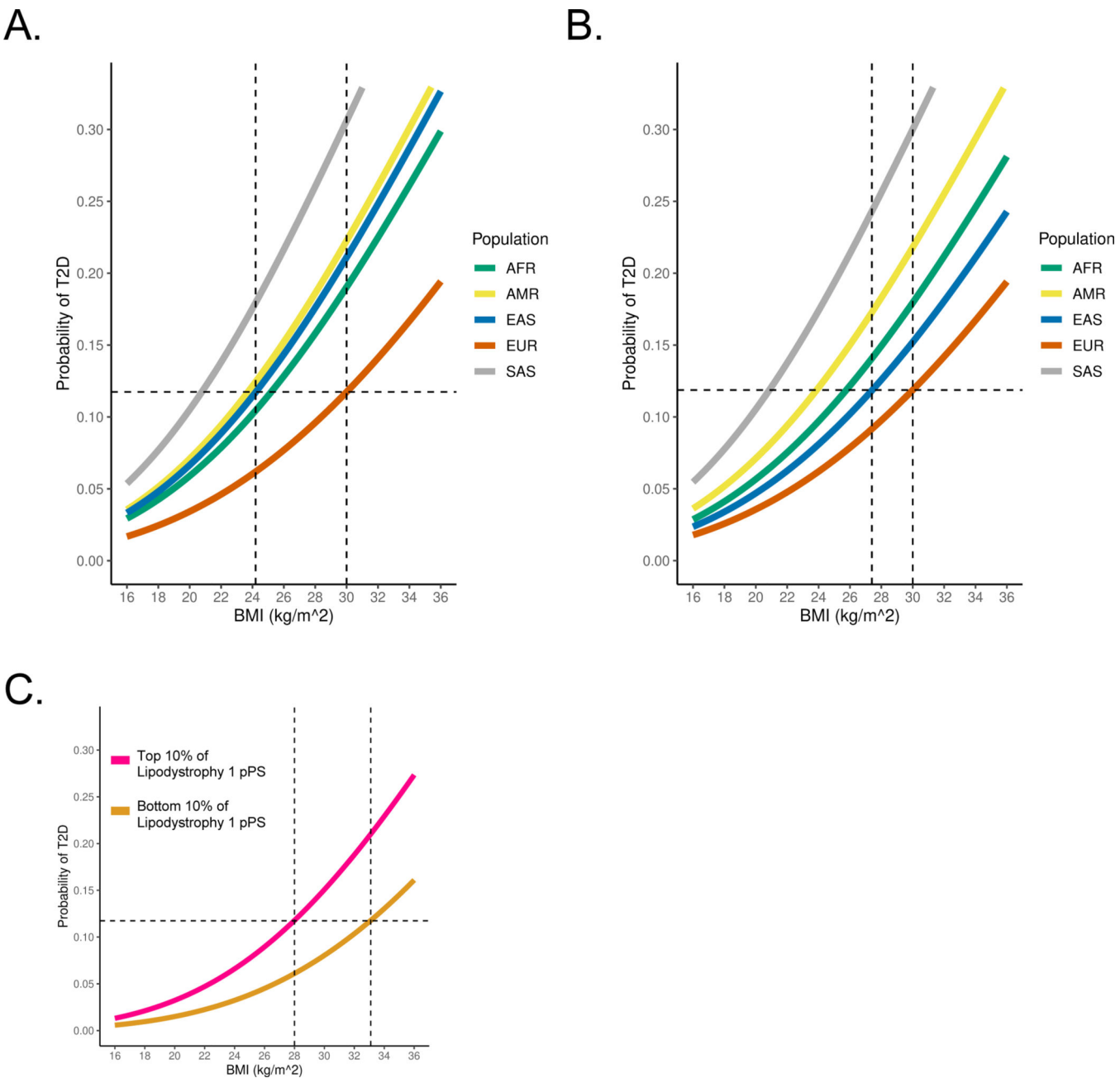
Extended Data Fig. 5.

Variation in distribution of multi-ancestry T2D genetic clusters across ancestry groups

.



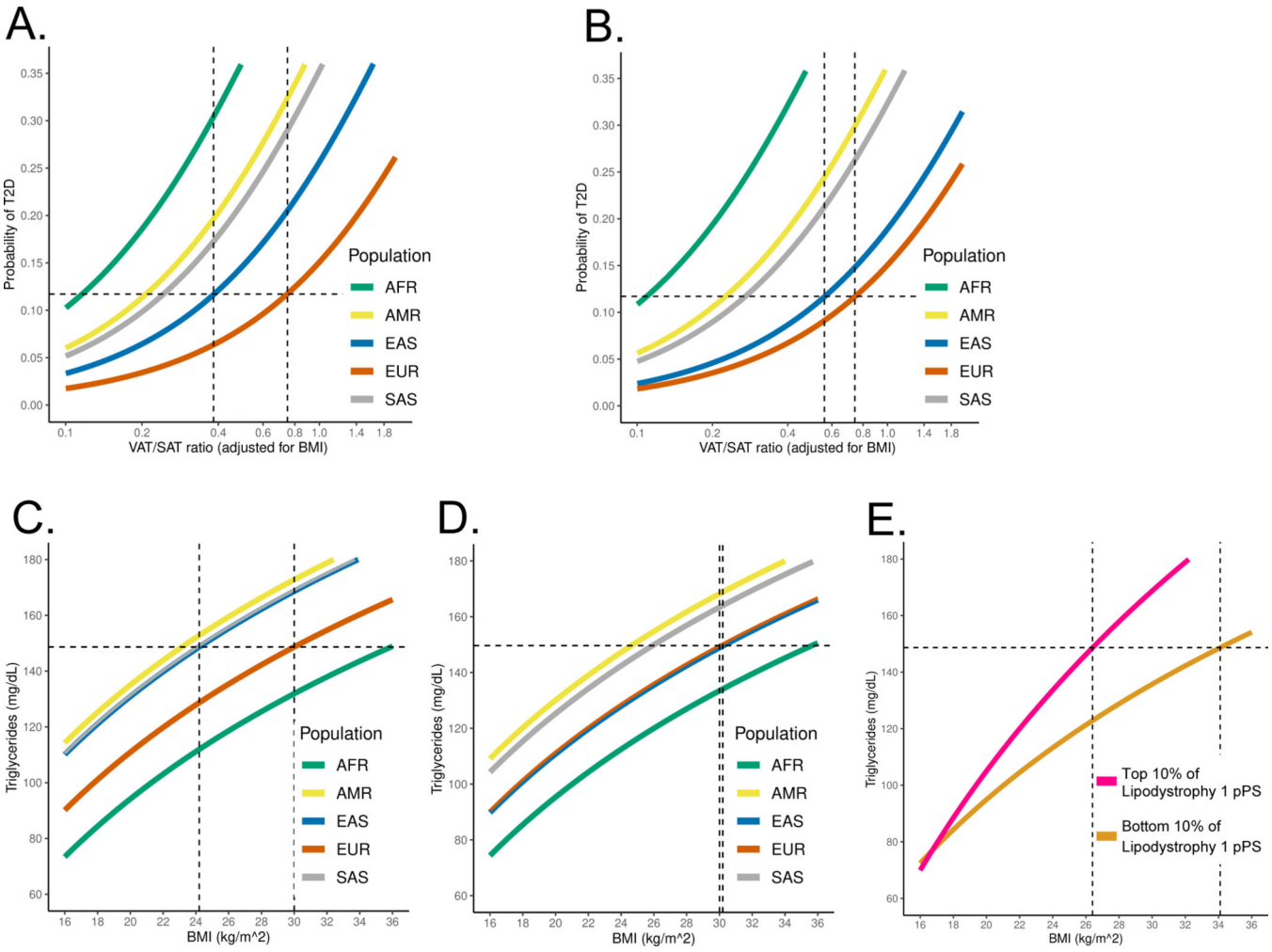
Extended Data Fig. 6.
Proportion of total T2D genetic risk attributable to each multiancestry T2D cluster.



Extended Data Fig. 7.
Validation of relationship between T2D genetic clusters, BMI, and T2D risk.

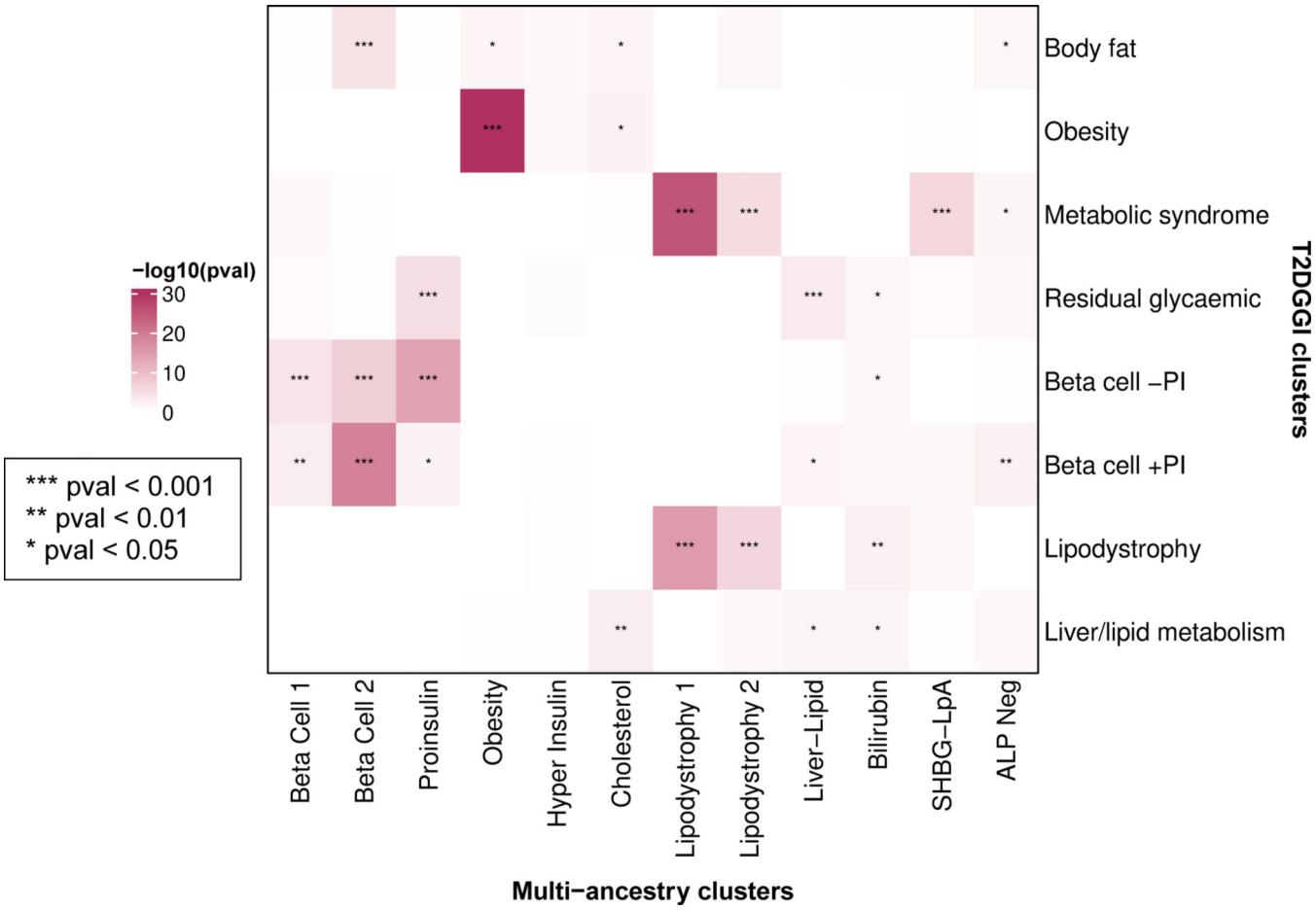
Extended Data Fig. 8.

Comparison of cluster-specific risk allele frequencies (RAF) in EUR and other ancestry groups.



Extended Data Fig. 9.

Ancestry-specific variation in adipose volume and triglycerides.



Extended Data Fig. 10.

Conservation of biological pathways between the multiancestry and T2DGGI clusters.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

AJD is supported by NIH/NIDDK T32 DK007028 and NIH/NIDDK F32 DK137487. KEW is supported by NIH K01DK133637. LS is supported by funds from the Ministry of Education and Science of Poland within the project “Excellence Initiative—Research University”, the Ministry of Health of Poland within the project “Center of Artificial Intelligence in Medicine at the Medical University of Bialystok” and American Diabetes Association grant 11-22-PDFPM-03. MC is supported by the Novo Nordisk Foundation (NNF21SA0072102), and NIDDK UM1 DK126185. JMM is supported by American Diabetes Association Innovative and Clinical Translational Award 1-19-ICTS-068, American Diabetes Association grant #11-22-ICTSPM-16 and by NHGRI U01HG011723. MSU is supported by NIDDK K23DK114551, NIDDK R03DK131249, and Doris Duke Foundation Award 2022063.

Thank you to the participants of the All of Us research program and MGB Biobank. In addition, we thank the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) for providing pre-publication access to GWAS summary statistics for post-challenge insulin resistance measures. Finally, we thank Josée Dupuis for assistance with statistical analysis.

References

1. Tobias DK et al. Second international consensus report on gaps and opportunities for the clinical translation of precision diabetes medicine. Nat. Med 29, 2438–2457 (2023). [PubMed: 37794253]
2. Misra S et al. Precision subclassification of type 2 diabetes: a systematic review. Commun. Med. (Lond.) 3, 138 (2023). [PubMed: 37798471]
3. Udler MS et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med. 15, e1002654 (2018).
4. Kim H et al. High-throughput genetic clustering of type 2 diabetes loci reveals heterogeneous mechanistic pathways of metabolic disease. Diabetologia 66, 495–507 (2023). [PubMed: 36538063]
5. Mahajan A et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. Nat. Genet 50, 559–571 (2018). [PubMed: 29632382]
6. Suzuki K et al. Multi-ancestry genome-wide study in >2.5 million individuals reveals heterogeneity in mechanistic pathways of type 2 diabetes and complications. medRxiv (2023) doi:10.1101/2023.03.31.23287839.
7. Martin AR et al. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet 51, 584–591 (2019). [PubMed: 30926966]
8. 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). [PubMed: 35551307]
9. 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). [PubMed: 32541925]
10. Spracklen CN et al. Identification of type 2 diabetes loci in 433,540 East Asian individuals. Nature 582, 240–245 (2020). [PubMed: 32499647]
11. BasuRay S, Wang Y, Smagris E, Cohen JC & Hobbs HH Accumulation of PNPLA3 on lipid droplets is the basis of associated hepatic steatosis. Proc. Natl. Acad. Sci. U. S. A 116, 9521–9526 (2019). [PubMed: 31019090]
12. Lee SM, Muratalla J, Sierra-Cruz M & Cordoba-Chacon J Role of hepatic peroxisome proliferator-activated receptor γ in non-alcoholic fatty liver disease. J. Endocrinol 257, (2023).
13. Getz GS & Reardon CA Apoprotein E and Reverse Cholesterol Transport. Int. J. Mol. Sci 19, (2018).
14. Zhang K et al. A single-cell atlas of chromatin accessibility in the human genome. Cell 184, 5985–6001.e19 (2021). [PubMed: 34774128]
15. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]
16. Pan-UKB team. https://pan.ukbb.broadinstitute.org (2020).
17. Caleyachetty R et al. Ethnicity-specific BMI cutoffs for obesity based on type 2 diabetes risk in England: a population-based cohort study. Lancet Diabetes Endocrinol 9, 419–426 (2021). [PubMed: 33989535]
18. Yaghootkar H, Whitcher B, Bell JD & Thomas EL Ethnic differences in adiposity and diabetes risk - insights from genetic studies. J. Intern. Med 288, 271–283 (2020). [PubMed: 32367627]
19. Ntuk UE, Gill JMR, Mackay DF, Sattar N & Pell JP Ethnic-specific obesity cutoffs for diabetes risk: cross-sectional study of 490,288 UK biobank participants. Diabetes Care 37, 2500–2507 (2014). [PubMed: 24974975]
20. Hsu WC, Araneta MRG, Kanaya AM, Chiang JL & Fujimoto W BMI cut points to identify at-risk Asian Americans for type 2 diabetes screening. Diabetes Care 38, 150–158 (2015). [PubMed: 25538311]
21. Rodriguez LA et al. Examining if the relationship between BMI and incident type 2 diabetes among middle-older aged adults varies by race/ethnicity: evidence from the Multi-Ethnic Study of Atherosclerosis (MESA). Diabet. Med 38, e14377 (2021).
22. Aggarwal R et al. Diabetes Screening by Race and Ethnicity in the United States: Equivalent Body Mass Index and Age Thresholds. Ann. Intern. Med 175, 765–773 (2022). [PubMed: 35533384]

23. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. Lancet 363, 157–163 (2004). [PubMed: 14726171]
24. Inker LA et al. New creatinine- and cystatin C-based equations to estimate GFR without race. N. Engl. J. Med 385, 1737–1749 (2021). [PubMed: 34554658]
25. Zhao W et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. Nat. Genet 49, 1450–1457 (2017). [PubMed: 28869590]
26. Goodarzi MO & Rotter JI Genetics Insights in the Relationship Between Type 2 Diabetes and Coronary Heart Disease. Circ. Res 126, 1526–1548 (2020). [PubMed: 32437307]
27. Sattar N et al. Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials. Lancet 375, 735–742 (2010). [PubMed: 20167359]
28. González-Lleó AM, Sánchez-Hernández RM, Boronat M & Wägner AM Diabetes and familial hypercholesterolemia: Interplay between lipid and glucose metabolism. Nutrients 14, 1503 (2022). [PubMed: 35406116]
29. Wei Y et al. Associations between serum total bilirubin, obesity and type 2 diabetes. Diabetol. Metab. Syndr 13, 143 (2021). [PubMed: 34876211]
30. Hansen M et al. Bile acid sequestrants for glycemic control in patients with type 2 diabetes: A systematic review with meta-analysis of randomized controlled trials. J. Diabetes Complications 31, 918–927 (2017). [PubMed: 28238556]
31. Glunk V et al. A non-coding variant linked to metabolic obesity with normal weight affects actin remodelling in subcutaneous adipocytes. Nat. Metab 5, 861–879 (2023). [PubMed: 37253881]
32. Fathzadeh M et al. FAM13A affects body fat distribution and adipocyte function. Nat. Commun 11, 1465 (2020). [PubMed: 32193374]
33. Li B-T et al. Disruption of the ERLIN-TM6SF2-APOB complex destabilizes APOB and contributes to non-alcoholic fatty liver disease. PLoS Genet. 16, e1008955 (2020).
34. ElSayed NA et al. 2. Classification and diagnosis of diabetes: Standards of care in diabetes-2023. Diabetes Care 46, S19–S40 (2023). [PubMed: 36507649]
35. Vyas DA, Eisenstein LG & Jones DS Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. N. Engl. J. Med 383, 874–882 (2020). [PubMed: 32853499]
36. Narayan KMV et al. Incidence and pathophysiology of diabetes in South Asian adults living in India and Pakistan compared with US blacks and whites. BMJ Open Diabetes Res. Care 9, e001927 (2021).
37. Narayan KMV & Kanaya AM Why are South Asians prone to type 2 diabetes? A hypothesis based on underexplored pathways. Diabetologia 63, 1103–1109 (2020). [PubMed: 32236731]
38. All of Us Research Program Investigators et al. The “All of Us” Research Program. N. Engl. J. Med 381, 668–676 (2019). [PubMed: 31412182]
39. Castro VM et al. The Mass General Brigham Biobank Portal: an i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. J. Am. Med. Inform. Assoc 29, 643–651 (2022). [PubMed: 34849976]
40. Kho AN et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J. Am. Med. Inform. Assoc 19, 212–218 (2012). [PubMed: 22101970]
41. Szczerbinski L et al. Algorithms for the identification of prevalent diabetes in the All of Us Research Program validated using polygenic scores – a new resource for diabetes precision medicine. bioRxiv (2023) doi:10.1101/2023.09.05.23295061.
42. Wakefield J A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am. J. Hum. Genet 81, 208–227 (2007). [PubMed: 17668372]
43. Crawford SL Correlation and regression. Circulation 114, 2083–2088 (2006). [PubMed: 17088476]
44. DiCorpo D et al. Type 2 Diabetes Partitioned Polygenic Scores Associate With Disease Outcomes in 454,193 Individuals Across 13 Cohorts. Diabetes Care 45, 674–683 (2022). [PubMed: 35085396]

45. Patel AP et al. Association of Rare Pathogenic DNA Variants for Familial Hypercholesterolemia, Hereditary Breast and Ovarian Cancer Syndrome, and Lynch Syndrome With Disease Risk in Adults According to Family History. JAMA Netw Open 3, e203959 (2020).

46. Magudia K et al. Population-Scale CT-based Body Composition Analysis of a Large Outpatient Population Using Deep Learning to Derive Age-, Sex-, and Race-specific Reference Curves. Radiology 298, 319–329 (2021). [PubMed: 33231527]

47. Bridge CP et al. A Fully Automated Deep Learning Pipeline for Multi-Vertebral Level Quantification and Characterization of Muscle and Adipose Tissue on Chest CT Scans. Radiol Artif Intell 4, e210080 (2022).

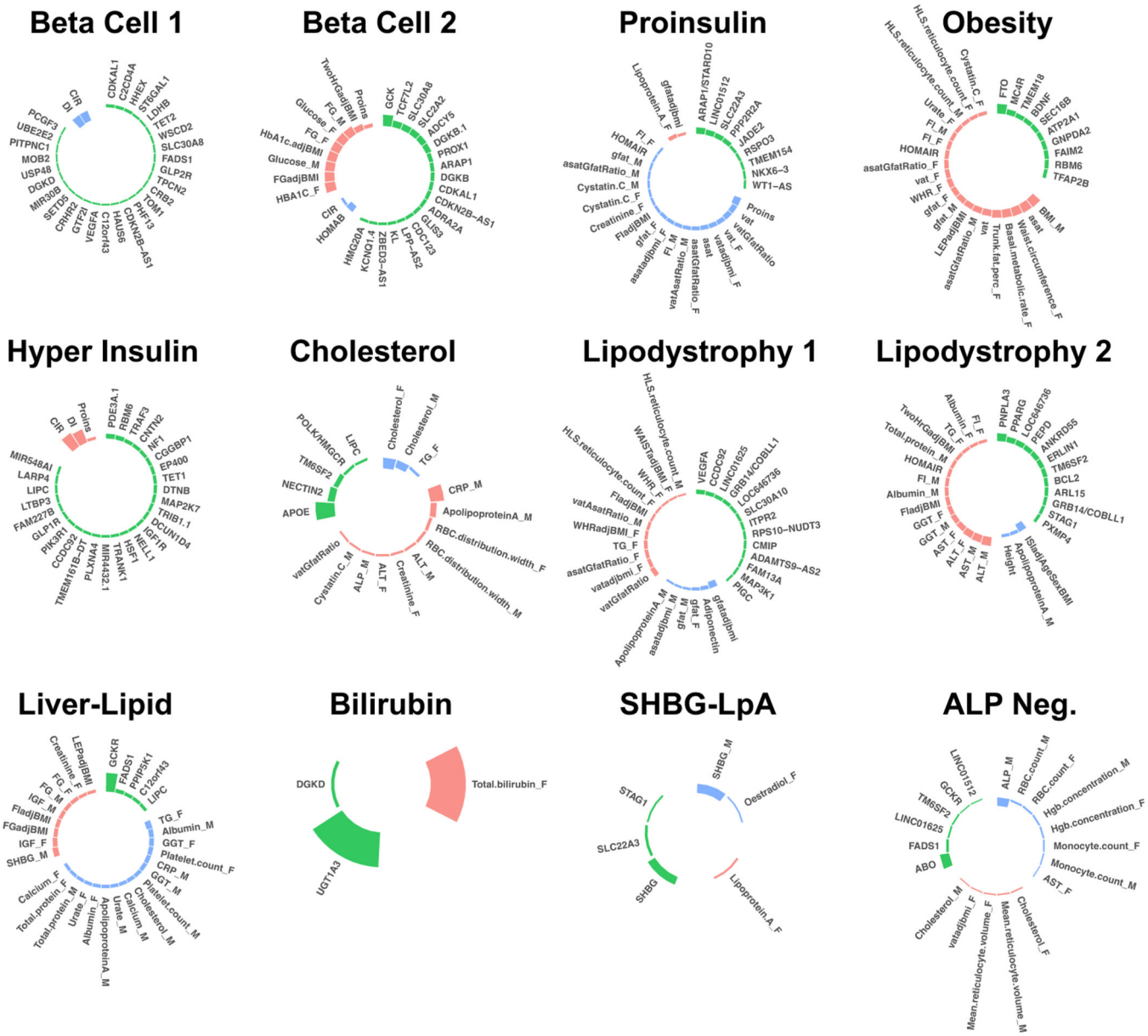


Fig. 1. Key loci and traits of multi-ancestry T2D genetic clusters

Each plot displays the top-weighted loci and traits within each multi-ancestry T2D genetic cluster. The length of the bars corresponds to the cluster weight determined by the bNMF algorithm. Green bars represent genetic loci, red bars represent traits with increased values, and blue bars represent traits with decreased values within each cluster. Female- and male-specific traits are appended with "_F" and "_M", respectively. A maximum of 30 elements (loci and traits) with the highest weights are displayed in each cluster. A legend for all abbreviations is included in Supplementary Table 3.

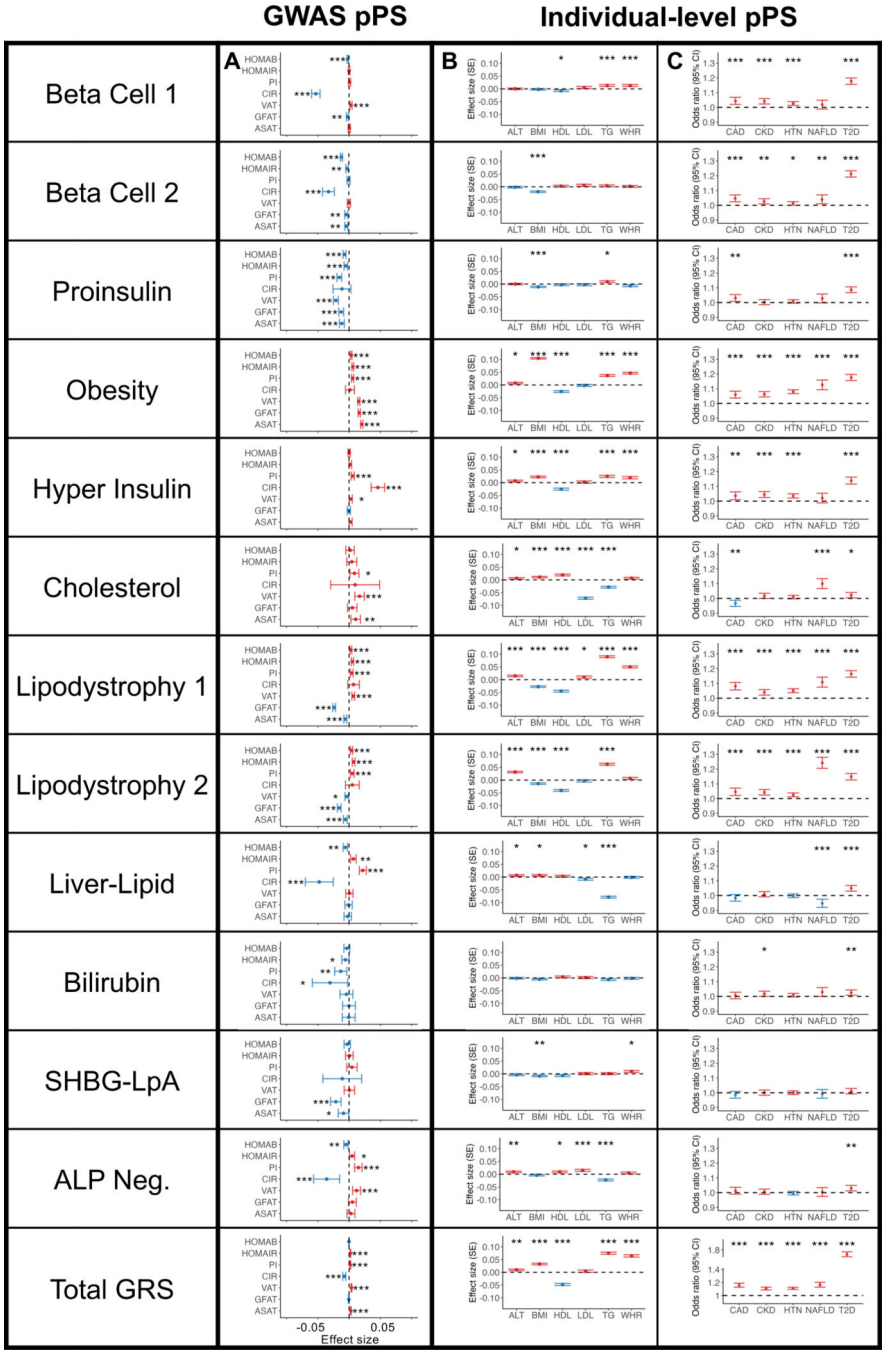


Fig. 2. Multi-ancestry T2D genetic cluster associations with continuous traits and clinical phenotypes

(A) Each plot displays associations between selected multi-ancestry T2D genetic clusters and selected continuous outcomes, based on GWAS-partitioned pPS. Each dot indicates the beta coefficient from a meta-analysis of GWAS summary statistics. Error bars represent the 95% confidence interval. PI, proinsulin; CIR, corrected insulin response; VAT, visceral adipose tissue; GFAT, gluteofemoral adipose tissue; ASAT, abdominal subcutaneous adipose tissue.

(B) Each plot displays cluster associations with selected continuous outcomes, based on individual-level pPS obtained from a meta-analysis of MGB Biobank and All of Us. Each outcome was normalized to a standard normal distribution. Each dot indicates the effect per one standard deviation increase in the pPS. Error bars represent the standard error from a linear regression model.

(C) Each plot displays cluster-specific odds ratios of selected clinical phenotypes, based on individual-level pPS obtained from a meta-analysis of MGB Biobank and All of Us. Each dot represents the odds ratios per one standard deviation increase in the pPS. Error bars represent the 95% confidence interval.

For all components, positive associations are colored in red and negative associations are colored in blue. *P* values were obtained from two-sided t tests and are indicated with asterisks (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$). A legend for all abbreviations is included in Supplementary Table 3. Complete statistics (including exact *P* values and the number of individuals measured for each phenotype) are provided in Supplementary Table 10 (Panel A), Supplementary Table 12 (Panel B), and Supplementary Table 15 (Panel C).

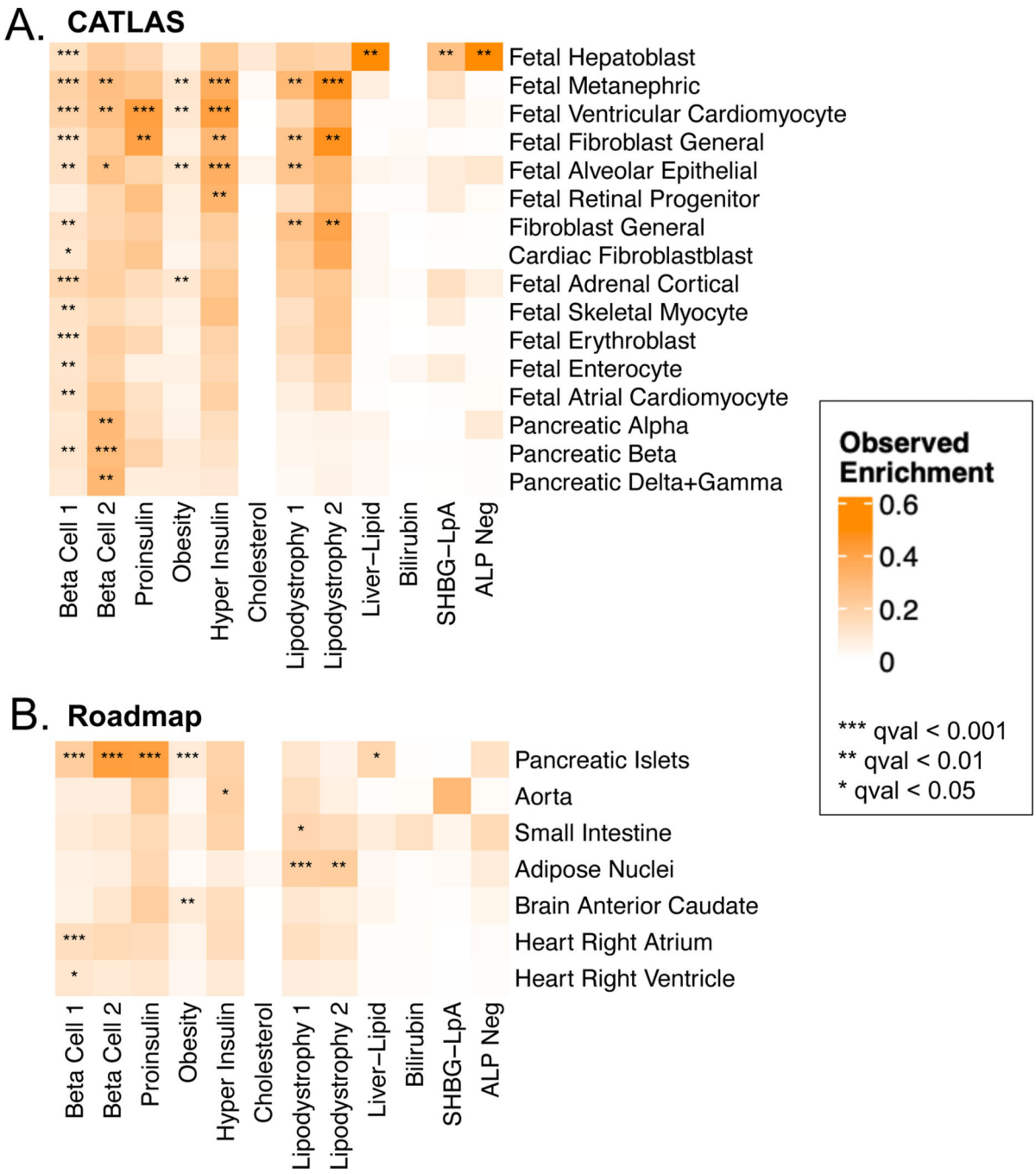


Fig. 3. Enrichment for cell type specific enhancers in multi-ancestry type 2 diabetes clusters.

Heatmaps display the significant cluster-specific enrichment of genomic annotations, represented by cumulative posterior probability, in (A) CATLAS single cell accessible chromatin data from 222 cell types and (B) Epigenomic Roadmap chromatin state calls from 28 cell types. *Q* values were corrected for false discovery rate (FDR). For both analyses, only cell types with at least one association of FDR < 0.1 are included in the figure, with additional data in Supplementary Table 17.

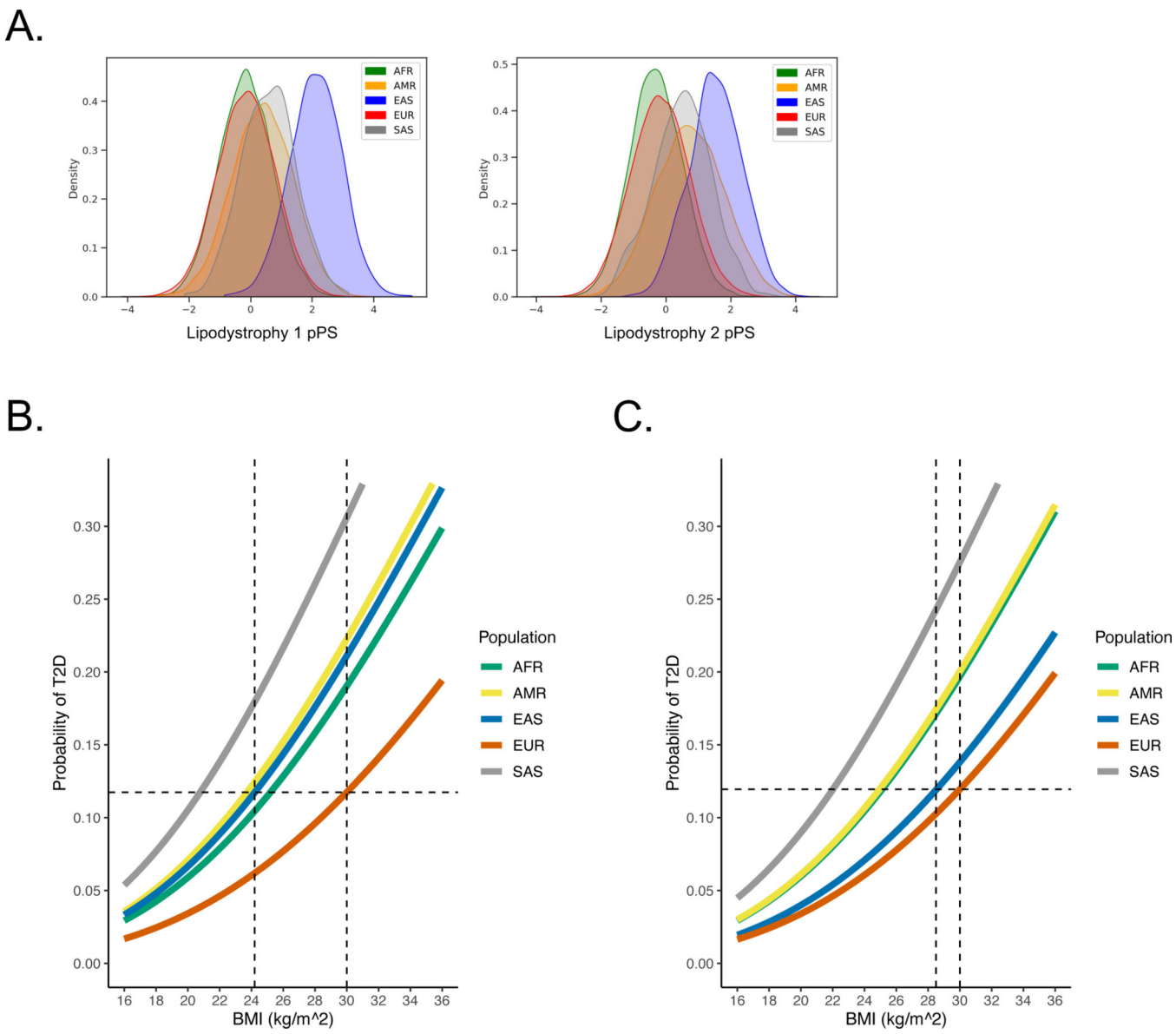


Fig. 4. Ancestry-specific relationship between T2D genetic clusters, BMI, and T2D risk

(A) Ancestry-specific distribution of Lipodystrophy 1 and Lipodystrophy 2 pPS (normalized to a standard normal distribution).

(B) Relationship between BMI and T2D risk (unadjusted), classified by genetic ancestry. T2D risk was assessed in a logistic regression model, controlling for age, sex, BMI, and genetic ancestry group. The horizontal dashed line represents the T2D risk for participants with European genetic ancestry and a BMI of 30 kg/m^2 (typically used to define obesity). The vertical dashed lines indicate the BMI thresholds needed to develop an equivalent risk of T2D in the European and East Asian ancestry groups.

(C) Relationship between BMI and T2D risk, adjusted for Lipodystrophy 1 pPS and Lipodystrophy 2 pPS.

All analyses were performed in a meta-analysis of MGB Biobank and All of Us.

Table 1.

Overview of multi-ancestry T2D genetic clusters. Refer to Supplementary Table 3 for trait abbreviations legend.

Cluster (# Variants)	Expected physiological impact	Key top-weighted traits	Key top-weighted loci	Suspected mechanism	Note
Beta Cell 1 (82)	Insulin deficiency	CIR (−), disposition index (−)	*CDKAL1, C2CD4A, HHEX, ST6GAL1, LDHB, TET2*	Beta cell function, glucose homeostasis	Recaptures part of Beta Cell cluster from Udler et al 2018 [3] and Beta Cell 1 from Kim et al 2022 [4]
Beta Cell 2 (40)	Insulin deficiency	HbA1c female (+), FGadjBMI (+), glucose male (+), proinsulin (+)	*GCK, TCF7L2, SLC30A8, SLC2A2, ADCY5, DGKB*	Beta cell function, insulin processing	Recaptures part of Beta Cell cluster from Udler et al 2018 [3] and Beta Cell 2 from Kim et al 2022 [4]
Proinsulin (16)	Insulin deficiency	PI (−), VAT (−)	*ARAP1/ STARD1, LINC01512*	Insulin synthesis	Recaptures Proinsulin cluster from Udler et al 2018 [3] and Kim et al 2022 [4]
Obesity (76)	Insulin resistance	BMI male (+), SAT (+), waist C female (+), Trunk fat % female (+)	*FTO, MC4R, TMEM18, BDNF*	Obesity-mediated insulin resistance	Recaptures Obesity cluster from Udler et al 2018 [3] and Kim et al 2022 [4]
Hyper Insulin (41)	Insulin resistance	DI (+), CIR (+)	*PDE3A, RBM6, TRAF3, CNTN2*	Insulin secretion, inflammation	Recaptures Hyper Insulin cluster from Kim et al 2022 [4]
Cholesterol (5)	Insulin resistance	CRP male (+), Cholesterol (−), Apolipoprotein A (+)	*APOE, NECTIN2, TM6SF2, POLK/ HMGCR*	*HMGCR* expression	New cluster in this study
Lipodystrophy 1 (47)	Insulin resistance	GFATAdjBMI (−), VAT/ GFAT (+), adiponectin (−)	*VEGFA, CCFC92, LINC01625/ CITED2, GRB14/ COBLL1, FAM13A*	Fat distribution-mediated insulin resistance	Recaptures Lipodystrophy cluster from Udler et al 2018 [3] and Kim et al 2022 [4]
Lipodystrophy 2 (29)	Insulin resistance	ALT (+), ISIadjAgeSexBMI (−), AST (+), GGT (+)	*PNPLA3, PPARG, LOC646736/ IRS1, PEPD, ANKRD55, ERLIN1*	Hepatic steatosis	New cluster in this study; split from previous Lipodystrophy cluster
Liver-Lipid (7)	Insulin resistance	TG female (−), SHBG male (+), IGF female (+), Albumin male (−)	*GCKR, FADS1, PPIP5K1*	Liver/lipid metabolism	Recaptures Liver-Lipid cluster from Udler et al 2018 [3] and Kim et al 2022 [4]
Bilirubin (2)	Unclear	Bilirubin (+)	*UGT1A3*	Bilirubin metabolism	New cluster in this study
SHBG-LpA (3)	Unclear	SHBG male (−), Lp(a) female (+), oestradiol female (−)	*SHBG, SLC22A3, STAG1*	SHBG and Lp(a) metabolism	Merged from LpA and SHBG clusters from Kim et al 2022 [4]
ALP Negative (6)	Insulin resistance	ALP (−), RBC count (−), Hgb concentration (−)	*ABO, FADS1*	ALP activity levels	Recaptures ALP Neg cluster from Kim et al 2022 [4]